Novel verbal instructions recruit abstract neural patterns of time-variable information dimensionality

Paula Pena, Ana F. Palenciano, Carlos González-García & María Ruz

Mind, Brain and Behavior Research Center, University of Granada, Granada, Spain

Correspondence: mruz@ugr.es

Competing Interests Statement: The authors declare no competing interests.

Abstract

Human performance is endowed by neural representations of information that is relevant for behavior, some of which are also activated in a preparatory fashion to optimize later execution. Most studies to date have focused on highly practiced actions, leaving largely unaddressed the re-configuration of neural information in novel settings, where unique task sets have to be generated from scratch. Using electroencephalography (EEG), this study investigated the dynamics of the content and geometry reflected on the neural patterns of control representations during novel instructed behavior. We designed a verbal instruction paradigm where each trial involved novel combinations of multi-component task information. By manipulating three task-relevant factors, we observed multiplexed coding throughout the trial, during both preparation and implementation stages. The temporal profiles were consistent with a hierarchical structure: whereas task information was active in a sustained manner, the coding of more concrete stimulus features was more transient. Data showed both high dimensionality and abstraction, particularly during instruction encoding and target processing. Our results suggest that whenever task content could be recovered from neural patterns of activity, there was evidence of abstract coding, with an underlying geometry that favored generalization. During target processing, where potential interference across factors increased, orthogonal configurations also appeared. Overall, our findings uncover the dynamic manner with which control representations operate during novel recombinatory unique scenarios, with changes in dimensionality and abstraction adjusting along processing stages.

Keywords: cognitive control, novel instructed behavior, task preparation and implementation, neural geometry

1. Introduction

Humans excel at following instructed commands, often guided by sentences that convey details about the relevance of pieces of information, and the rules associating them with the required actions. This ability, tightly linked to cognitive control mechanisms, is most useful in novel and changing scenarios. For example, we can play a new card game shortly after we receive the instructions. However, we know very little about how the human brain encodes novel task information to achieve such success. Theoretical models propose that top-down cognitive control mechanisms organize and maintain the relevant information by creating "task sets" (Miller & Cohen, 2001; Sakai, 2008), with empirical evidence linking these processes to the Multiple Demand Network (Duncan, 2010). Recently, the study of the neural geometry of activity patterns in these regions has gained relevance, examining whether control codes are structured to favor separability or generalization of information (Fusi et al., 2016). The neural mechanisms supporting the flexible generation of novel behaviors remain an elusive topic in the literature, however. The current study used complex novel verbal instructions, where diverse elements have to be recombined with a novel configuration each time (Cole et al., 2011; Ruge et al., 2019; Palenciano et al., 2019a), to examine the abstraction and generalizability of the neural patterns of information codes linking verbal commands and efficient responses.

Previous studies have shown that control-related information is anticipated when preparing to implement a task and later reactivated during task implementation (Grootswagers et al., 2018; Manelis & Reeder, 2015; Sakai & Passingham, 2003; Peñalver et al., 2023). Nevertheless, most of this evidence comes from repetitive and highly practiced tasks, which limits flexibility and constrains extrapolation to more novel contexts cornerstone of human cognitive control. The small body of research incorporating novel tasks has demonstrated that information provided by instructions can be decoded both during preparation (Muhle-Karbe et al., 2017; González-García, 2017) and task implementation (Hartstra et al., 2011; Palenciano et al., 2019a). However, these studies are based on fMRI recordings which, despite of providing rich anatomical information, are blind to the underlying finer-grained temporal dynamics. While few studies have used EEG to investigate instructed behavior (Formica et al., 2021; Formica et al., 2022), none so far has combined temporally resolved multivariate analyses to characterize the coding of the informational content of novel task instructions. Hence, how different streams of information available in complex instructions are structured in time is unknown.

A highly relevant descriptor of control representations is their underlying neural geometry (Badre et al., 2021) and their balance between neural flexibility and stability (Bernardi et al., 2020; Bhandari et al., 2024). On the one hand, control codes require high dimensionality, with

separable representations for all the task inputs and their combinations, while reducing interference (Rigotti et al., 2013; Fusi et al., 2016). On the other hand, control representations must be abstract, with compressed low-dimensional geometries that support generalization of task-relevant components to novel situations (Bhandari et al., 2024). Some studies have claimed that geometric arrangement depends on the nature of the task paradigm (Musslick & Cohen, 2021), whether it demands that information is shared between different task combinations, or there is a preference for reducing representational overlap (Kadohisa et al., 2023; Flesch et al., 2022). Given this dichotomy, two architectures have been described to portray task requirements in the representational space, a parallel geometry that favors generalization of information, and an orthogonal organization that reduces potential conflict (Stokes et al., 2020). Yet, no consistent geometry model for novel task coding has been reported to date.

Some have theorized that the representational space in novel settings should be high-dimensional, encoding all task input combinations in a separable manner, with such dimensionality decreasing with learning (Bhandari et al., 2024). While a few studies have supported this with monkey data (Wójcik et al., 2023) or artificial neural networks (Farrel et al., 2022), this perspective, agnostic to the task-relevant information, is not in line with the fast acquisition and learning efficacy of humans in novel contexts (Musslick & Cohen, 2021). An opposite line of reasoning proposes the existence of an initial low-dimensional coding scheme, in which generalization of pre-existing control representations can be re-used and transferred across contexts (Badre et al., 2021; Verbeke & Verguts, 2022; Cole et al., 2013). Similarly to the already mentioned caveat of research on decoding the content of novel instructions, no study to date has examined the neural geometry of novel tasks on humans. More specifically, how information dimensionality and abstraction of control representations unfold over time when encoding and implementing novel action plans is currently unknown.

In this study, we collected electroencephalographic (EEG) data of humans while they followed novel instructed complex verbal commands. The instructions manipulated task demands, upcoming target categories, and the relevant features used to respond. We addressed how the instructions' content organizes the neural space with Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008), and the coding of such content with time-resolved Multivariate Pattern Analysis (MVPA; Grootswagers et al., 2018). Temporal Generalization Analysis (King & Dehane, 2014) informed about the stability of control codes, whereas time-wise Cross-Classification Generalization Performance (GGCP) was used as an index of abstraction. We also analyzed the dimensionality of the neural codes across time (Bernardi et al., 2023) and lastly, used an innovative model-based RSA approach to inquire if the representational

structure in a low-dimensional manifold portrayed a geometrical arrangement favoring either generalization or interference reduction across task demands (Muhle-Karbe et al., 2023).

2. Results

Task paradigm and behavioral results

Forty participants completed a verbal instruction task with abstract rules that allowed the combinatorial generation of new statements for each trial. They had to respond to whether the instructed statement was fulfilled or not by a pair of subsequent target images (Figure 1A). Instructions were presented in four different screens of 200 ms each. The first two indicated whether the participant had to either *integrate* information, attending to both target stimuli, or *select* one while ignoring the other. They also indicated the stimuli that were to appear, which could be *animate* or *inanimate* targets. The third screen referred to the feature they had to pay attention to, either the *color* or the *shape* of the frame surrounding the images, whereas the fourth and last screen indicated the key to press in case the instruction was fulfilled. After a pre-target interval, a pair of images was presented (200 ms) and participants had to respond. Additional catch trials (12% on every block) were added to ensure that participants were actively maintaining all instructions, where participants had to detect (by pressing simultaneously “A” and “L”) when any of the target images were different from the ones instructed. They had to complete 16 blocks of the task, interspersed with 8 localizer blocks (not analyzed in this study), with all conditions fully counterbalanced within each block. Participants were able to perform the task with efficiency (Figure 1B). Regular trials had a mean Accuracy of 0.91 (SD = 0.05) and mean Reaction Times of 1.13 seconds (SD = 0.22). Catch trials also add a good overall performance, with a mean Accuracy of 0.780 (SD = 0.16; Reaction Times: M = 1.09 s, SD = 0.18), which confirms that participants were attending to all the information given by the instructions.

