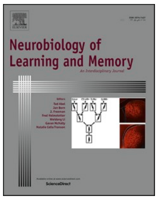
Neurobiology of Learning and Memory 206 (2023) 107860

Contents lists available at ScienceDirect

Neurobiology of Learning and Memory

journal homepage: www.elsevier.com/locate/ynlme



The emergence of task-relevant representations in a nonlinear decision-making task

N. Menghi [a,b,*], F. Silvestrin [a], L. Pascolini [a], W. Penny [a]

[a] University East Anglia, School of Psychology, UK
[b] Max Planck for Human Cognitive and Brain Sciences, Department of Psychology, Germany

ARTICLE INFO

Keywords:
Representation learning
Decision-making
EEG
Representation similarity analysis

ABSTRACT

This paper describes the relationship between performance in a decision-making task and the emergence of task-relevant representations. Participants learnt two tasks in which the appropriate response depended on multiple relevant stimuli and the underlying stimulus-outcome associations were governed by a latent feature that participants could discover. We divided participants into good and bad performers based on their overall classification rate and computed behavioural accuracy for each feature value. We found that participants with better performance had a better representation of the latent feature space. We then used representation similarity analysis on Electroencephalographic (EEG) data to identify when these representations emerge. We were able to decode task-relevant representations in a time window emerging 700 ms after stimulus presentation, but only for participants with good task performance. Our findings suggest that, in order to make good decisions, it is necessary to create and extract a low-dimensional representation of the task at hand.

1. Introduction

As we learn, we create representations of the world, which we use to make decisions in different contexts. These representations are informed and reconstructed by sensory input into sensory-independent spaces where knowledge is encoded in a map-like format (Behrens et al., 2018; Park et al., 2020; Niv, 2019). Inputs can be conceived as coordinate points in a map, where their relative positions are defined according to relevant features (Viganò et al., 2021; Constantinescu et al., 2016). For example, we could represent felidae according to four physical dimensions: size, fur, wilderness and coat colour. Leopards and cheetahs would score similarly in these dimensions, taking a close position on the map. Cats, on the other hand, would take a fairly distant spot. In these maps, distances code for similarity, so that similar inputs are positioned closer than dissimilar ones (Theves et al., 2019).

1.1. *Contextual reinstatement*

Neuroscientific evidence shows that neural representation for stimuli occupying near positions in these maps is similar (O`Keefe and Speakman, 1987; Bellmund et al., 2018). Following up with the cat example, the neural code associated with cheetahs would be very similar to the one associated with a leopard, as both have similar features. In addition, every time an element of these maps is recalled, we do not only recall information associated with the event itself, but also information associated with neighbouring elements (its context). This process is called contextual reinstatement (Davachi, 2006) and has been associated with behavioural effects like memorization and recall in word list tasks (Manning et al., 2011), generalization, inference (Morton et al., 2020; Zeithamova et al., 2012) and probabilistic reward learning (Luyckx et al., 2019).

1.2. *Dimensionality*

Cognitive maps are typically high-dimensional if their purpose is to provide accurate reconstructions of sensory inputs, but low-dimensional if they are to be used in specific tasks (Radulescu et al., 2021; Badre et al., 2021). This compression into a lower dimension can be seen as rule extraction or structure learning that provides an encoding of the relationship between inputs, actions and outcomes (Benna and Fusi, 2021; Penny et al., 2022; Braun et al., 2010). Neural activity in high-dimensional maps can represent different inputs as orthogonal activity patterns whereas neural patterns in lower-dimensional maps can identify and group a set of inputs thereby facilitating generalization and

[*] Corresponding author at: University East Anglia, School of Psychology, UK.
E-mail address: menghi@cbs.mpg.de (N. Menghi).

https://doi.org/10.1016/j.nlm.2023.107860
Received 9 March 2023; Received in revised form 26 September 2023; Accepted 6 November 2023
Available online 10 November 2023
1074-7427/© 2023 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

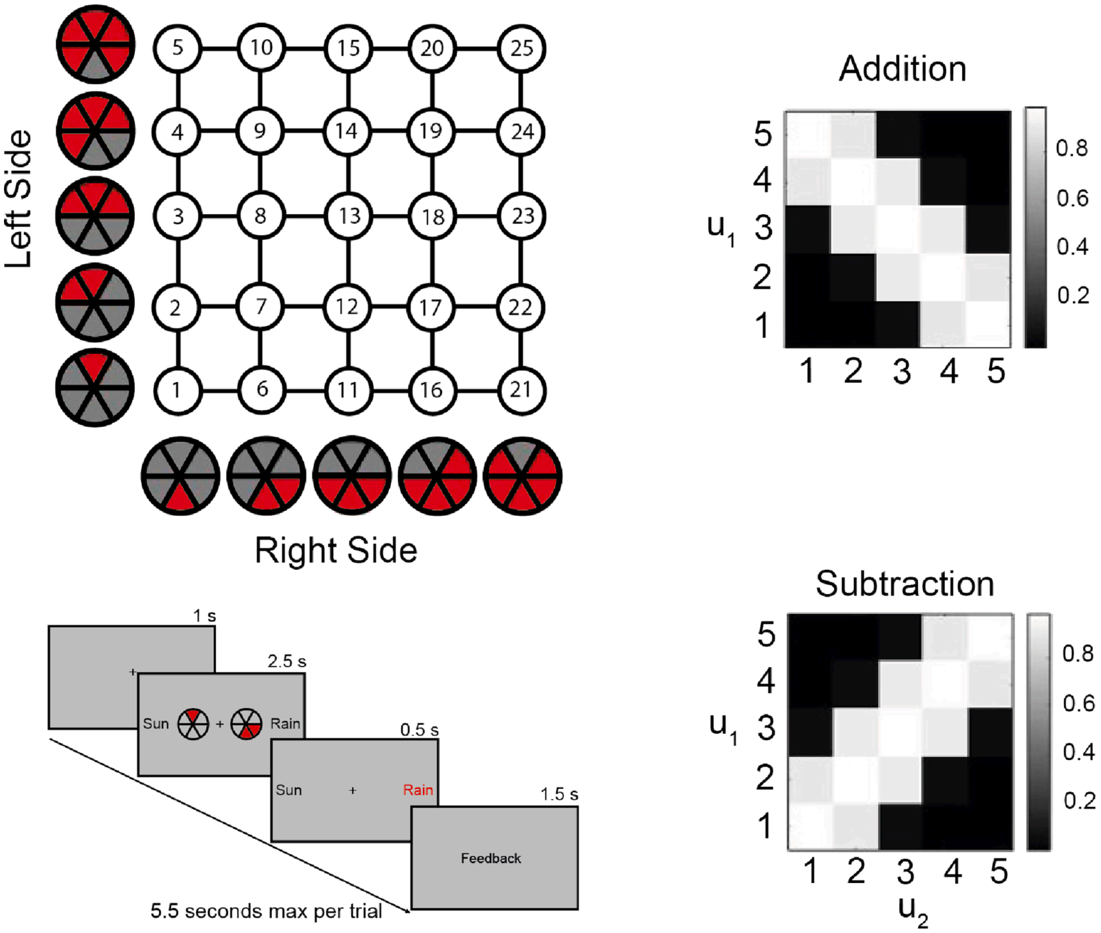


Fig. 1. **Stimuli, Trial Structure and Stimulus-Outcome Mappings** *The top left panel shows the experimental stimuli. Each pie on the left side can be combined with each pie on the right side, creating 25 potential stimulus configurations. The bottom left panel shows the trial structure. In the right panel, the gray scale image plots the Sun Outcome probability (given button press "sun"), as a function of the number of red slices in the right side pie, u_1, and left side pie, u_2.*

transfer learning (Menghi et al., 2021; Radulescu et al., 2021). Coming back to the felidae example, high-dimensional maps can differentiate between individual animal belonging to the felidae family whereas low-dimensional maps could group them based on their different subfamilies and breeds.

1.3. *Temporal dynamics of representation learning*

The temporal dynamic of hierarchical/abstract representations has been investigated mostly by capitalizing on existing categories. Categorization can be made at different levels of abstraction, such as the high dimensional recognition of a single animal belonging to the felidae family, breed or subfamilies. It could be a specific feline, like my cat, Shi, or a breed like a Maine coon or an animal belonging to a different subfamily like a lion. Current theories suggest that stimulus processing in different levels of abstraction occurs in parallel at around 200 ms after stimulus onset (Fabre-Thorpe, 2011). The experience here could play a role. Well-consolidated categories could have separate access and parallel processing. It has been only recently that the attention has been directed to representation learning and their temporal dynamics (King and Dehaene, 2014; Hubbard et al., 2019). Multivariate approaches such as Representation Similarity Analysis (RSA) (Sols et al., 2017; Kriegeskorte et al., 2008), temporal generalization (King and Dehaene, 2014; Wolff et al., 2017) and classification (Hubbard et al., 2019) have been used to describe and characterize the dynamics of these representations over time. Recent studies compared stimuli belonging to different tasks and categories, like numbers and symbols, but organised in the same one-dimensional manifold. Common representations here would index an abstract map where stimuli belonging to different tasks can be compared, facilitating generalization and transfer of knowledge. Representations of items within a category emerged at about 100 ms after stimulus onset and a later, abstract representation between categories, representation at about 300 ms to 650 ms (Luyckx et al., 2019; Teichmann et al., 2018). Our goal in this study is to use EEG and multivariate data analysis to uncover the times at which task-relevant low-dimensional representations emerge.

1.4. *Learning task and hypotheses*

We designed an experiment that allowed us to assess the multiple cognitive processes that unfold over time during nonlinear decision making. We used a revised version of the Weather Prediction Task (Knowlton et al., 1994) in which participants learnt the association between configurations of graphical pies and a weather outcome (sun or rain). However, a major difference is that in our tasks there are hidden features in the stimulus-outcome mappings that can be discovered by participants. All study participants learnt two tasks each with a different latent feature. We hypothesized that participants who performed well built an abstract and low-dimensional representation of the task.

2. Methods

2.1. *Participants*

A total of 25 students from the University of East Anglia (mean age = 20.88, SD = 4.94, 7 males, 28 females) participated in the

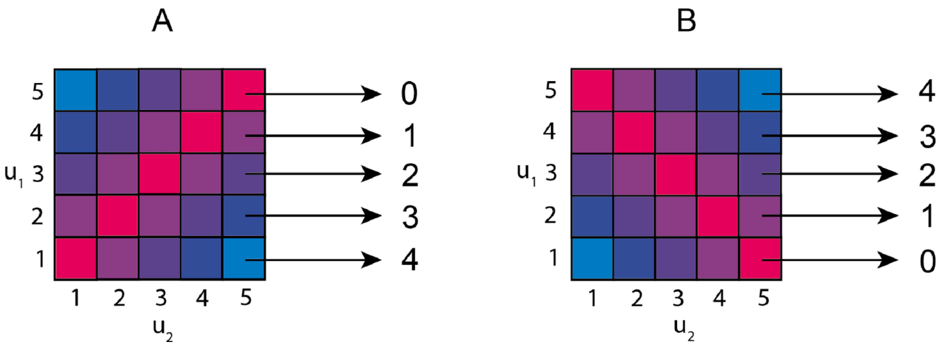


Fig. 2. Feature Values in (A) Subtraction and (B) Addition tasks. *Each task structure can be described by a one-dimensional manifold determined by the feature value computed over the number of slices of the two pies,* u_1 *and* u_2*. This feature is subtraction for the subtraction task and addition for the addition task. Extremal feature values (0 and 4) map onto extremal outcome probabilities, of 0 and 1. Intermediate feature values map onto intermediate outcome probabilities where it is less clear what decision should be made.The outcome probabilities (after deciding sun) are:* $p(outcome|feature = 0) = .978, p(outcome|feature = 1) = .90, p(outcome|feature = 2) = .056, p(outcome|feature = 3) = .001, p(outcome|feature = 4) = 0$*. We hypothesise that participants who do well at the task will have identified the underlying task manifold.*

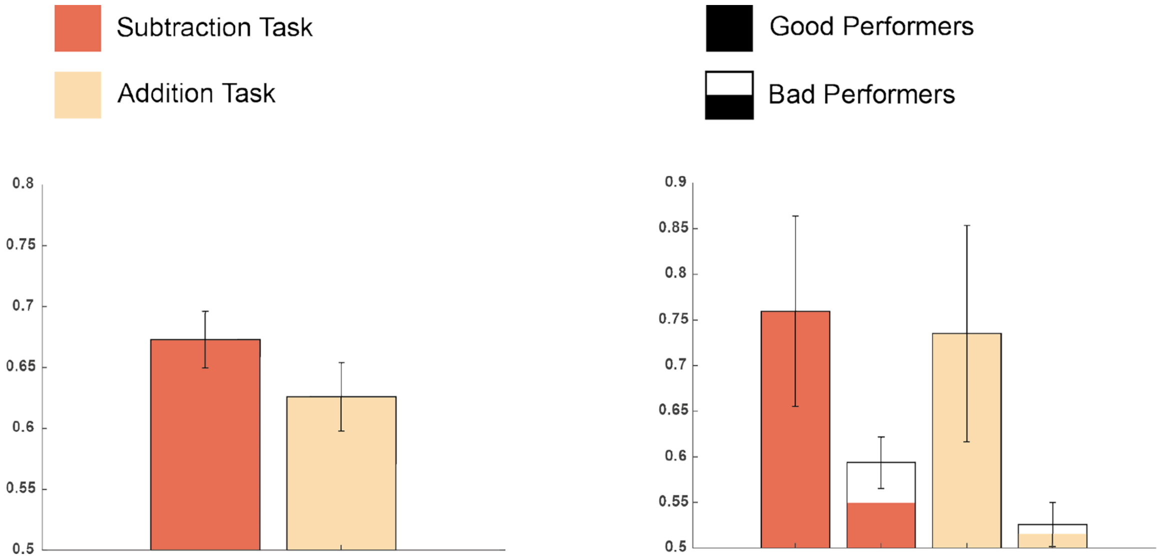


Fig. 3. Overall Accuracy The bar plots on the left show the mean accuracies for subtraction and addition tasks. The bar plots on the right show the mean accuracies for subtraction and addition tasks for good and bad performers..

experiment. All of them were naive to the purpose of the experiment. Data from one participant became unavailable due to EEG-computer synchronization errors. Data from a further participant was discarded because their performance was below chance level in both tasks. We performed our analysis on the remaining sample of 23 participants (mean age = 20.86, SD = 5.15, 6 males). All participants gave informed written consent, and the study procedure was approved by the local institutional review board of the University of East Anglia, UK. At the end of the experiment, participants received course credits for their participation.

2.2. EEG acquisition and preprocessing

BrainProduct actiCAP was used to record EEG signals from 63 electrodes plus one additional electrode used as a horizontal electrooculogram (hEOG). EEG electrodes were placed following the standard 64-channel arrangement, FT9 was used as hEOG and FT10 as Iz. All electrode impedances were kept below 25 $k\Omega$. EEG signals were recorded at a sampling rate of 1000Hz. Preprocessing was carried out using Fieldtrip toolbox for MATLAB (Oostenveld et al., 2011). Continuous data were highpass filtered at 0.1Hz and re-referenced to the common average. The data were epoched from 500 ms before the onset of the stimulus to 1.5 s following it. We visually inspected these epochs to remove trials containing muscle activity and electrical artifacts, and identified bad electrodes which were then interpolated to the weighted average of neighbouring electrodes. A maximum of 2 non-neighbouring electrodes were interpolated per participant. We interpolated one electrode for 9 participants, two for 4 participants and none for the remaining 10. Fast Independent Component Analysis (fastICA) (Comon, 1994) was then performed on the epoched data. ICA components were visually inspected to reject eye blinks, eye movements and sustained high-frequency noise. No trials were discarded during this procedure. Furthermore, we performed baseline correction based on the whole epoch as the period pre-onset may have contained task-related cognitive activity. EEG epochs were then low-pass filtered with a cut-off of 100Hz and notch filtered at 50Hz to remove mains artefact. Finally, we visually reinspected the epochs to ensure no artifact remained. Rejected trials and hEOG signals were excluded from all further analyses.

2.3. Apparatus and stimuli

The experiment was performed in a dimly lit room with participants seated 60 cm away from a computer display with their head supported by a chin-rest. Stimuli were presented on a 23-inch HP Elite Display 240c monitor using the Psychophysics Toolbox (http://psychtoolbox.org/) (Brainard, 1997) for Matlab (Mathworks) running on Windows 7.

Two virtual ”pies” (1 x 1 degrees of visual angle) were displayed at 1 degree from the central fixation point. Each pie was divided into six slices with from one up to five slices that could be filled with red colour, making a total of twenty-five combinations, as shown in Fig. 1. The stimuli were presented on a dark grey background.

2.4. Procedure

The experiment was composed of two consecutive tasks with two different mappings which we refer to as ”addition” or ”subtraction”, in counterbalanced order. The tasks are described in the following section. As shown in Fig. 1, each trial started with a black fixation cross presented at the center of the screen for 1000 ms. Afterwards, the stimuli appeared and stayed on screen for 2500 ms maximum or until a response was made. Responses were made on a standard keyboard, the letter ”g” indicating a prediction of sun and ”j” predicting rain. Responses not given within the required time constitute ”missed trials”. Right after button press, confirmation of the choice was given for 500 ms. Finally, feedback was provided, saying ”correct” if the prediction was correct, ”incorrect” if it was not and ”too slow” if they missed the trial (no response within 2500 ms). At the end of each block of trials, participants were required to keep their eyes on a fixation cross for one minute. Participants were explicitly instructed to maintain their gaze fixed on the central fixation cross throughout the task. This instruction was reinforced before the task began, and participants were reminded to avoid unnecessary eye movements.

In order to test participants’ knowledge about the task, at the end of each task, we asked them how they approached it and at which point in time they started approaching it that way. At the end of the experiment, we probed participants with two questionnaires, one per task, to assess their explicit knowledge of the task. The experiment plus preparation took about one hour and a half to complete. Each task lasted about thirty minutes.

2.5. Stimulus-outcome mappings

Two different Stimulus-Outcome Mappings maps were used during the experiment. These mappings were defined by an operation, addition or subtraction, that reduces the value of the configuration to a single feature value (see Fig. 2). In the ”addition” task, participants needed to make a decision based on the sum of the number of pies, whereas in the ”subtraction” task they needed to make a decision based on the difference. Both tasks were defined using a probabilistic mapping in which the log-odds of the outcome, y_t in Eq. 1, was a quadratic function of stimulus characteristics, the number of slices (u in the equation below). (see Fig. 3)

$$\log\left[\frac{p(y_t=1)}{p(y_t=0)}\right]=(u_t-\mu)^T W(u_t-\mu)+w_0 \quad (1)$$

$$\begin{aligned} W = &\ 2.4\times\begin{bmatrix}-0.71 & w_d\\ w_d & -0.71\end{bmatrix}\\ \mu = &\ [3,3]^T\\ w_0 = &\ 2\\ u_t = &\ [u_{left},u_{right}]^T\end{aligned}$$

The w_d parameter was arbitrarily chosen so that the flipping of its sign in this mapping produced either the addition or subtraction task depicted in Fig. 1 (right panel), where $w_d=0.71$ produces the subtraction map and $w_d=-0.71$ produces the addition map. The subtraction task can be approximately described with a single logical clause: ”decide Sun for same number of slices” or ”decide Sun if difference in number of slices is zero”. For the addition task we have ”decide Sun if the sum of the slices is 6 - a full pie”. Another way to describe the tasks is to ask what is the discriminatory feature; these are addition and subtraction, respectively.