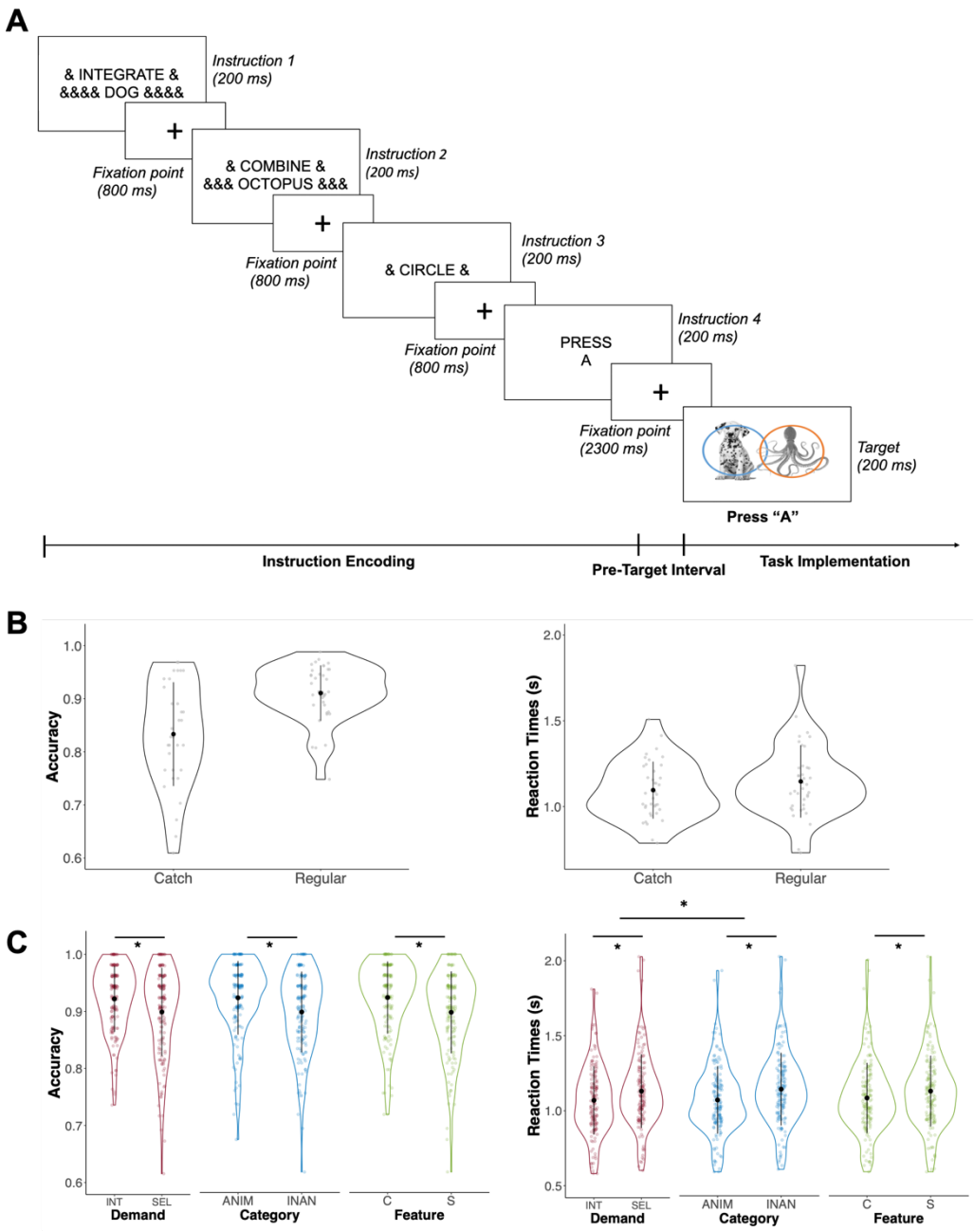

Figure 1. Task paradigm and behavioral results. (A) Example of a sequence of trial events. In this case, the sequential instructions indicate that if both the dog and octopus are framed by a circle, the key "A" has to be pressed. If this was an example of a catch trial, the target images presented could be dog and violin, instead of dog and octopus. The original instructions were presented in Spanish. **(B)** Mean Accuracy (left) and Reaction Times (right) of catch and regular trials. **(C)** Repeated-measures ANOVA results for Accuracy (left) and Reaction Times (right) per condition. Black dots refer to the mean of the condition, and the black error bars span one standard deviation above and below the mean. Colored dots depict the mean per condition of every participant. Asterisks indicate significant main effects of the ANOVAs ($p < 0.01$). Abbreviations: INT, Integration; SEL, Selection; ANIM, Animate; INAN, Inanimate; C, Color; S, Shape.

We explored whether the three main task components modulated behavior. The two behavioral measures (Accuracy and Reaction Times) in regular trials were analyzed separately with two within-subjects repeated-measures ANOVAs, including as factors Task Demand (integrate or

select information), Target Category (animate or inanimate), and Target Relevant Feature (attend to the color or shape surrounding the images). The results can be seen in Figure 1C. Task Demand had a significant main effect on both Accuracy ($F_{38,1}$ = 9.87; p < 0.01; η_p^2 = 0.21) and RT ($F_{38,1}$ = 30.51; p < 0.01; η_p^2 = 0.45), with more accurate (integration: M = 0.92, SD = 0.06; selection: M = 0.90 , SD = 0.08) and faster responses (integration: M = 1.07, SD = 0.23; selection: M = 1.13, SD = 0.24) for integration trials. Target Category also showed a significant main effect on Accuracy ($F_{38,1}$ = 25.91; p < 0.01; η_p^2 = 0.41) and RT ($F_{38,1}$ = 86.67; p < 0.01; η_p^2 = 0.70), with more accurate (animate: M = 0.92, SD = 0.07; inanimate: M = 0.90, SD = 0.07) and faster responses (animate: M = 1.06, SD = 0.23; inanimate: M = 1.14, SD = 0.24) for animate trials. Lastly, we also found a significant main effect of Target Relevant Feature on Accuracy ($F_{38,1}$ = 27.52; p < 0.01; η_p^2 = 0.42) and RT ($F_{38,1}$ = 70.32; p < 0.01; η_p^2 = 0.65), with more accurate (color: M = 0.92, SD = 0.06; shape: M = 0.90, SD = 0.07) and faster responses (color: M = 1.08, SD = 0.24; shape: M = 1.12, SD = 0.24) for color trials. Additionally, we found a significant interaction term between Task Demand and Category on Reaction Times ($F_{38,1}$ = 14.05; p < 0.01; η_p^2 = 0.27). Post hoc tests showed that the effect of Task Demand was significant for both Category conditions, with a larger effect for inanimate (t_{77} = -7.66; p < 0.01) than animate trials (t_{77} = -4.42; p < 0.01). All the remaining interactions in terms were not significant (p > 0.05). Thus, our findings reveal that the task components mostly had independent effects on behavior.

Novel task information organizes the underlying neural space at different temporal dynamics

Multivariate model-based Representational Similarity Analysis (RSA, Kriegeskorte et al., 2008) allowed to compare the neural data with the theoretical RDMs of the three instructed variables and examine how well each model reflects the neural coding space along the entire trial epoch. Neural Representational Dissimilarity Matrices (RDMs; Figure 2B) were created from a cross-validated Pearson's coefficient measure (1 - r), and later fitted into a multiple linear regression as dependent variables, with the theoretical RDMs of each of the three task components as regressors (Figure 2A). The time-resolved results showed that Task Demand, the higher-level task component, had a significant effect during most of the instruction stage while the remaining lower-level variables did so transiently and only in specific time windows, as displayed in Figure 2D.