2.6. Experimental design

We assess the effect of task using a within-subject design with two levels of the factor task (addition and subtraction). All participants did both addition and subtraction tasks. Each task was composed of 250 trials (10 repetitions per stimulus configuration) divided into 5 blocks. Given that participants are required to make Sun/Rain decisions and learn incrementally via feedback, this is reminiscent of the classic Weather Prediction Task (Knowlton et al., 1994). However, a major difference is that in our tasks there is a hidden structure in the stimulus-outcome mappings that can be discovered by participants.

3. EEG Data analysis

We performed univariate and multivariate analyses of the data to get a deeper insight into the relationship between task-relevant representations and participants’ performance. First, we performed event-related potentials (ERPs) analysis and a General Linear Model (GLM) on a participant group level dividing Subtraction and Addition tasks. Second, we performed representation similarity analysis on both the group level and on good and bad performers (splitted according to median accuracy) to investigate whether the representation created varied according to performance.

3.1. ERP

We performed a cluster-based permutation test of univariate within-group analysis of variance (ANOVA), comparing the five different feature values in addition and subtraction tasks. We then performed a cluster-based permutation test of within-group t-tests to compare the activity related to sun/rain categories. Cluster-based permutation testing on all the electrodes and the whole epoch was implemented using the FieldTrip software. The cluster-forming threshold as the threshold for statistical testing were set to an alpha level of 0.05 two-tails. Condition labels were randomly permuted 1000 times with the Monte Carlo method, following the default method implemented in FieldTrip. This provides an automatic method for finding significant clusters, corrected for multiple comparisons, that does not depend on a priori selection of time window and electrodes.

3.2. GLM

We then set up a GLM (Friston et al., 2007; Dobson et al., 2018) with a dependent variable given by the Stimulus Epoch EEG signal and independent variables corresponding to the feature subspace (i.e. taking the sum or difference of the number of pies) and task. The GLM had two regressors: feature values and an intercept. The intercept is a column of 1’s (the associated regression coefficient will compute the mean of the EEG signal over trials and so corresponds to the standard ERP). We ran the model for each participant, at each time point and each electrode. The dependent variables were $[N_{trials}\times 1]$ vectors of the EEG signal for each participant, time point and electrode. The corresponding GLM design matrices were of dimension $[N_{trials}\times P]$ where P is the number of regressors. Feature value regressor was set to have zero mean and unit variance. The estimated regression coefficients for each subject were then entered into a group-level analysis using the summary-statistic approach (Friston et al., 2007). At the group level, a cluster-based

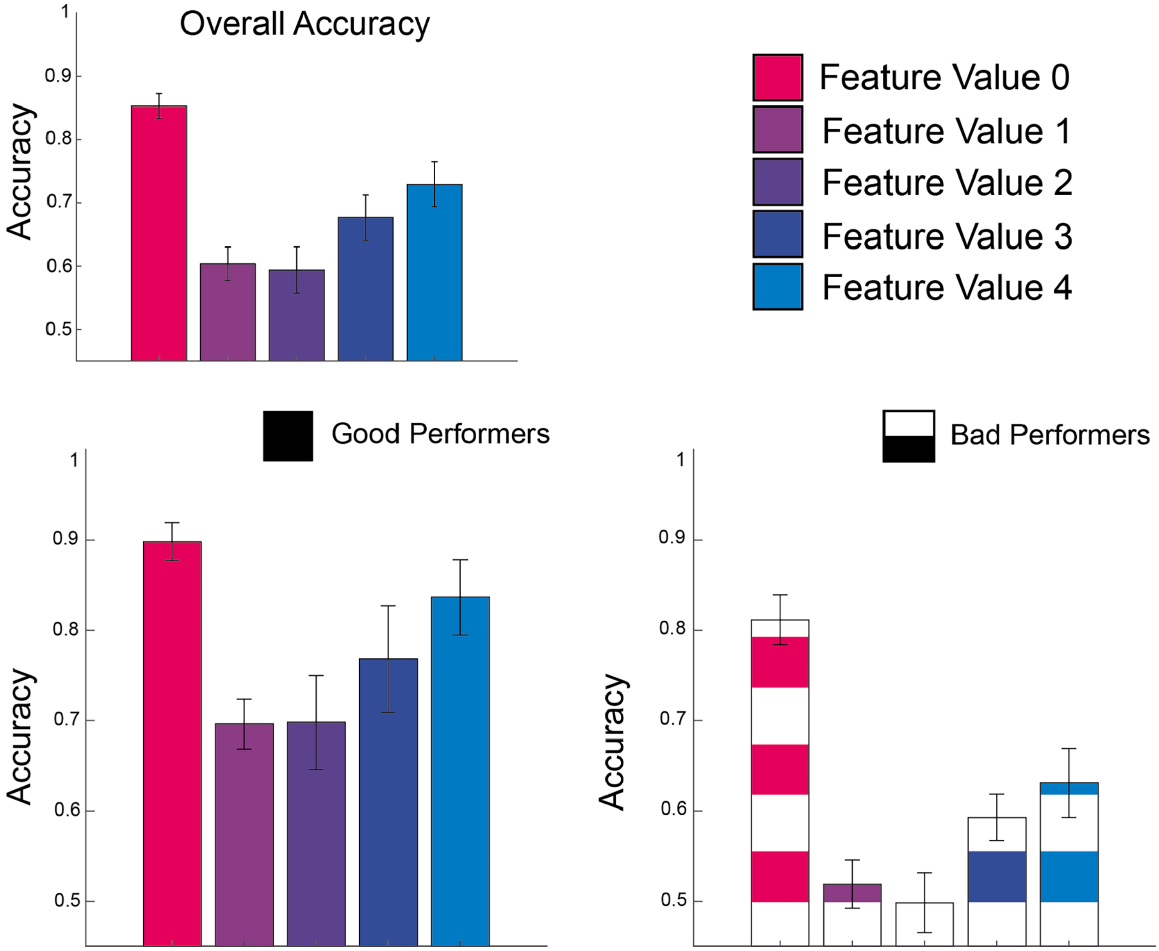


Fig. 4. Subtraction Task: Overall, Good and Bad performers accuracies by Feature Value The table show the overall mean accuracies, error represented as the standard error of the mean .

nonparametric test was implemented, following the procedure described in the papers by Samaha and colleagues, Balestrieri and Colleagues and Maris and Ostenveld (Samaha et al., 2017; Balestrieri and Busch, 2022; Maris and Oostenveld, 2007). Briefly, we multiplied a random participants subset by -1, we computed the cluster statistic by selecting all the contiguos points with a cluster-forming threshold of $p < 0.05$ two tails, creating a distribution of randomly generated cluster statistics. All the analyses were computed using Fieldtrip Toolbox (Oostenveld et al., 2011) and Matlab built-in functions (MATLAB, 2018).

3.3. Representational similarity analysis

As shown in Fig. 1, our experiment used $C = 25$ different stimulus configurations, each being a unique combination of number of slices in the left and right 'pies'. Here we used RSA to identify the relationship between these configurations and the multivariate (63-channel) EEG signals as they evolve over time (Kriegeskorte et al., 2008). We downsampled the EEG epochs to 100 Hz and selected the peri-stimulus signal from -100 ms to 1500 ms with respect to stimulus onset. To construct the neural dissimilarity matrix we computed the averaged neural response per configuration and then calculated their Spearman correlational distance. We then formed two model matrices by calculating the Euclidean distance between the configurations in two dimensions, defined by the slices in the two pies and in one dimension defined by the distance in the feature values. The Stimulus-bound model and the task-relevant model represent respectively a bidimensional and a compressed representation of the task. For each time point, we correlated (spearman) the neural dissimilarity matrix with our models. For all time points, significance was determined non-parametrically at the group level by a cluster-based permutation approach (cluster-forming threshold of $p < 0.05$ two tails), corrected significance level $p < 0.05$ (Two Tails) (Maris and Oostenveld, 2007). We calculated the clusters of time points in which configurations could be discriminated.

3.4. RSA - classifier approach

Due to constraints in statistical power, in addition to employing the conventional RSA analysis, we also employed a classifier RSA approach (Cichy et al., 2014). The classifier RSA approach provided a valuable alternative enabling us to investigate the distinctions between good and bad performers with enhanced granularity.

For each time point, t we used a Support Vector Machine (SVM) classifier to discriminate stimulus configuration i from configuration j with $i = 1..C, j = 1..C$ and formed an "EEG Decoding Matrix" of dimension $25 \times 25 \times 161$. Entry $[i, j, t]$ in this matrix corresponds to the decoding accuracy computed using 10-fold cross-validation. We used all data epochs from all participants (i.e. up to 10 trials per configuration per subject over 23 subjects with trials removed if participants provided no response or EEG signals were corrupted).