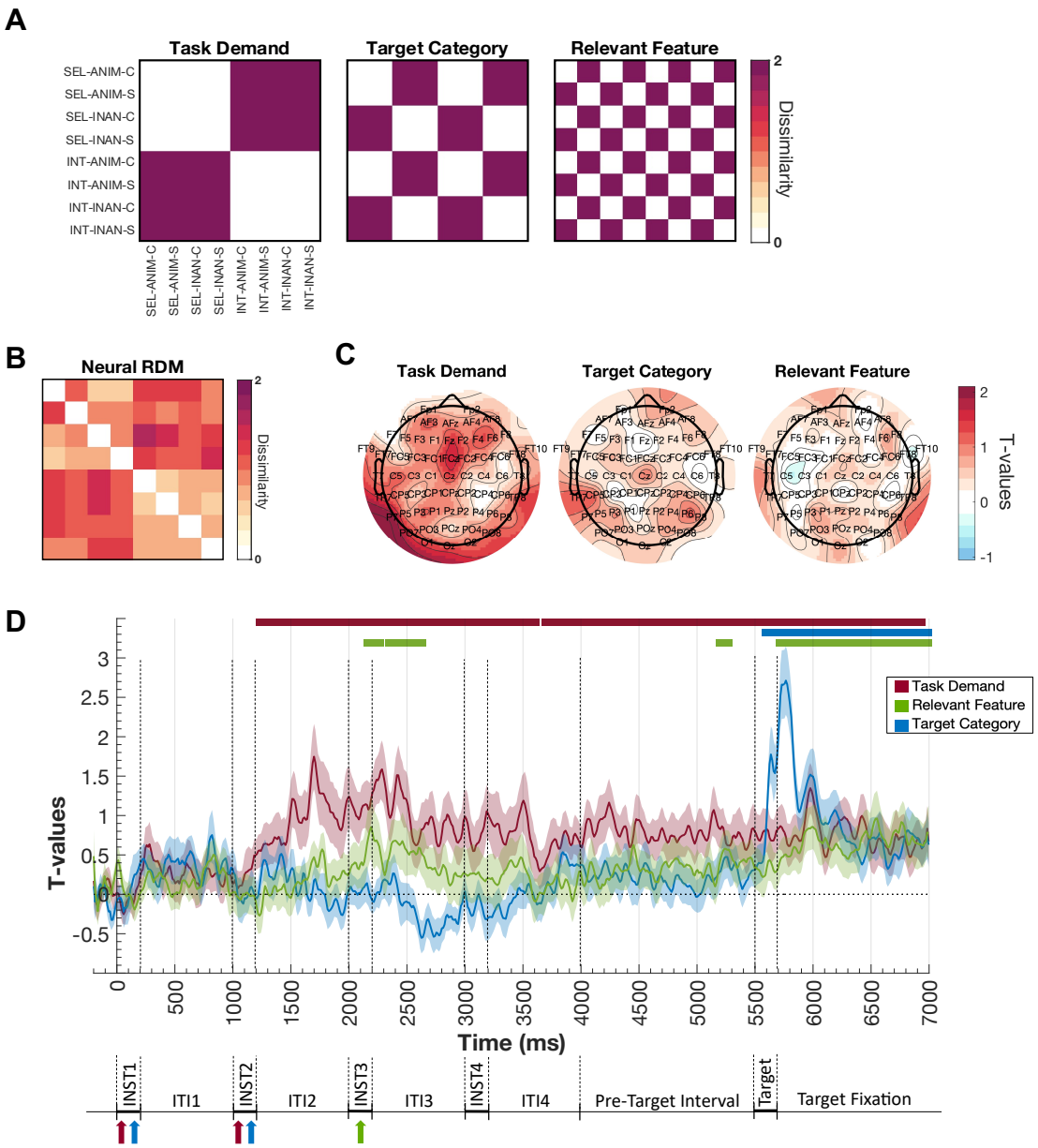

Figure 2. Task-relevant information organizes neural coding space: Representational Similarity Analysis (RSA). **(A)** 8x8 RDM of the theoretical models for the RSA of Task Demand, Target Category, and Target Relevant Feature. As the color bar illustrates, darker colors refer to maximal dissimilarity (2), and lighter colors to minimal dissimilarity (0). **(B)** Example of a neural RDM of one participant at one time-point. **(C)** RSA results per channel of the Task Demand, Target Category, and Target Relevant Feature model, representing the electrode's topography of each model for visualization purposes. We performed Representational Similarity Analysis separately for each channel, using the time points as features. The empirical matrices per channel were fitted into a linear regression with the theoretical models as regressors. Colors refer to different t-values, with red illustrating higher t-values. **(D)** RSA time-resolved results. Each darker-colored line represents the t-values after fitting the three theoretical models with a multiple linear regression, illustrating the unique share of variance explained by each model. Lighter shading refers to the standard error associated with the mean t-values. The horizontal-colored lines above refer to the statistical significance against zero (dotted line) after implementing cluster-based permutation analysis. The results are accompanied with a simplified illustration of the task paradigm, where colored arrows indicate in which instruction's screen the information of each task component is provided. Abbreviations: INT, Integration; SEL, Selection; ANIM, Animate; INAN, Inanimate; C, Color; S, Shape; A; INST, Instruction; ITI, Inter-instruction Interval.

Novel task information is encoded in the neural patterns with fluctuating temporal stability during preparation and execution

To study the coding of content of the task components in neural patterns, we used binary classifiers with a Linear Discriminant Analysis algorithm independently for every variable (Task Demand, Target Category, and Target Relevant Feature). In line with the results of the Representational Similarity Analysis (RSA), the Task Demand could be decoded during most of the epoch (Fig. 3A). First, two early peaks were significant during the processing of the first instruction (starting at 760 and 950 ms; colored arrows at the bottom of Fig. 3A). Next, a wide significant cluster extended during the remaining of instruction presentation (for a 3000 ms window), most of the pre-target interval, and target processing. The highest Area Under the curve (AUC) values for Task Demand were found during the processing of the second and third sections. Second, the Target Category classifier performance was significantly above chance in two shorter clusters during the instruction phase, right after the presentation of the second and third instruction screens (for 500 and 150 ms windows, respectively), which were those indicating the relevant category and feature. There was a small ramping-up of the classifier's precision right before the target onset (starting at 5250 ms) and the curve was steadily significant after target presentation (5540 ms onwards; Fig. 3A). Last, the classifier decoding of the Target Relevant Feature was significant right after the third instruction screen (for a 350 ms window; Fig. 3A), where the feature information was displayed. No significant decoding was detected during the pre-target interval or after target presentation.

The preceding findings do not address the degree to which neural coding patterns are stable over time. To further examine whether common or independent task representations are recruited across the preparatory and implementation stages, we used a temporal generalization approach (King & Dehane, 2014). The classifier was trained at one time point and tested on all the time points of the epoch, allowing the comparison of different patterns of brain activity throughout the entire trial (see Figure 3B).

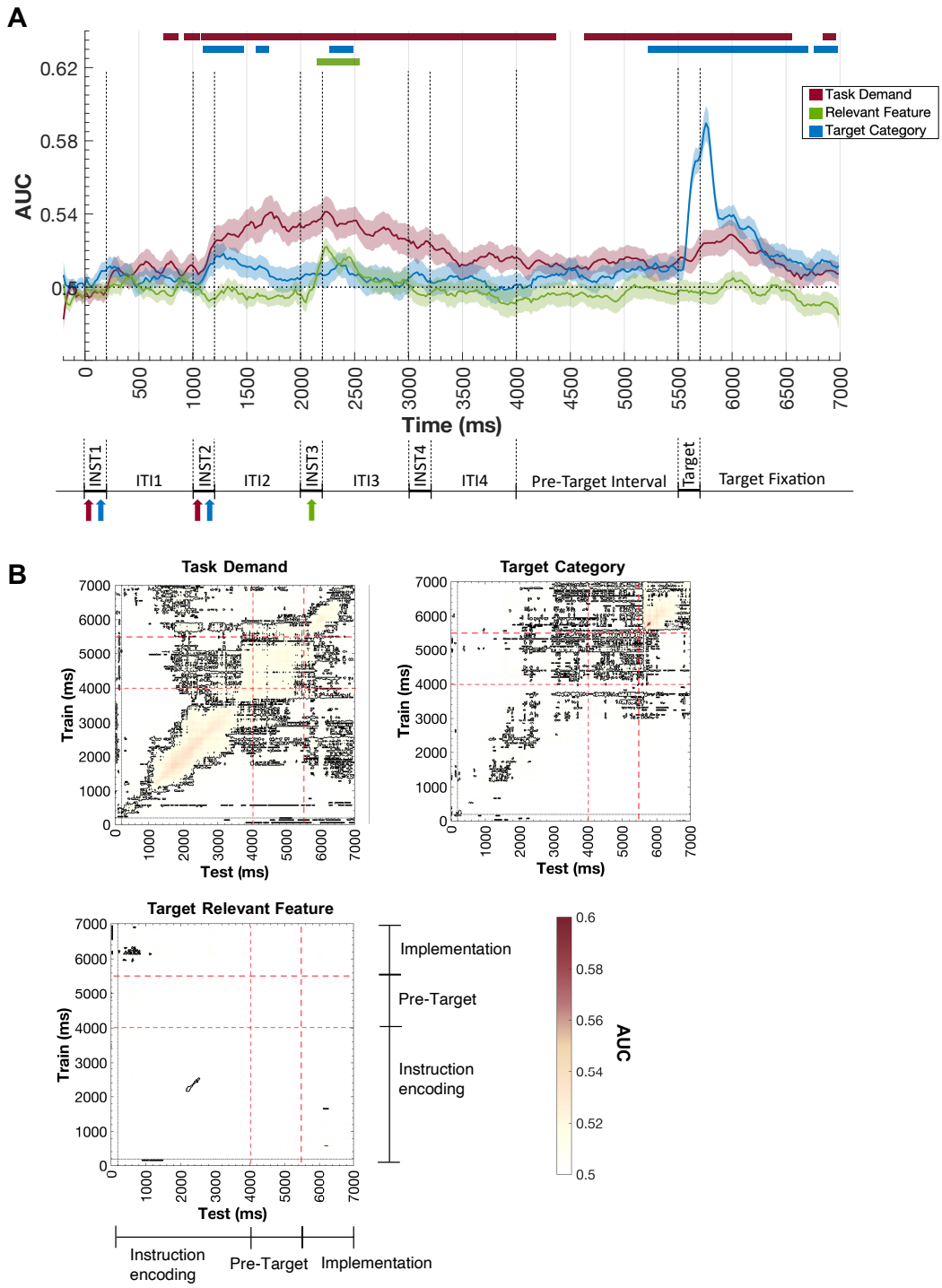

Figure 3. Raw Voltage MVPA Results **(A)** Results of the time-resolved classifications of the instruction-following task. Darker thin lines illustrate Area Under the Curve values (AUC), and lighter shading their variance. The horizontal-colored lines refer to the statistical significance against chance (dotted line). The figure shows the results of training and testing as described in *Time-resolved MVPA* (Methods section). The results are accompanied with a simplified illustration of the task paradigm. **(B)** Results for the temporal generalization analyses: Task Demand, Target Category, and Target Relevant Feature Temporal Generalization Matrices. The vertical axis indicates the training time bins, and the horizontal axis the testing time bins. The color bar indicates the AUC value of the classifier, and black lines denote significant clusters. Red dotted lines indicate the partitions between instruction encoding, pre-target interval, and task implementation. Abbreviations: INST, Instruction; ITI, Inter-instruction Interval.