For all time points, significance was determined non-parametrically at the group level by a cluster-based permutation approach comparing

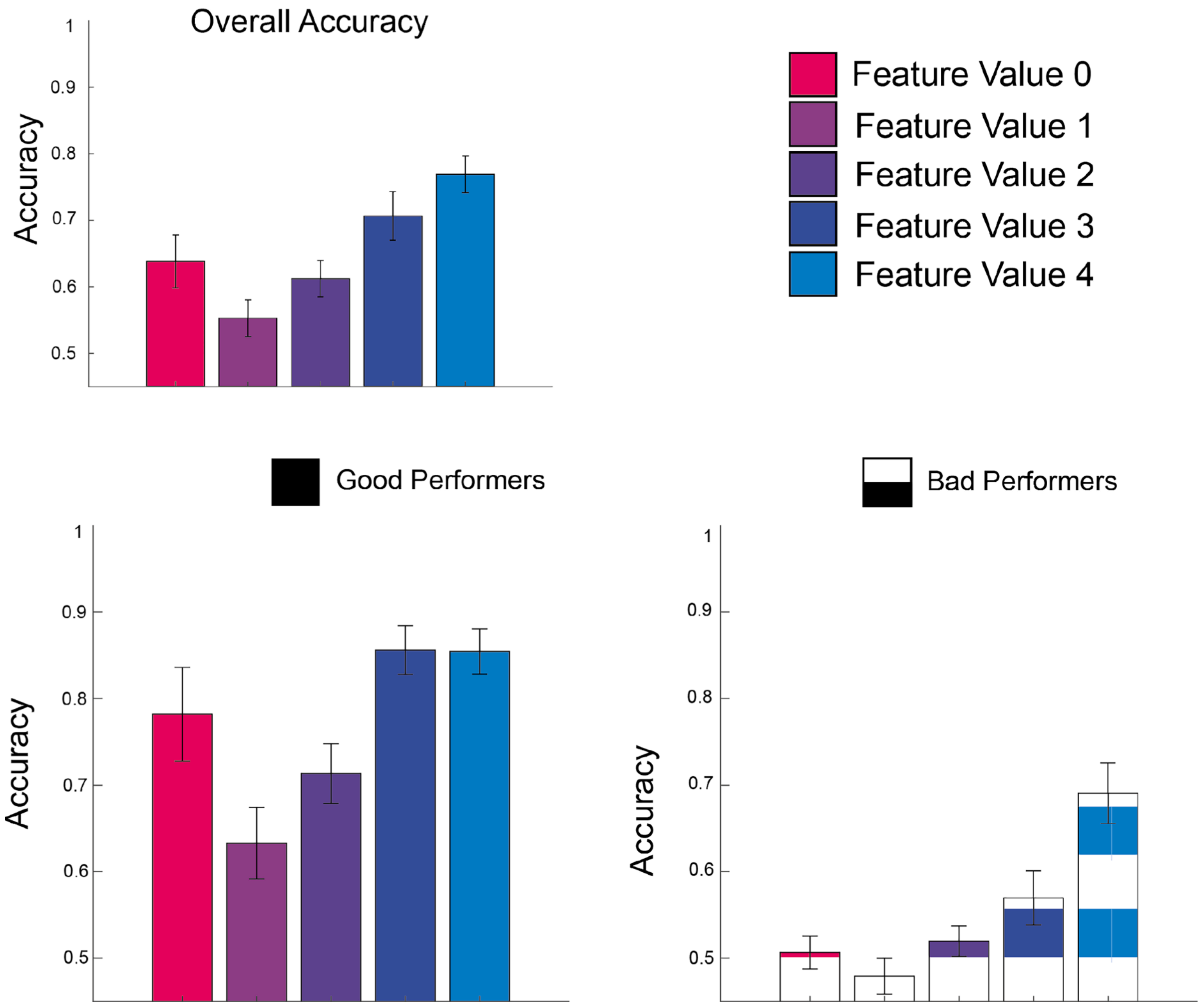


Fig. 5. Addition Task: Overall, Good and Bad performers accuracies by Feature Value The table show the overall mean accuracies, error represented as the standard error of the mean .

all rows of the decoding matrix using within or between t-tests (cluster-forming threshold of $p < 0.05$ two tails), corrected significance level $p < 0.05$ (Two Tails) (Maris and Oostenveld, 2007). We calculated the clusters of time points in which configurations could be discriminated.

3.4.1. *Time course of structure decoding*

To determine when various task representations emerge, we computed decoding accuracy over various partitions of the EEG Decoding Matrix. For each task (subtraction or addition), we partitioned the decoding matrices of good and bad performers in two different ways. The first we called the ”feature space distance” (see Section 4.3.2 below), we partitioned the decoding matrix based on the distance between feature values (see Fig. 2). We thereby compared pairs of configurations that were nearby in feature space (with feature-value discrepancies of 0 or 1) versus those that were far away (2,3 or 4). The second (see Section 4.3.3 below) partitioned the decoding matrix according to the category of the configuration (Sun or Rain). We thereby compared pairs of configurations belonging to the same or different categories.

4. Results

Our main hypothesis is that better performance will be associated with a more accurate representation of task structure (see Fig. 2). In order to identify the representations that participants created, we divided the analysis into two parts. First, in our analysis of the behavioural data, we calculated participants’ overall performance on each task and broke this down into accuracy for good versus bad performers based on a median split. We followed this up by computing accuracy as a function of feature value. Second, in our analysis of the EEG data, we ran a representational similarity analysis to relate the representation created to participants performance.

4.1. *Behavioural results*

4.1.1. *Overall accuracy*

Accuracy was computed as the correct rate over all 250 trials in each task. We performed a within-subject t-test to see if performance in the two tasks was different. We found that participants performed better in the subtraction task compared to the addition task, albeit t-test only showed a strong trend in that direction ($t(22) = 1.95$, $p(\text{one tail}) = 0.062$). We then divided participants into good and bad performers through a median split. Good performers in the subtraction task did not perform differently than good performers in the addition task ($t(20) = 0.51$, $p = 0.613$). Bad performers in the subtraction task performed better than bad performers in the addition task ($t(22) =6.366$, $p < 0.001$).

4.1.2. *Subtraction task accuracy*

Accuracy was computed as the correct rate over all trials in the subtraction task and over all trials at each feature value. We performed within-subject t-tests to see if performance was above chance and found

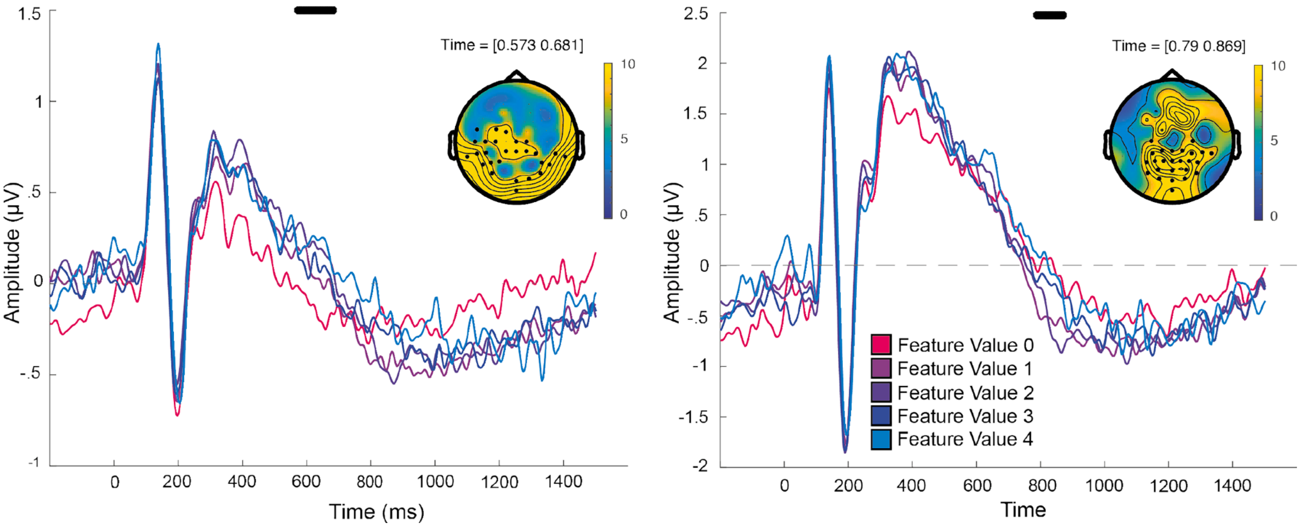


Fig. 6. Categorical Coding: Scalp maps and ERP time series. *The left (right) panel shows ERP time series and topography of the first (second) cluster. Both topographies indicate F-values on a scale from 0 to 10 and ERPs are colour-coded as in Fig. 2 (e.g. magenta indicates a feature. value of zero).*

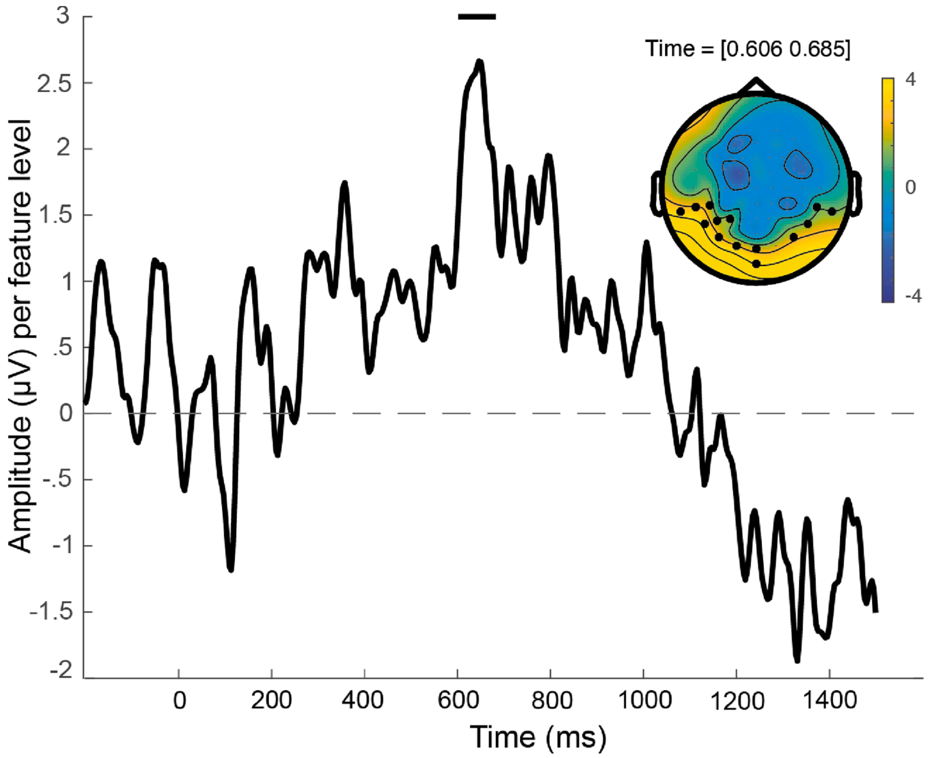


Fig. 7. Metric Coding: Scalp map. *The plot shows the time series and topography of the positive cluster. The t-values are on a scale from −4 to 4.*

that participants performed better than chance at every feature value (See the top panel of Fig. 4). For a summary of accuracies and results of t-tests, see table in the supplementary materials.

We then divided participants into good and bad performers through a median split. Good performers performed better than chance at every feature value. (See bottom left panel in Fig. 4). For a summary of accuracies and results of t-tests, see table in the supplementary materials.

Bad performers performed above chance at extremal feature values (0, 3 and 4) but not at intermediate values. (See bottom righ panel in Fig. 4). For a summary of accuracies and results of t-tests, see table in the supplementary materials.

4.1.3. *Addition task accuracy*

Accuracy was computed as the correct rate over all trials in the addition task and over all trials in each subspace value. We performed within-subject t-tests to see if performance in each subspace was above chance and found significant results except at feature value 1. (See top panel in Fig. 5), for a summary of accuracies and results of t-tests, see table in the supplementary materials.

We then divided participants into good and bad performers through a median split. Good performers performed above chance in each subspace value. (See bottom left panel in Fig. 5), for a summary of accuracies and results of t-tests, see table in the supplementary materials.

Bad performers were above chance only at extremal feature value 4. (See bottom left panel in Fig. 5), for a summary of accuracies and results

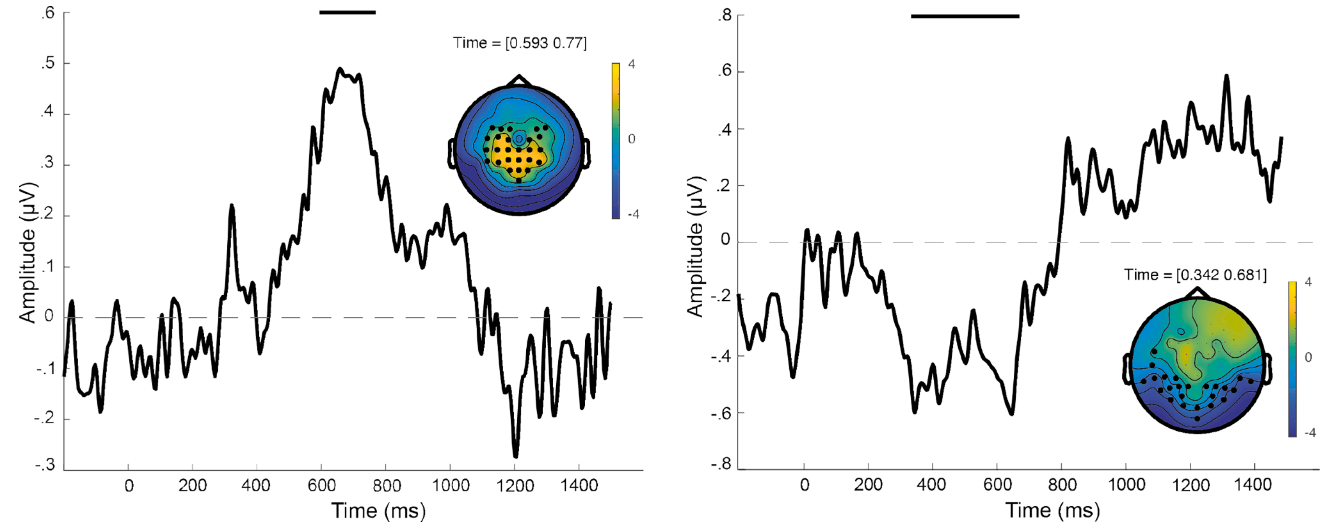


Fig. 8. Same-Different coding: Scalp maps and ERP time series. *The left (right) panel shows ERP time series and topography of the positive (negative) cluster. The topographies indicate t-values on scales between* -4 *and 4.*

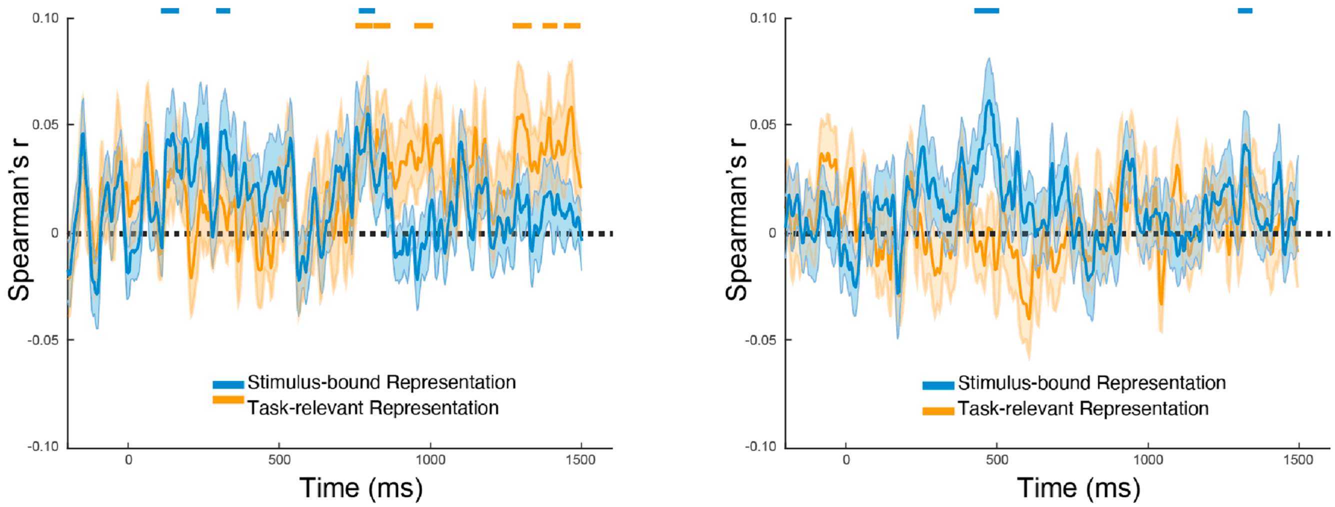


Fig. 9. Stimuls-bound and task-relevant representation of the task *Time course of the correlation between stimulus-bound and task-relevant models and the neural dissimilarity matrix. The left panel shows the results for the subtraction task, the right one for the addition task.*

of t-tests, see table in the supplementary materials.

4.1.4. *Reaction times*

Reaction time was computed over all 250 trials in each task. Participants took on average 1.05 s (σ_M =.05) to respond in Subtraction task and 1.10 s (σ_M =.04) in addition task. We performed a within-subject t-test to see if performance in the two tasks was different. We found no differences between the two tasks (t(22) = 1.63, p = 0.116). See the supplementary materials for a more detailed visual representation of reaction time across trials.

4.2. *Univariate EEG results*

We first test for the existence of a categorical representation of feature values - that is, where different levels of feature value (0,1,2,3,4) result in different ERP traces. This can be tested by looking at the main effect of feature value in a standard ANOVA. We then test for a metric representation in which the magnitude of the ERP trace changes linearly with feature value, which can be tested for using a regression approach in which a single independent variable encodes feature value. Finally, we test to see whether a feature value of zero is encoded differently to other feature values. For the subtraction task this corresponds to a ”same-different” encoding (do the pies have the same or different number of coloured slices?).

4.2.1. *Categorical coding of feature value*

In the subtraction task, cluster permutation analysis of variance revealed two significant clusters as shown in Fig. 6. The first is an occipito-temporal and central cluster between 570 ms and 680 ms after stimulus onset. The second is an occipito-parietal cluster between 790 ms and 870 ms. No results were found for the addition task.

4.2.2. *Metric coding of feature value*

The regressors for this analysis were X_1 = Feature Value and X_2 = Offset. The dependent variable was the Stimulus Epoch EEG signal. We run one model per task (addition or subtraction). Cluster permutation analysis revealed a significant cluster for the subtraction task and none for the addition task. We found an occipito-temporal cluster between 605 ms and 685 ms after stimulus onset (see Fig. 7). The significant time points and channels found here are coherent with the occipito-temporal cluster found in the previous analysis.