First, regarding Task Demand (Fig. 3B upper-left) we observed off-diagonal significant classification, reflecting the sustained activity of the same patterns across time. Two connected clusters of significant generalization during the encoding of instructions (1000-2000 and 2000-3000 ms) suggest common representations within these two intervals, with activity patterns during the pre-target interval also displaying generalization (4000-5500 ms; middle quadrant of the matrix). During the task implementation phase, there was a small cluster of significant generalization after the target presentation (6000-7000 ms; upper-right quadrant of the matrix). The cross-classification across different task stages shows small significant clusters with generalized activation between the encoding of the last components of the instructions (2000-4000 ms) and both pre-target (middle-left and bottom-medium quadrants of the matrix) and task implementation (upper-left and bottom-right) intervals, similar cross-decoding was also found across the pre-target and implementation stages (upper-medium and middle-right). However, the temporal profile of these findings does not follow a sustained pattern reflecting a complete temporal persistence of neural codes during the whole task trial. Rather, this suggests a certain degree of stability of part of the neural patterns coding across the three task stages.

As for the analysis of Target Category (Fig. 3B upper-right), transient epochs of significant decoding were detected during the presentation and processing of the instructions; without generalization across the different instructions' screens. Several points of significant cross-classification were evident during the pre-target interval, but with no apparent unified cluster. This emerged after target onset (5500-7000 ms), indicating stable neural patterns during the task implementation period. Additionally, there were small clusters of significant cross-decoding when the classifier was trained during target processing (5500-7000 ms) and tested throughout instructions and pre-target intervals (upper-left and middle quadrants of the matrix). Besides, several significant points of cross-classification between the later instruction encoding and the pre-target interval indicated generalization between these two time periods. Overall, cross-classification results suggest that the animacy of the target category was coded with representations sustained to some extent. Last, the Target Relevant Feature matrix (Fig. 3B lower-left) only showed points of significant decoding on the diagonal during the encoding of the third instruction (2000-3000 ms).

Abstract neural codes of different task-relevant information emerge with sustained and transient dynamics

After ascertaining that neural patterns code task-relevant information conveyed by the novel instructions, we addressed whether these patterns were represented in an abstract format, generalizing across different contexts. For that purpose, we performed Cross-Condition

Generalization Performance (CCGP; Bernardi et al., 2020) for the three task variables in a time-resolved manner. We split the data into train and test sets for every possible combination of the conditions and performed linear classifiers, leaving 6 conditions for training (3 vs. 3) and 2 for testing (1 vs. 1) on every iteration, so that training and testing sets always consisted of different conditions. Figure 4B illustrates the CCGP value, the average performance of the 16 cross-classifications (see *Figures 4 - Figure Supplement 2*, *Figure 4 - Figure Supplement 3* and *Figure 4 - Figure Supplement 4* for the classifier performance per cross-classification).

CCGP findings illustrate different temporal profiles of abstraction for each of the task components. The variable Task Demand engaged common neural codes along most of the trial, since the mean Area Under the Curve (AUC) of the CCGP was significantly above chance for almost all the time-points of the epoch. There was a peak of cross-generalized decoding during instruction encoding and, although the AUC decreased on the pre-target interval, it was still significantly decoded above chance. Thus, Task Demand recruited in a stable fashion common neural codes that support generalization, corroborating that the higher-level variable of the task was represented abstractly in a prolonged fashion. In fact, this temporal profile of results seems to concur with those of the regular Time-Resolved MVPA, suggesting that abstracted neural patterns underpin the coding of this task component throughout the trial window. Note that the cross-validation and cross-decoding schemes followed in these two analyses do not allow for direct statistical comparison due to the different number of observations used for testing and training the classifiers, which could lead to spurious results. Instead, we aimed at characterizing their potentially common temporal profiles through Pearson correlations, by including the results of all time points at once. The average coefficient across participants after Fisher's Z-transform indicated a significant and strong relationship between the two time courses (r_{avg} = 0.84, combined p-value using Fisher's method, $p < 0.01$).

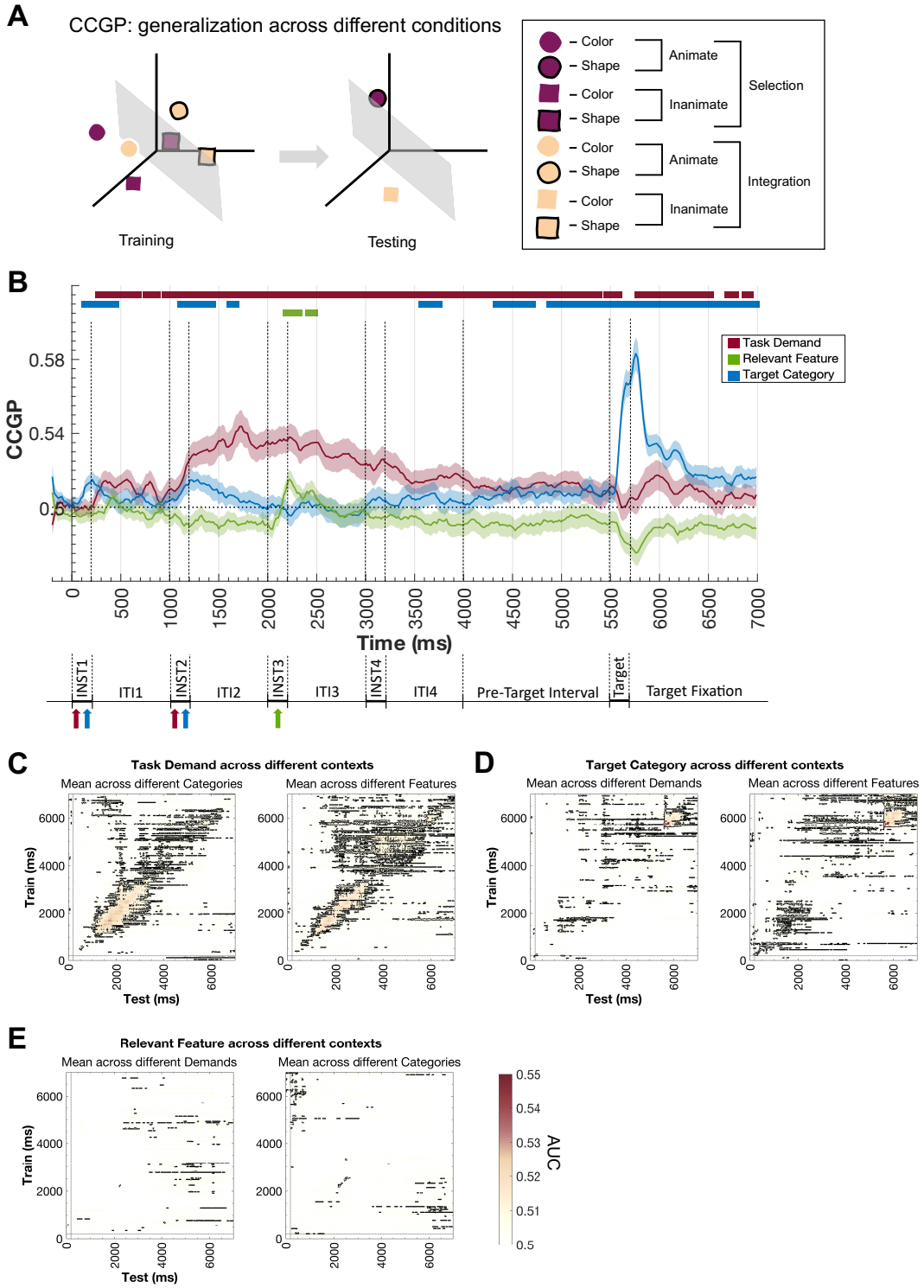

Figure 4. Temporal dynamics of abstract neural codes. **(A)** CCGP analysis rationale. Data were split according to the condition labels, so that training and testing sets always consisted of different conditions. This process was iterated for all possible splits of the data (see *Figure 4 - Figure Supplement 1* for full depiction*)*. **(B)** Cross-Classification Generalization Performance (CCGP) of the task components. The darker-colored lines represent the mean Area Under the Curve (AUC) after performing a linear classifier over all possible combinations of train and test data sets for each variable: Task Demand (magenta), Target Category (blue), Target Relevant Feature (green). Lighter shading refers to the variance associated with the performance measure. The horizontal-colored lines refer to the statistical significance against chance (dotted line) after implementing cluster-based permutation analysis. The vertical axis indicates the training time bins, and the horizontal axis the testing time bins. The color bar indicates