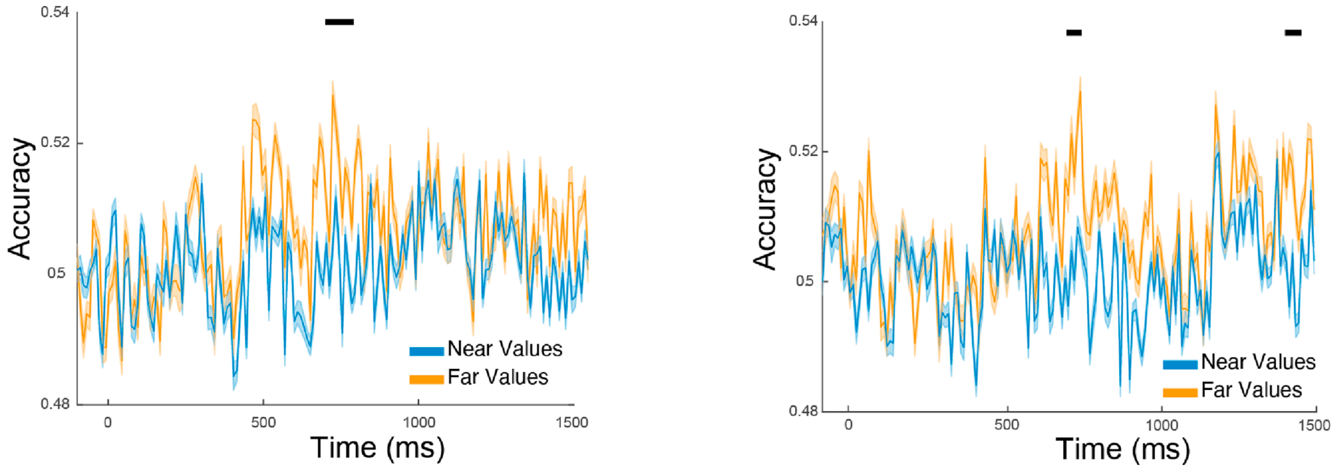


Fig. 10. Feature space distance for good performers in Subtraction (left) and Addition (right) tasks. *Time course of decoding the difference between near and far feature values. The horizontal bars above represent the significant clusters between the two conditions.*

4.2.3. Same-different coding

We then compared ERPs for the configuration belonging to feature value 0 (the main diagonals in Fig. 2), to all other feature values. We found two clusters in the subtraction task (see Fig. 8). A positive, central cluster between 590 ms and 770 ms and a negative, occipito-temporal cluster between 340 ms and 680 ms. This corresponds to a same-different encoding. There were no significant clusters in the addition task.

4.2.4. Categories

Finally, we computed ERPs related to trials predicting Sun and Rain and we found no significant results in neither subtraction or addition tasks. This shows that the differences in feature value are not due to the associated outcome but rather to how these features are processed.

4.3. Multivariate EEG results

4.3.1. Representation similarity analysis

Fig. 9 shows the correlation between the neural dissimilarity matrices and the Stimulus-bound model and task-relevant model averaged across all 25 stimulus configurations. Before and just after stimulus presentation, grand average decoding accuracy fluctuated around the chance level. In the subtraction task, the stimulus-bound representation reached significance at 124 ms (124–156 ms), followed by a cluster at 304 ms (304–324 ms) and the last cluster at 776 ms (776–804 ms). The task-relevant representation reached significance at 764 ms (764–800 ms), followed by a cluster at 820 ms (820–856 ms) a cluster at 960 ms (960–996 ms) and three late clusters (1284–1324 ms; 1384–1408 ms; 1456–1484 ms). In the addition task classification reached significance at 440 ms (440–496 ms), followed by a late cluster at 1313 ms (1313–1336 ms). No significant clusters were found for the task-relevant representation. Thus, multi-variate analysis of EEG data revealed the temporal dynamics of the task representation. First, a stimulus-bound representation emerges, providing a reconstruction of the stimulus map. Later on, compressed representation emerges, providing a reconstruction of the structure of the task at hand. We do not find an effect of the task-relevant representation in the addition task. This is likely attributed to the influence of bad performers who did not create such a representation masking so any underlying representation differences. In response, we have adopted the RSA classifier approach to elucidate the representation disparities between good and bad performers.

4.3.2. RSA - Classifier approach

To validate the efficacy of our algorithm, we conducted a group-level analysis. In order to substantiate the functionality of the algorithm, we present these results in the supplementary materials. This additional evidence serves as a demonstration of the algorithm's effectiveness in producing outcomes consistent with our prior observations, affirming its utility in decoding neural representations. The results presented in the next paragraphs are computed on good and bad performers.

4.3.3. Feature space distance

The left panel in Fig. 2 shows how the subtraction feature is related to the stimulus (configuration) space. Similarly, the right panel in Fig. 2 shows the same for the addition feature.

Here, we partitioned the decoding matrix based on the distance between these feature values. We computed average accuracy over ”near values” in the feature space (at distances 0 and 1) and ”far values” in the feature space (at distances 2, 3 and 4) by averaging the EEG decoding matrix over the relevant stimulus configurations. We then compared decoding accuracies for near versus far values as a function of peri-stimulus time with the results reported in Fig. 10. Generally, discrimination is worse for near values, implying that neural representations are more similar than for far values. We found a significant cluster for good performers in the subtraction task, significant at 700 ms (700–790 ms) but none for bad performers. Similarly, we found two significant clusters for good performers in the addition task, significant at 690 ms (690–740 ms) and at 1400 ms (1400–1450 ms) and none for the bad performers. As a control check, we tried to decode addition space distances during the subtraction task and subtraction space distances during the addition task, but found no significant clusters. This implies that only the task-relevant representations were engaged. In summary, we find the emergence of task-relevant representations at about 700 ms post-stimulus. The significant time points are close, even if they do not match, to the timeframe found with the GLM analyses.

4.3.4. Categories

We partitioned the decoding matrix based on the two categories and compared configurations predicting the same outcome to configurations predicting different outcomes. No significant results were found in good and bad performers in neither subtraction nor addition tasks. This shows that these representations are not related to a categorical distinction (Sun/Rain) but rather to how sensory stimuli are processed.

5. Discussion

To perform any task, it is necessary for the brain to process potentially high-dimensional sensory input so as to extract the relevant low-dimensional data features necessary to make good decisions. Here, we found evidence for the emergence of such task-relevant representations

Table 1
Subtraction Task: Overall accuracy by Feature Value The table show the overall mean accuracies, μ, the standard error of the mean, σ_M, and the results of the t-tests.

Feature Value	Overall Accuracy	Stats
0	$\mu = .85$, $\sigma_M = .09$	t(22) = 39.97, p < 0.001
1	$\mu = .60$, $\sigma_M = .13$	t(22) = 20.36, p < 0.001
2	$\mu = .59$, $\sigma_M = .17$	t(22) = 14.46, p < 0.001
3	$\mu = .67$, $\sigma_M = .17$	t(22) = 16.90, p < 0.001
4	$\mu = .73$, $\sigma_M = .17$	t(22) = 18.78, p < 0.001

Table 2
Subtraction Task: Good performers accuracy by Feature Value The table show the overall mean accuracies, μ, the standard error of the mean, σ_M, and the results of the t-tests.

Feature Value	Good Performers	Stats
0	$\mu = .90$, $\sigma_M = .07$	t(10) = 18.21, p < 0.001
1	$\mu = .69$, $\sigma_M = .09$	t(10) = 6.77, p < 0.001
2	$\mu = .70$, $\sigma_M = .17$	t(10) = 3.62, p = 0.004
3	$\mu = .77$, $\sigma_M = .19$	t(10) = 4.32, p = 0.001
4	$\mu = .83$, $\sigma_M = .14$	t(10) = 7.69, p < 0.001

Table 3
Subtraction Task: Bad performers accuracy by Feature Value The table show the overall mean accuracies, μ, the standard error of the mean, σ_M, and the results of the t-tests.

Feature Value	Bad Performers	Stats
0	$\mu = .81$, $\sigma_M = .09$	t(11) = 10.86, p < 0.001
1	$\mu = .52$, $\sigma_M = .09$	t(11) = 0.68, p = 0.509
2	$\mu = .50$, $\sigma_M = .11$	t(11) = −0.05, p = 0.958
3	$\mu = .59$, $\sigma_M = .09$	t(11) = 3.46, p = 0.005
4	$\mu = .63$, $\sigma_M = .13$	t(11) = 3.29, p = 0.007

Table 4
Addition Task: Overall accuracy by Feature Value The table show the overall mean accuracies, μ, the standard error of the mean, σ_M, and the results of the t-tests.

Feature Value	Overall Accuracy	Stats
0	$\mu = .64$, $\sigma_M = .19$	t(22) = 3.38, p = 0.002
1	$\mu = .55$, $\sigma_M = .13$	t(22) = 1.86, p = 0.075
2	$\mu = .61$, $\sigma_M = .13$	t(22) = 3.97, p < 0.001
3	$\mu = .70$, $\sigma_M = .17$	t(22) = 5.53, p < 0.001
4	$\mu = .77$, $\sigma_M = .13$	t(22) = 9.42, p < 0.001

Table 5
Addition Task: Good performers accuracy by Feature Value The table show the overall mean accuracies, μ, the standard error of the mean, σ_M, and the results of the t-tests.

Feature Value	Good Performers	Stats
0	$\mu = .78$, $\sigma_M = .18$	t(10) = 4.96, p < 0.001
1	$\mu = .63$, $\sigma_M = .14$	t(10) = 3.06, p = 0.012
2	$\mu = .71$, $\sigma_M = .11$	t(10) = 5.90, p < 0.001
3	$\mu = .85$, $\sigma_M = .09$	t(10) = 12.11, p < 0.001
4	$\mu = .85$, $\sigma_M = .08$	t(10) = 12.96, p < 0.001

Table 6
Addition Task: Bad performers accuracy by Feature Value The table show the overall mean accuracies, μ, the standard error of the mean, σ_M, and the results of the t-tests.

Feature Value	Bad Performers	Stats
0	$\mu = .50$, $\sigma_M = .06$	t(11) = 0.32, p = 0.749
1	$\mu = .48$, $\sigma_M = .07$	t(11) = −0.96, p = 0.356
2	$\mu = .52$, $\sigma_M = .06$	t(11) = 1.07, p = 0.305
3	$\mu = .57$, $\sigma_M = .11$	t(11) = 2.14, p = 0.053
4	$\mu = .69$, $\sigma_M = .12$	t(11) = 5.21, p < 0.001

in EEG.

Univariate data analysis provided evidence for categorical representations (Fig. 4) of latent feature value in a central cluster (570–680 ms), occipito-temporal cluster (570–680 ms) and occipito-parietal cluster (790–870 ms). We also found evidence for a metric representation (Fig. 5) of latent feature value in an occipito-temporal cluster (600–685 ms). Further analysis (Fig. 6) showed that some of this activity was likely driven by a same-versus-different representation of stimuli, although these representations started earlier and lasted longer (positive activation in occipito-temporal cluster from 340–680 ms, and negative activation in central cluster from 590–770 ms). All of these findings were for the subtraction task only.