the AUC value of the classifier, and black lines denote significant clusters. **(C, D, E)** Results of Temporal Generalization analysis of the task components through different contexts with cross-classification (different conditions for train and test). **(C)** Mean Temporal Generalization Matrices after classifying integration vs. selection (Task Demand) through different Target Categories (left) and different Relevant Features (right). **(D)** Mean Temporal Generalization Matrices after classifying animate vs. inanimate (Target Category) through different Task Demands (left) and different Relevant Features (right). **(E)** Mean Temporal Generalization Matrix after classifying color vs. shape (Relevant Feature) through different Task Demands (left) and different Target Categories (right). See *Figure 4 - Figure Supplement 2, Figure 4 - Figure Supplement 3,* and *Figure 4 - Figure Supplement 4* for the MVPA results per dichotomy of Figure 4B. See *Figure 4 - Figure Supplement 5* for the correlation distributions across the pair-wise comparisons of regular MVPA and CCGP results. See *Figure 4 - Figure Supplement 6*, *Figure 4 - Figure Supplement 7*, and *Figure 4 - Figure Supplement 8* for the results per direction of the classification of Figure 4C, 4D, and 4E.

Target Category showed cross-generalization transiently during the earlier stages of the trial and steadier after target presentation, with a high CCGP peak right after target onset. The results suggest that Target Category was represented abstractly only at specific time intervals, which match the time periods where the content of this variable was decodable in the data. In this case, compared to the regular MVPA results, the average Pearson's correlation coefficient across participants after Fisher's Z-transform was r_{avg} = 0.83 (combined p-value using Fisher's method, $p < 0.01$). This suggests that whenever the Target Category was explicitly coded in the neural patterns, it was, at least partially, in a generalizable format. Lastly, the Target Relevant Feature showed a significant CCGP at a short interval during the encoding of the third instruction, when feature information was being provided, again matching the decoding results during this time period (see previous time-resolved MVPA results, average Pearson coefficient across participants was r_{avg} = 0.77, $p < 0.01$). Thus, this portrays abstract transient coding of feature information.

To ensure that the correlation results were robust and not due to the temporal structure of noise, we repeated the correlation of regular decoding and abstraction results but with different variables (e.g., MVPA results of Task Demand and CCGP results of Target Category), for a total of 6 pair-wise comparisons. The results showed negative, close to zero correlation coefficients, with the average Pearson coefficients ranging from r_{avg} = -0.10 to r_{avg} = -0.17 across the 6 comparisons (p-values range from p = 0.11 to 0 = 0.19; see *Figure 4 - Figure Supplement 5* for the distribution of correlations across all pair-wise comparisons). Thus, the strong correlations from MVPA and CCGP of the same variable (described in the previous paragraphs) were not merely a by-product of the temporal noise structure in the data.

To investigate whether these abstract neural codes were also temporally stable and transferable across the trial epoch, we performed a Temporal Generalization analysis with cross-classification. We obtained generalization matrices for each task variable when training

326 and testing on different conditions of the remaining variables. The analyses were performed in
327 both directions of classification (train on condition A and test on B, and vice-versa) and later
328 averaged (see *Figure 4 - Figure Supplement 6*, *Figure 4 - Figure Supplement 7* and *Figure 4 -*
329 *Figure Supplement 8* for the results per direction of classification).

330 For the Task Demand we observed similar generalization patterns both across different Target
331 Categories and Relevant Features (Figure 4C), with a unified cluster of generalization during
332 the processing of the second and third instructions' screens, and points of cross-classification
333 prior to and after target onset. The only difference of note was a thinner instruction encoding
334 cluster (fewer instances of cross-classification) across different Relevant Features (Fig. 4C
335 right). In comparison with the regular temporal generalization results (with the same data for
336 training and testing; see Temporal Generalization Analysis results), a smaller cluster of
337 generalizable neural patterns during target processing was detected, which suggests that the
338 Task Demand after the target appeared was coded with more independent neural patterns
339 depending on the Category and Feature of the trial's target. Together with the CCGP results,
340 the generalization matrices of Task Demand corroborate that this task component engaged
341 shared neural codes across different conditions, and that the abstract format of the variable
342 was maintained along the trial.

343 Results of cross-classifying the Target Category across Task Demands and Relevant Features
344 showed a resembling profile (Figure 4D). There was a cluster of generalization during target
345 processing, coherent with the previous results showing that the Target Category was strongly
346 coded in that time window (see RSA and Time-resolved MVPA results). Additional points of
347 cross-classification during instruction encoding and between target processing and both
348 instruction encoding and the pre-target interval were found. In accordance with what was
349 described previously, the neural patterns of Target Category reflected an abstract format
350 transiently during target processing, but not so evidently during the early stages of the trial.
351 Last, the Target Relevant Feature only exhibited segregated points of significant classification
352 that did not depict any recognizable generalization pattern (Figure 4E). As we described with
353 the CCGP results, this task variable was only explicitly coded in the data for a very brief time
354 period during the encoding of the third instruction screen.

355 ***The dimensionality of the neural codes shifts during the trial***

356 We examined the complexity of information build-up with instruction encoding, preparation and
357 task implementation across the whole trial. Shattering Dimensionality (SD) analysis was used
358 (Bernardi et al., 2020) by calculating the mean classifier performance of the 35 balanced
359 dichotomies obtained from mixing all task conditions. Figure 5B illustrates these results, where

two time intervals revealed higher SD values. First, during instruction encoding, before the onset of the second instruction's screen until the onset of the fourth, the overall separability of the neural codes increased. We observed similar results on a second instance, before target onset, and during its processing. During the pre-target interval, in between these two time periods of higher separability, the SD decreased drastically, going back to the initial low-dimensional state. Note that although the performance measure of the classifier had a generally low value in comparison to the previous time-resolved MVPA analysis, this is an expected result when considering the cross-classification approach implemented (Bernardi et al., 2020).

Since the Shattering Dimensionality represents the mean performance of classification between the dichotomies of the data set, it could be the case that some dichotomies with higher values of decoding are over-represented, skewing the results. To avoid this, we performed an additional measure of dimensionality that was based on the number of dichotomies decoded significantly above chance (Rigotti et al., 2013; Bernardi et al., 2020). As it can be seen in Figure 5C, the temporal dynamics were very similar to the original SD results. There was an increase in the number of decoded dichotomies during instruction encoding, which diminished during the pre-target interval, and had a later peak around target onset.