With multivariate data analyses we found a two-dimensional representation of the task configurations from 100 ms from stimulus onset providing evidence of a faithful reconstruction of the task at hand (Fig. 9). In the subtraction task we found evidence of a compressed representation supporting the hypothesis of the emergence of a task-relevant structure. However, we did not uncover such a representation in the addition task. This absence may be attributed to the influence of bad performers who did not learn a clear representation, potentially masking any underlying structural differences among participants (see performance the supplementary materials for a figure showing bad performers accuracy across trials).

For good performers only, we found significant discrimination of trials with far but not near feature values (Fig. 10). This was evident for both addition (690–740 and 1400–1450 ms) and subtraction tasks (700–790 ms). These findings are consistent with the notion of metric representations, as observed in the univariate analysis, indicating that highly differentiated cortical codes are easier to discriminate; a pattern strong in good performers but weak in poor performers.

Importantly, we found no evidence for representations of the difference subspace during the addition task, or the addition subspace during the subtraction task. Also, it was not possible to identify the category associated with the stimuli (see Section 4.2.4 on ”category”). This suggests that the task we designed is represented over continuous dimensions rather than two categorical categories.

Taken together our findings support the idea of task-relevant representations emerging in central and occipito-temporal sensors at about 600–900 ms post-stimulus.

5.1. *Emergence of abstract representations*

Low-dimensional maps, like the task-relevant representation we found, are thought to be abstracted from the sensory information and coded in a way that facilitates generalization and transfer learning across tasks. Luyckx and colleagues (2019), in a multitask experiment compared the EEG activity during a numerical decision task and a learning task. They found that numerical values, as outcome probabilities, are coded in the brain according to value similarity, starting from about 100 ms after stimulus onset. Interestingly, from about 300 ms to 650 ms, this code is abstracted and aligned between the two tasks to represent a shared concept of magnitude. As with our paper, they found faster (stimulus-bound) and slower (abstracted) processes. Similarly, Teichmann and colleagues (2018) found stimulus-bound and shared abstract processes in the encoding of numbers from 1 to 6 in two different formats. However, differently from the aforementioned studies, in our experiment, the faster and stimulus-bound process was shared between tasks, as the stimuli were the same. Instead, the slow process was task-relevant as the two tasks did not have a common structure. This difference motivates a future experiment where EEG activity could be recorded while participants are tested in two consecutive tasks, that do or do not have a common structure (see Menghi et al.

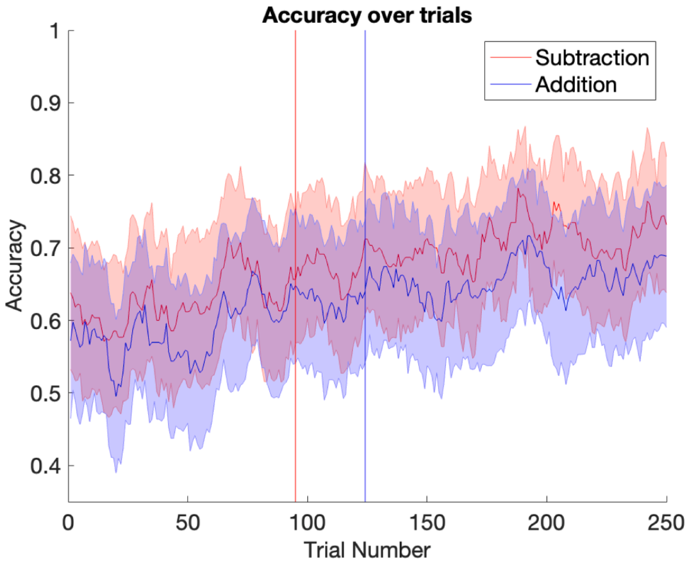


Fig. 11. Accuracy Over trials in Subtraction and Addition tasks. *Learning accuracy is computed with a moving average with a window of 10 trials. The error bars in the figure correspond to the standard error of the mean. Additionally, a line on the graph indicates the average point in time at which participants declared their respective strategies.*

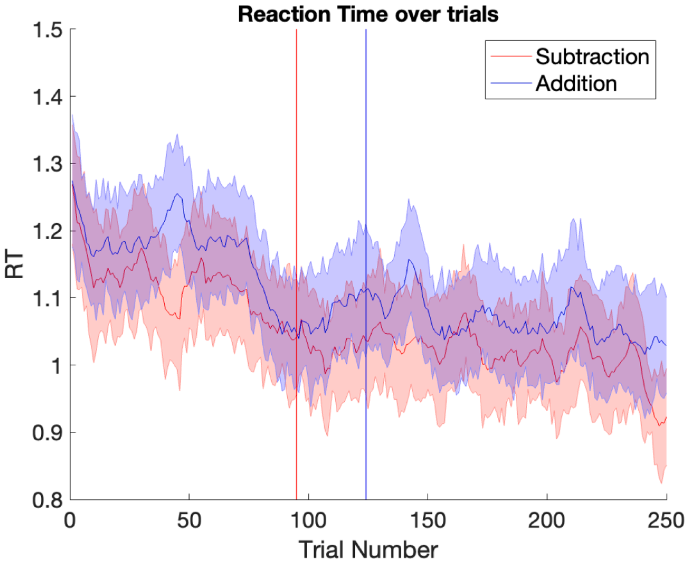


Fig. 12. Accuracy Over trials in Subtraction and Addition tasks. *Learning accuracy is computed with a moving average with a window of 10 trials. The error bars in the figure correspond to the standard error of the mean. Additionally, a line on the graph indicates the average point in time at which participants declared their respective strategies.*

2021). Such an experiment would show how the stimulus-bound and task-relevant processes we have identified might be differentially involved in transfer learning. Moreover, to fully understand the dynamics of representation emergence and transfer learning, future investigations could include a post-learning task that directly probes the stability and generalization of the acquired representation across new contexts. It would also be interesting to track the emergence of these representations during the learning process. Unfortunately, due to sample size limitations (where a single block does not contain enough repetitions of each stimulus configuration) this was not possible in the current experiment. Nevertheless, a highered powered experiment might help explain the weaker representation manifest for the addition task as being due to slower learning.

5.2. Category boundaries

Research has established that classification accuracy tends to diminish as data points move closer to category boundaries (Braunlich et al., 2017). Our behavioural data already evidenced how performance close to the boundaries tends to drop. Bad performers' classification rate was at the chance level in feature values corresponding to the boundaries. Bad performers appeared to rely on a simplified rule, centering their focus exclusively on learning the feature value 0 as opposed to feature values 1, 2, 3, and 4. This deviation from a comprehensive consideration of feature values underscores the challenges associated with representing and effectively leveraging the low-dimensional structure inherent to the task. This effect is corroborated by our

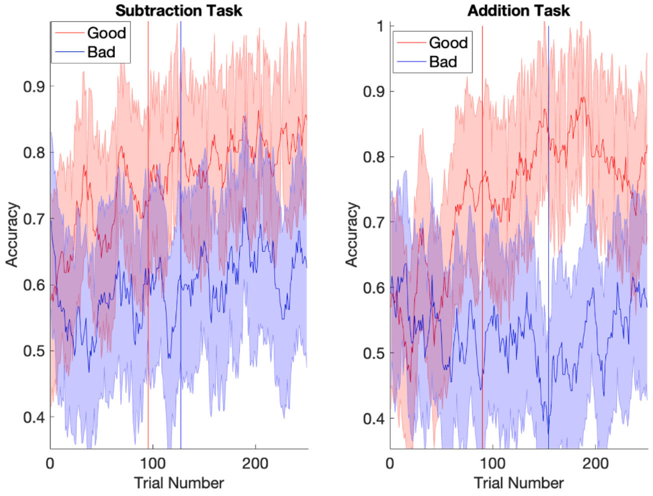


Fig. 13. Accuracy Over trial for Good and Bad performers in Subtraction and Addition tasks. *Learning accuracy is computed with a moving average with a window of 10 trials. The error bars in the figure correspond to the standard error of the mean. Additionally, a line on the graph indicates the average point in time at which participants declared their respective strategies.*

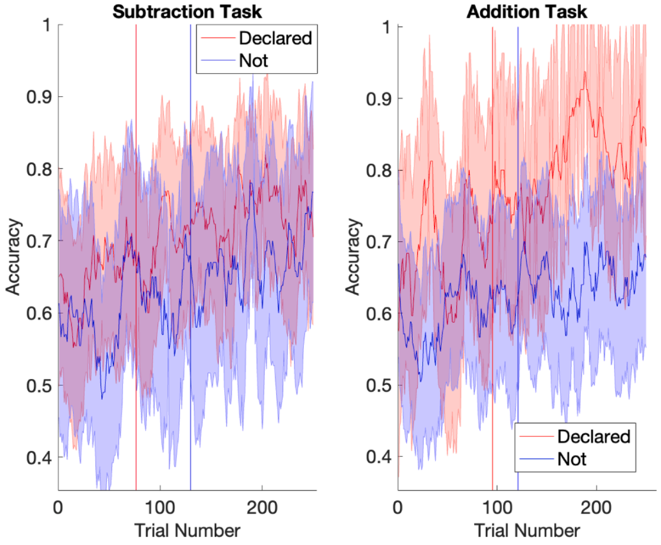


Fig. 14. Accuracy Over trials for Declarative and Non-declarative participants in Subtraction and Addition tasks. *Learning accuracy is computed with a moving average with a window of 10 trials. The error bars in the figure correspond to the standard error of the mean. Additionally, a line on the graph indicates the average point in time at which participants declared their respective strategies.*

behavioural data, as evidenced by the U-shaped patterns in Figs. 4 and 5 from our experiment. These patterns vividly illustrate the difficulty participants encounter in grasping and making use of the underlying task structure as they move away from category boundaries.