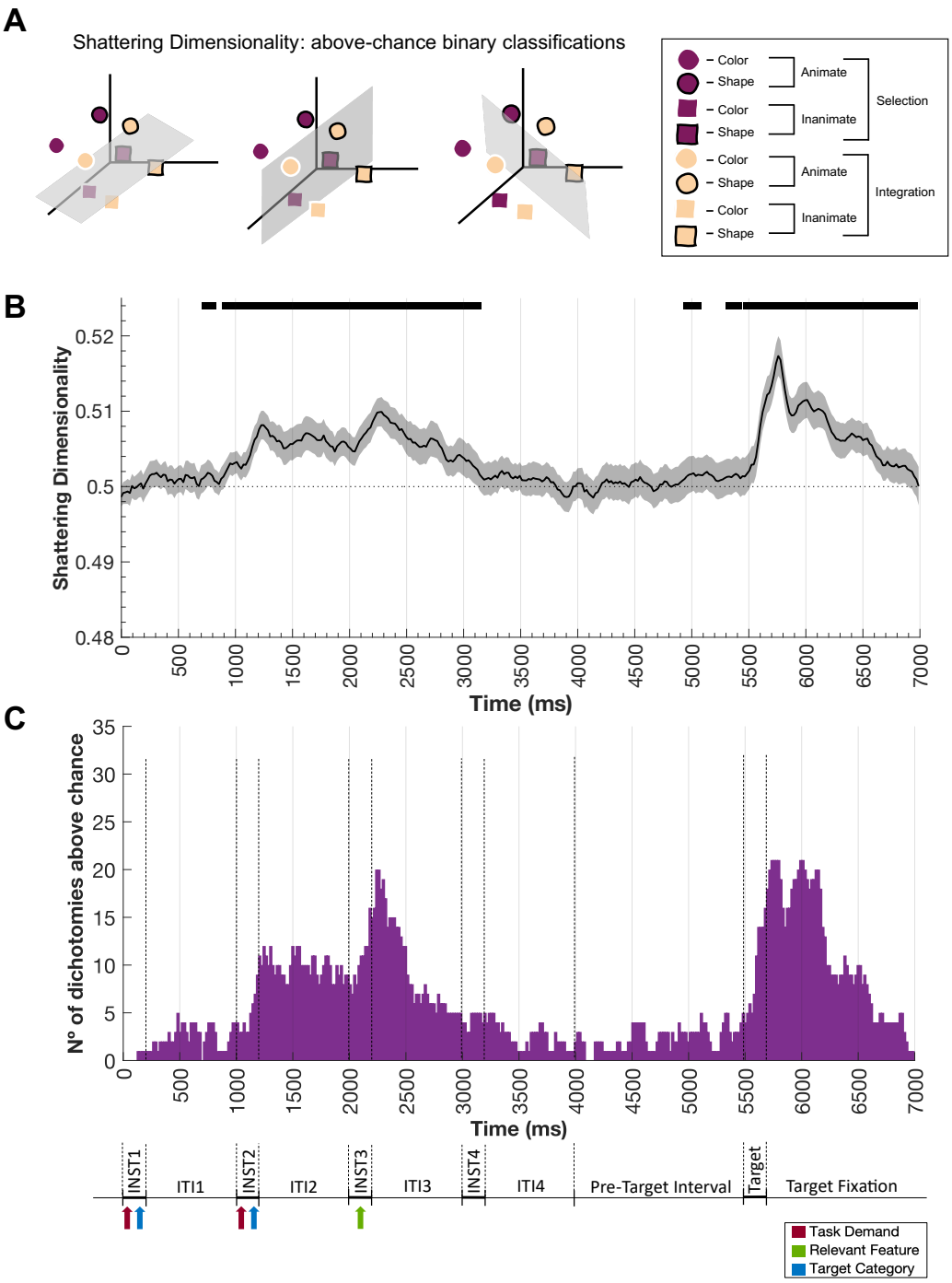

Figure 5. Temporal dynamics of neural dimensionality. **(A)** Analysis rationale. Data from the full-crossing of the variables were split into all possible balanced dichotomies (see *Figure 5 - Figure Supplement 1* for the full depiction of data splitting). After performing classifiers on every dichotomy, the more ways the data can be separated, the higher the dimensionality of the neural space. **(B)** Shattering Dimensionality (SD) results. The darker line represents the mean Area Under the Curve (AUC) after applying a linear classifier over all possible balanced dichotomies, obtained from mixing the task conditions. Lighter shading refers to the variance associated with the performance measure. The horizontal line refers to the statistical significance against chance (dotted line) after implementing cluster-based permutation analysis. See *Figure 5 - Figure Supplement 2* for the MVPA results per dichotomy. **(C)** Number of dichotomies decoded significantly above chance. A linear classifier was performed with each of the 35 balanced dichotomies obtained from mixing the task variables. For every time point, the significantly decoded dichotomies were summed. The results are accompanied with a simplified illustration of the task paradigm. Abbreviations: INST, Instruction; ITI, Inter-instruction Interval.

The neural representational format suggests a parallel geometry favoring generalization

Building on previous research reporting representational strength of task demands and contexts in the coding of instructions (Sobrado et al., 2022; Muhle-Karbe et al., 2017), we further evaluated the spatial configuration of the activity patterns coding the two Task Demands. For that purpose, we employed model-based RSA (Kriegeskorte et al., 2008) and compared two alternative theoretical geometries: one where the low-dimensional structure of integration and selection was organized in parallel planes (offset model; Figure 6A left), favoring generalization of information, and another conformed by planes laying at 90 degrees (orthogonal model; Figure 6A right), reducing interference of conflicting information.

The results indicated that the offset model explained the most variance for the majority of the trial (Figure 6D). This model started predicting the pairwise distances of the conditions very early on, for a brief period during the encoding of the first instruction's screen. Later, it had significant t-values again during the encoding of the second and third screens (for a 1300 ms window). During late encoding of the fourth screen, there was a slight increment of the t-values, with the significantly above-zero values maintained for the remainder of the trial. Prior to the presentation of the target, the results of the offset geometry exhibited a ramping-up pattern, peaking on target processing (approx. at 6000 ms). The orthogonal model, with integration and selection laying at a 90º angle, had significant results only after target presentation. Looking at the temporal profile of its t-values, the orthogonal model started with values near chance, then the model's explained variance diminished, with t-values below zero during the preparatory stage, instruction encoding and pre-target interval. After target onset, there was an increase in the model predictability, with t-values reaching significant above-zero rates. In summary, these findings suggest a geometry in the neural space in which integration and selection are represented separately, but with a configuration that allows the generalization of task information, with a shift towards orthogonality after target presentation.

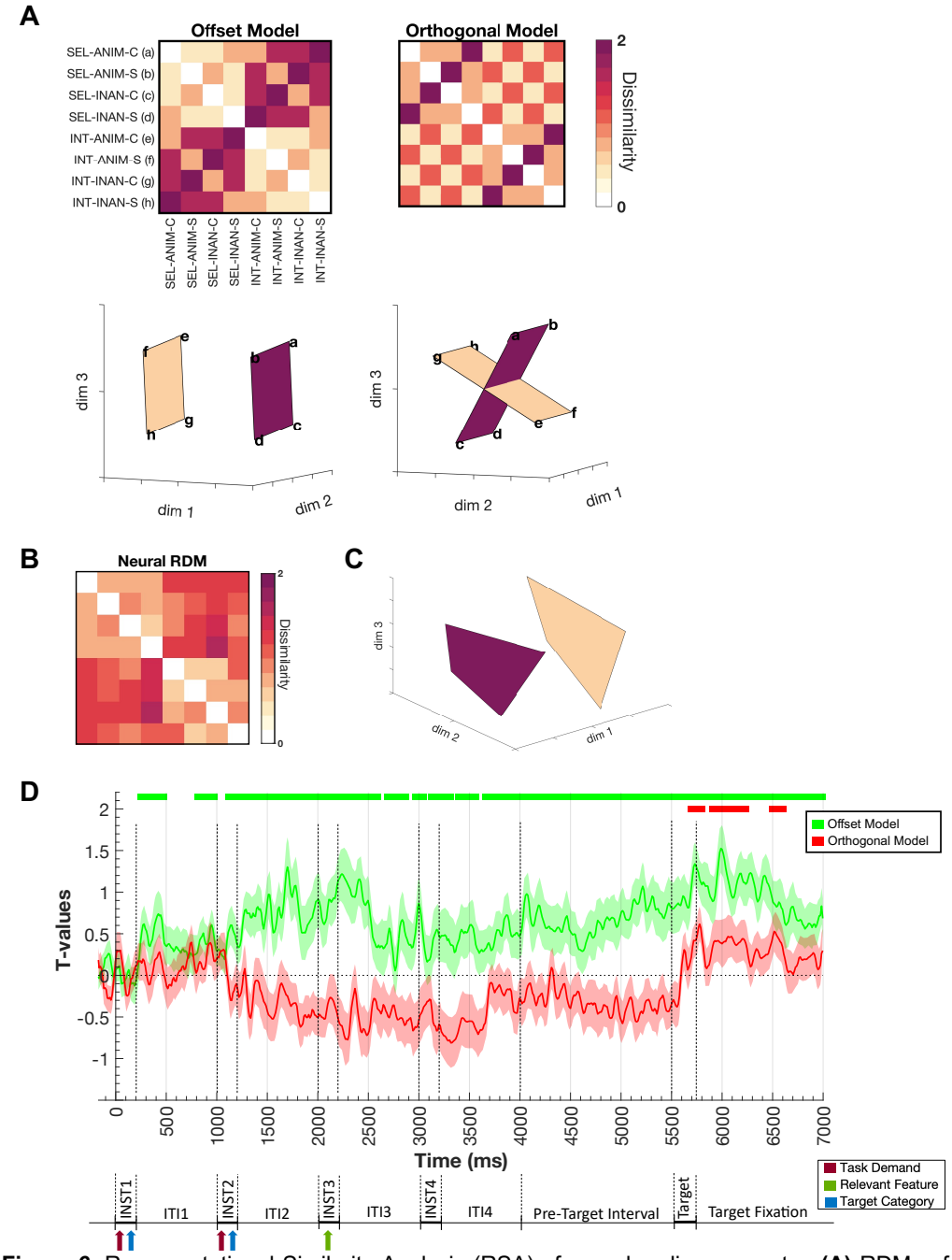

Figure 6. Representational Similarity Analysis (RSA) of neural coding geometry. **(A)** RDMs of the two geometry models and their Multidimensional Scaling plots (Offset and Orthogonal Model, respectively). As the color bar illustrates, darker colors refer to maximal dissimilarity (2), and lighter colors to minimal dissimilarity (0). The figures below depict the two alternative geometries, projected on a three-dimensional space. The darker purple plane refers to the selection activity patterns (with each vertex indicating a condition, e.g., a = selection - animate - color), and the beige plane to the integration activity patterns. **(B)** Example of a Neural RDM of one participant at a specific time point. **(C)** Multidimensional Scaling (MDS) plot of the previous neural RDM reduced to 3 dimensions. MDS was used for visualization purposes only. **(D)** RSA results per geometry model. Each darker-colored line represents the t-values after fitting the two geometry dissimilarity models to a multiple linear regression, illustrating the unique share of variance explained by each model. Lighter shading refers to the variance associated with the t-values. The horizontal-colored lines refer to the statistical significance against zero (dotted line) after implementing cluster-based permutation analysis. The results are accompanied with a simplified illustration of the task paradigm. Abbreviations: INT, Integration; SEL, Selection; ANIM, Animate; INAN, Inanimate; C, Color; S, Shape.

3. Discussion

This study investigated how multi-component novel verbal instructions organize neural patterns during task preparation and execution, in a temporally-resolved manner. By characterizing the temporal unfolding of the content and geometry reflected by activity patterns, our findings show how the neural codes shift in abstraction and dimensionality corresponding to the varying nature with which task-relevant components are coded by the human brain.

Our study is the first to characterize in time the neural coding of information underlying multi-component novel verbal instructions, since previous investigations decoding novel control representations employed mainly fMRI (Palenciano et al., 2019a; Muhle-Karbe et al., 2017; González-García et al., 2017). Time-resolved MPVA results revealed that neural patterns carried information about the content indicated by the verbal instructions during both task preparation and implementation. This fits with theoretical models that address the temporal nature of cognitive control, distinguishing *proactive* preparation for upcoming demands and *reactive* target implementation (Braver et al., 2007). We observed content-specific activity patterns during both instruction encoding and preparation, consistent with previous evidence of preparatory activity before target presentation (Palenciano et al., 2019a; Dumontheil et al., 2011; Muhle-Karbe et al., 2017; Hartstra et al., 2011; Hartstra et al., 2012; González-García et al., 2017; Formica et al., 2021; Formica et al., 2022), related to the active maintenance of goal-relevant information. Additionally, we also observed specific activity patterns after target presentation, associated with reactive control processes (Palenciano et al., 2019b; González-García et al., 2017). Results of the Temporal Generalization Analysis challenged the independence of these two control modes, as we observed significant transfer of the neural codes across task preparation and implementation. Two of the three variables manipulated, task demands and target animacy, showed stable underlying patterns that transferred through control epochs (Palenciano et al., 2019b; Ruge et al., 2019), stressing the intermixed nature of proactive and reactive control modes.

There were notable differences in the temporal profiles of the codes underlying different instruction components. MVPA findings revealed that the broader, more abstract variable of task demands was encoded earlier and maintained in a sustained way, while more specific components (target category and relevant feature) did so fleetingly. This divergence between variables was also portrayed in the representational space, when we directly addressed the share of variance uniquely attributable to each component. Results showed that task demand explained more variability, and during longer periods of time. Previous research has proposed hierarchical processing of task sets (Badre & Nee, 2018; Cole et al., 2011) with studies

showing stronger representations for higher-level variables (Woolgar et al., 2011a; Woolgar et al., 2011b; Cellier et al., 2022). Specifically, manipulating broader task demands or contexts induces highly separable representational schemes in both repetitive (Muhle-Karbe et al., 2017; Woolgar et al., 2011a; Sobrado et al., 2022; Stokes, 2013) and novel tasks (González-García et al., 2017; Bourguignon et al., 2018; Palenciano et al., 2019a). Extending these results, our findings further elucidate this structure in terms of lower-level variables. Since the relevant feature was tied to the stimulus-response mapping, it would be reasonable that its processing constituted one of the last steps of the task set hierarchy. The less sustained coding of color and shape information before target presentation in our task would be in line with this notion.

We further characterized control representations investigating their neural geometry through abstraction and dimensionality. Cross-Condition Generalization Performance results revealed that all the instruction components were at some point coded in an abstract format, generalizing across different training and testing conditions. Generally, when the content of each instruction component was decoded, there was evidence of abstract coding. Thus, the information of lower-level variables was portrayed in an abstract manner whenever it appeared during critical time intervals. Previous fMRI studies have reported sustained abstract task representations during planning (Vaidya et al., 2021) and execution (Bhandari et al., 2024), with similar results from Deep Neural Networks in an instruction task (Riveland & Pouget, 2024). Among our variables, task demand showed more stable CCGP, results corroborated by the cross-classified Temporal Generalization analysis. Given the potential interference among target elements, which could be linked to differential responses, the neural coding of the task set could minimize representational overlap but also facilitate the sharing of information across task combinations. To address the geometry that underlies these competing task constraints we employed RSA. Our results suggest that the neural space was broadly organized by an offset geometry, with integration and selection configured in separate, parallel planes, supporting generalization. Only after target presentation, when a completely parallel representation could impair performance, did we find evidence of orthogonal configuration, which could help to minimize the potential interference among competing elements. Thus, although the verbal instructions appeared to induce a neural format that facilitated transfer of information across task demands, there was a certain degree of constraint on this configuration during the implementation stage. In this regard, recent experimental evidence has reported alignment of different task contexts in the neural state space to facilitate generalization of conceptual information (Sheahan et al., 2021). To further elucidate this context-dependent organization, the geometry of task contexts and its theoretical implications should remain a focus of research.

Regarding the temporal dynamics of the complexity of information coding, we identified two main intervals with high Shattering Dimensionality. In these, more information dichotomies were separable, indicating higher-dimensional neural codes. Neural dimensionality was not stable during the trial; instead, this dynamically changed along trial stages. When information had already been instructed, neural patterns became less separable. This compression in dimensionality could support the generalization of the main instruction-relevant components in a preparatory fashion. However, considering the concurring reduction of decoding performance, an alternative explanation may be that information was not explicitly maintained during that time interval. Upon target presentation, neural coding recovered a high-dimensional state, decreasing potential interference among components. Studies employing repetitive tasks describe rapid dimensionality expansion after stimuli presentation, with neural codes morphing to create higher-dimensional states (Parthasarathy et al., 2017; Kikumoto et al., 2020; Kikumoto et al 2023; Aoi et al., 2020). Our results reveal how this information dimensionality dynamically changes across different stages of novel instruction encoding and implementation, with an expansion-compression interplay that may serve as a control mechanism to solve the “curse of dimensionality” (Bellman, 1957) while minimizing interference.

The overall pattern of dimensionality results informs the debate on the neural geometry of novel tasks. While some authors suggest that dimensionality reduction should take place over learning (Farrel et al., 2022; Wójcik et al., 2023; Bhandari et al., 2024), others argue that novel settings may start with less dimensions and abstract codes (Badre et al., 2021; Verbeke & Verguts, 2022; Cole et al., 2013). In contrast, we observed an interesting pattern featuring high dimensionality alongside generalization across conditions. Discrepancies may stem from the nature of the tasks used in most dimensionality studies, involving slower trial-and-error learning (Wójcik et al., 202; Flesch et al., 2022), while verbally instructed tasks can generate complex task sets very early on. Additionally, it has been shown that task settings with a similar structure where information can be reused, or with high task complexity, favor generalization (Aoi et al., 2020, Flesch et al., 2022; Badre et al., 2021; Woolgar et al., 2015). The instructions employed in the current experiment maintained a consistent structure and were highly demanding. For the case of novel verbal instructions, our results do not support neither an initial high (Wójcik et al., 2023) nor low dimensional state (Verbeke & Verguts, 2022), but a more complex representational landscape for novel scenarios. Further research will be needed to better characterize this interesting temporal dimensionality dynamic profile during human instructed performance.