Data Availability

EEG data and analyses implemented in Matlab (Mathworks Inc) software are available from https://github.com/Nich0Me/EEG_Nonlinear_DM.

CRediT authorship contribution statement

N. Menghi: Conceptualization, Methodology/Study design, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review and editing, Visualization. **F. Silvestrin:** Investigation, Data curation. **L. Pascolini:** Investigation, Data curation. **W. Penny:** Conceptualization, Methodology/Study design, Software, Formal analysis, Writing – original draft, Writing – review and editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to my data/code at the end of my manuscript, before the references

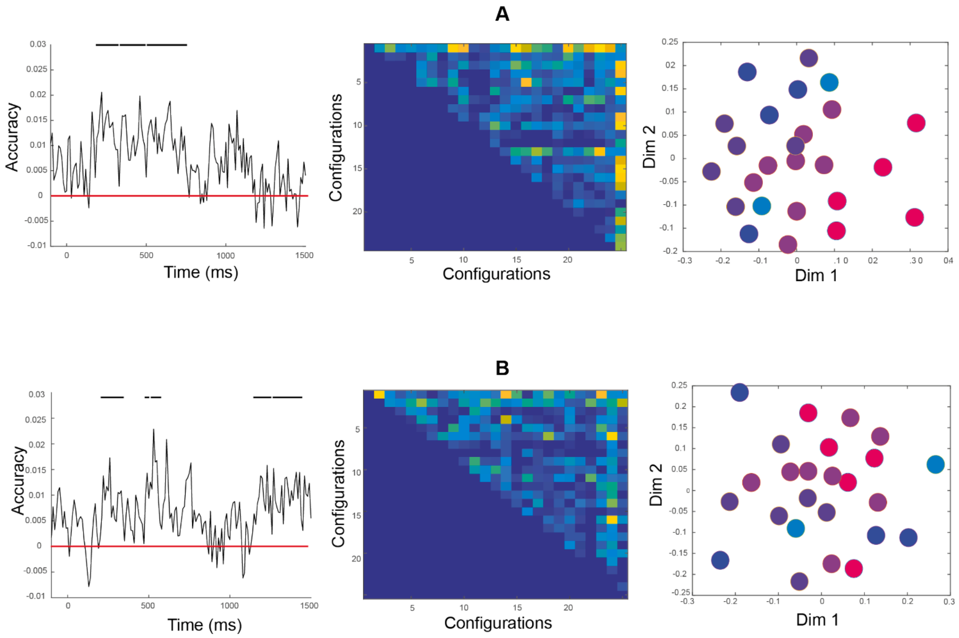


Fig. 15. Timecourse of decoding accuracy among configurations, the structure of decoding matrices and MDS spaces in (A) Subtraction and (B) Addition tasks. *The left panel illustrates the time course of overall decoding. The horizontal bars above represent the significant clusters. The* $[i,j]$*th entry in the EEG Decoding Matrices (central panels) correspond to the cross-validated accuracies with which stimulus configuration i and can be discriminated from configuration j (with yellow denoting highest accuracy). These accuracies have been averaged over time points containing significant effects (see left panels). The right panel illustrates the first two dimensions of the MDS in the EEG decoding matrix, according to the feature value (see Fig. 2).*

Appendix A. Supplementary Material

A.1. Summary of Behavioural Results

See Table 1–6.

A.2. Learning Curve

In this supplementary section, we provide additional insights into the learning accuracy across trials. Here we present a description of learning accuracy and reaction time across different trials. This figure shows accuracy over trials computed as a moving average over 10 trials. (see Figs. 11 and 12).

A.2.1. Good and bad performers

Here we divide participants based on their performance, providing additional insights on the learning pattern of these two groups. (see Fig. 13).

A.2.2. Declarative Knowledge

Here we divide participants into declarative and non-declarative based on their correct declaration of the task rule, providing additional insights on the learning pattern of these two groups. (see Fig. 14).

A.3. Multivariate EEG Results

A.3.1. Decoding of configurations

Fig. 15 shows the decoding accuracy averaged across all 25 stimulus configurations. Before and just after stimulus presentation, grand average decoding accuracy fluctuated around the chance level. In the subtraction task, classification reached significance at 190 ms (190–320 ms), followed by a cluster at 340 ms (340–490 ms) and the last cluster at 510 ms (510–750 ms). In the addition task classification reached significance at 210 ms (210–340 ms), followed by a cluster at 480 ms (480–570 ms), a cluster at 1150 ms (1150–1250 ms) and a last one at 1270 ms (1270–1440 ms). Thus, multi-variate analysis of EEG data revealed the temporal dynamics of the visual processing of the different configurations in the brain.

A.3.2. Multidimensional Scaling

Because it is difficult to directly make sense of the 25 × 25 × 161 EEG decoding matrix, we used multidimensional scaling (MDS) to project the data into a two-dimensional space of the first two dimensions of the solution, such that similar representation are grouped together and dissimilar ones far apart. MDS is a method to visualize the level of similarity of individual objects contained in a distance matrix (here the decoding matrix), whereby objects are automatically assigned coordinates in space so that distances between objects are preserved. For the purpose of MDS we averaged the EEG decoding matrix over those time points shown to be significant using the non-parametric permutation tests.

References

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences, 38*, 20–28.

Balestrieri, E., & Busch, N. A. (2022). Spontaneous alpha-band oscillations bias subjective contrast perception. *Journal of Neuroscience, 42*(25), 5058–5069.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? organizing knowledge for flexible behavior. *Neuron, 100*(2), 490–509.

Bellmund, J. L., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science, 362*(6415), eaat6766.

Benna, M. K., & Fusi, S. (2021). Place cells may simply be memory cells: Memory compression leads to spatial tuning and history dependence. *Proceedings of the National Academy of Sciences, 118*(51). e2018422118.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Braun, D. A., Mehring, C., & Wolpert, D. M. (2010). Structure learning in action. *Behavioural brain research, 206*(2), 157–165.

Braunlich, K., Liu, Z., & Seger, C. A. (2017). Occipitotemporal category representations are sensitive to abstract category boundaries defined by generalization demands. *Journal of Neuroscience, 37*(32), 7631–7642.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience, 17*(3), 455–462.

Comon, P. (1994). Independent component analysis, a new concept? *Signal processing, 36* (3), 287–314.

Constantinescu, A. O., O'Reilly, J. X., & Behrens, T. E. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science, 352*(6292), 1464–1468.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current opinion in neurobiology, 16*(6), 693–700.

Dobson, J., A., Barnett, G., A. (2018). An introduction to generalized linear models. Chapman and Hall.

Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in psychology, 2*, 243.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W. (Eds.). (2007). Statistical parametric mapping: The analysis of functional brain images. Academic Press.

Hubbard, J., Kikumoto, A., & Mayr, U. (2019). EEG decoding reveals the strength and temporal dynamics of goal-relevant representations. *Scientific reports, 9*(1), 1–11.

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in cognitive sciences, 18* (4), 203–210.

Knowlton, J., B., Squire, R., L., Gluck, A., M. (1994). Probabilistic classification learning in amnesia. Learning and Memory, 1, 106–120.

Kriegeskorte, N., Mur, M., Bandettini, P.A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 4.

Luyckx, F., Nili, H., Spitzer, B., & Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning. *Elife, 8*, e42816.

Manning, J. R., Polyn, S. M., Baltuch, G. H., Litt, B., & Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, 108*(31), 12893–12897.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods, 164*(1), 177–190.

MATLAB. (2018). *version 9.7.0.1190202* ((r2019b).). Natick, Massachusetts: The MathWorks Inc.

Menghi, N., Kacar, K., & Penny, W. (2021). Multitask learning over shared subspaces. *PLoS Computational Biology*.

Morton, N. W., Schlichting, M. L., & Preston, A. R. (2020). Representations of common event structure in medial temporal lobe and frontoparietal cortex support efficient inference. *Proceedings of the National Academy of Sciences, 117*(47), 29338–29345.

Niv, Y. (2019). Learning task-state representations. *Nature neuroscience, 22*(10), 1544–1553.

O'Keefe, J., & Speakman, A. (1987). Single unit activity in the rat hippocampus during a spatial memory task. *Experimental brain research, 68*(1), 1–27.

Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.-M. (2011). Fieldtrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Computational intelligence and neuroscience, 2011.

Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map making: constructing, combining, and inferring on abstract cognitive maps. *Neuron, 107*(6), 1226–1238.

Penny, W.D., Menghi, N., Renoult, L. (2022). Cluster-based inference for memory-based cognition. bioRxiv.

Radulescu, A., Shin, Y. S., & Niv, Y. (2021). Human representation learning. *Annual Review of Neuroscience, 44*(1), 253–273.

Samaha, J., Iemi, L., & Postle, B. (2017). Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Consciousness and Cognition, 54*, 47–55.

Sols, I., DuBrow, S., Davachi, L., & Fuentemilla, L. (2017). Event boundaries trigger rapid memory reinstatement of the prior events to promote their representation in long-term memory. *Current Biology, 27*(22), 3499–3504.

Teichmann, L., Grootswagers, T., Carlson, T., & Rich, A. N. (2018). Decoding digits and dice with magnetoencephalography: evidence for a shared representation of magnitude. *Journal of cognitive neuroscience, 30*(7), 999–1010.

Theves, S., Fernandez, G., & Doeller, C. F. (2019). The hippocampus encodes distances in multidimensional feature space. *Current Biology, 29*(7), 1226–1231.

Viganò, S., Rubino, V., Di Soccio, A., Buiatti, M., & Piazza, M. (2021). Grid-like and distance codes for representing word meaning in the human brain. *NeuroImage, 232*, 117876.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature neuroscience, 20*(6), 864–871.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron, 75*(1), 168–179.