Despite its contributions, our study is not without limitations. The adaption of the paradigm to an EEG setting required splitting instruction encoding stages into sequential words, while most

previous studies present novel instructions all at once (González-García, 2017; Palenciano et al., 2019a), as it also happens in more natural environments. This arrangement may have had an impact on the structure underlying the neural codes observed, restricting comparisons with previous results. Thus, future studies may take this into consideration for further specifications. Regarding the nature of the novelty of our paradigm, while the whole instructions were different and new on every trial, the overarching task structure repeated across the session, and the specific elements were recursively reused, as it is commonly done in this type of studies (Cole et al., 2011; Ruge et al., 2019). On this note, practice probably plays a role in the paradigm. The nature of our task could potentially have driven higher-order learning, which may have affected the coding scheme observed. For example, Ruge et al. (2019) found significant changes in task representations between early and late trials of a novel instructed task, and decreases in representational strength of task components due to familiarity have been reported (Woolgar et al., 2011a; Woolgar et al., 2011b), accompanied by a strengthening of abstract higher-level task conjunctions (Kikumoto et al., 2024). Nonetheless, our approach was intended to offer the opportunity to study not only how novel task sets are configured in single trials, but also how the brain compositionally adds sub-components to build a complex task set over time and how these trial-by-trial changes interact with its overall structure. Finally, using EEG limits conclusions on the spatial source of the brain activations. Previous studies have shown the involvement of frontoparietal areas of the Multiple Demand Network (Duncan, 2010) in cognitive control processes during novel tasks (Muhle-Karbe et al., 2017; González-García et al., 2017; Palenciano et al., 2019a). Fusion analysis of EEG and fMRI data (Cichy & Oliva, 2020) would help pinpoint the brain regions involved in the temporal dynamics described above.

In conclusion, the present study sheds light on how information is selectively represented and flexibly reconfigured to meet demands instructed verbally. Results show how instruction contents are coded with an abstract format that allows generalization across contexts, using at the same time high-dimensional patterns that vary depending on stages of information processing. Further studies should extrapolate these findings to novel scenarios with an overall different task structure, and complement them with more spatially precise neuroimaging techniques.

4. Methods

4.1 Participants

Forty participants (mean age = 21.85, range: 18-27, SD = 2.28; 26 women and 14 men) received economic compensation for taking part in the experiment, varying from 30 to 35 euros,

according to their average performance on the task. They were all native Spanish speakers, right-handed, with normal or corrected vision and no history of neurological issues. Participants that were not able to complete the practice session to criterion (see below) received a compensation of 5 euros. Additionally, the data of one participant had to be excluded due to low quality of the EEG recording, resulting in a final sample size of 39. All participants signed a consent form approved by the local Ethics Committee (reference 1584/CEIH/2020). The experiment was carried out at the Mind, Brain and Behavior Research Center (CIMCYC) of the University of Granada.

The sample size was selected based on a previous fMRI study where a similar experimental paradigm was employed. We compared our targeted sample size against the one obtained with a power analysis focused on a behavioral effect size. We used the software PANGEA (Westfall, 2015) to detect a modest effect size (Cohen's d = 0.3) of the variable Task Demand in the behavioral data (Reaction Times and Accuracy). Our sample size of 40 would achieve a statistical power of 90,5%.

4.2 Experimental Design

The experiment consisted of a main instruction-following task with a 2x2x2 within-subject design, where the independent variables were manipulated according to different hierarchical levels, with Task Demand (integration vs. selection) setting the goal of the task, and Target Category (animate vs. inanimate) and Target Relevant Feature (color vs. shape) as lower-level variables. The dependent variables were behavioral (Reaction Times and Accuracy) and electrophysiological (Voltage Values). Additionally, we ran a localizer task with a 4x2 within-subject design, manipulating the Target Subcategory of the stimuli (sea animal, land animal, musical instrument or tool) and Trial Type (repeat or non-repeat trials). This task was implemented for other analyses, not included in the current article, and thus will no further be discussed.

4.3 Apparatus and Stimuli

We generated 512 novel verbal instructions by fully crossing the three independent variables. Each instruction consisted of an "if… then" statement, indicating a condition about the two upcoming targets, together with the required response in case the condition was fulfilled. After the instructions, participants were shown pairs of targets surrounded by frames, which were generated by combining 8 images, 4 colors and 4 shapes. Images were taken from a total pool of 16 and were either animate (land animals: horse, bear, dog or cat and sea animals: whale, shark, octopus or jellyfish) or inanimate (tools: screwdriver, saw, hammer or drill and instruments: flute, saxophone, violin or guitar) objects. All stimulus images were retrieved from

the Google image search engine using the CCBY 4.0 filter. Half of the images were presented during the practice session (the first two of each subcategory mentioned), while the remaining were used for the experiment. Besides, each image was framed with a colored (green, pink, orange or blue) shape (triangle, circle, rhomboid or square).

The instructions gave information about the specific features of either one or the two target stimuli that had to be selected or integrated to respond. Task Demand specified whether the condition of the statement applied to both target stimuli (integration), or whether one of them had to be selected while ignoring the other (selection). Regarding the variable of Target Category, the instructions pointed to the category of the upcoming stimuli, either animate or inanimate. The variable of Target Relevant Feature indicated the specific features of the stimulus frames to pay attention to, either color (pink, green, blue, orange) or shape (square, rhomboid, circle, triangle). Finally, the instructions indicated the motor responses of either a left or right index button press. Half of the instruction sets referred to integration statements, whereas the remaining referred to selection statements; they were also equivalent in the rest of the parameters described (animate-inanimate stimuli; color-shape features), see Figure 1 for an example of the paradigm.

To minimize horizontal eye movements while reading during EEG recording, the instructions were sequentially presented. The first and second screens indicated if the information from both stimuli had to be integrated ("Integra" and "Combina", respectively, which translate to "Integrate" and "Combine"; ~ 20º x 3.5º); or if the information of only one stimulus had to be selected ("Atiende" and "Ignora", respectively, which translate to "Attend" and "Ignore"; ~ 20º x 3.5º). Note that to match the presentation structure across task conditions, each Task Demand condition uses two different labels. They also included the two concrete stimuli that the instruction applied to (belonging to either animate or inanimate categories; ~ 20º x 3.5º). The third screen indicated the specific feature to focus on, either a color ("Rosa", "Verde", "Naranja" or "Azul"; pink, green, orange, or blue, respectively; ~ 20º x 3.5º), or a shape ("Cuadrado", "Círculo", "Triángulo" or "Rombo"; square, circle, triangle, or rhomboid, respectively; ~ 20º x 3.5º). The fourth and last screen specified the motor response associated with the statement ("Pulsa A" or "Pulsa L"; which translates to "Press A" and "Press L"; ~ 9º x 3.5º). For example, an instruction for integration would be "Integrate dog, combine octopus, circle, press A"; here, the participant would have to pay attention to both the dog and the octopus and if they were framed by a circle, they should press A (otherwise, press L). An example of a selection instruction would be "Attend drill, ignore guitar, orange, press L" (which would indicate to "attend to the drill object, ignore the guitar. If the drill is surrounded by an orange shape, press L, otherwise press A"). To ensure that on all trials, each possible

instruction had the same size on screen (first, second and third instructions, separately), meaningless symbols were added ('&') to match the number of characters (see Figure 1A).

The experiment presentation and behavioral data collection were done with Psychtoolbox running on Matlab on a Microsoft PC screen of 61 x 34 cm. Participants were seated at approximately 65 cm from the screen.

4.4 Procedure

Once participants arrived at the lab, they received overall instructions and performed a practice session. Only participants obtaining a minimum of 80% behavioral accuracy, for a maximum of 8 blocks, were invited to the EEG session (80% of participants reached this criterion). Practice sessions lasted from 10 to 40 minutes. Afterwards, we set up the EEG cap and participants performed the experiment that lasted approximately 1 hour and 40 minutes. Instruction-following and localizer blocks were interspersed throughout the session. The main task blocks were further differentiated according to the particular stimulus features that participants responded to (e.g., orange and blue frames in half of the blocks and pink and green frames in the other half). The whole experiment had a total of 24 blocks with self-pacing rest times between them. Of these, 16 blocks were of the main task and 8 of the localizer. To control for the effect of block order, the transitions between block types (main task with features 1, main task with features 2, and localizer task) were counterbalanced within participants, so that all transitions were equally probable. In total, the whole experimental session lasted approximately 3 hours.

Figure 1 displays a schematic representation of a trial of the main task. Each one of the four screens was on display for 200 ms and was followed by a fixation point for 800 ms. After the 4 screens (with their respective succeeding fixation intervals), participants had an additional 1500 ms pre-target interval. Then, the targets were shown for 200 ms (~ 16º x 10.5º), with participants being instructed to respond as accurately and fast as possible. Participants had a 2800 ms interval afterwards to respond. This way, trials were divided into three phases: instruction encoding, pre-target and task implementation.

Each of the 16 blocks of the task contained 32 trials, giving a total of 512. For each block, there were 4 observations per each of the eight experimental conditions, resulting from the full crossing of the three independent variables. Additionally, for each block, we balanced the experimental conditions against a series of control variables that enabled creating a rich instructions pool, such as the target identity being integrated or selected (e.g., cat, drill…), the congruency of the target in regards to the instruction (both fulfilled it, both did not fulfill it, or only one fulfilled it), the specific frame features that participants responded to (e.g., green,

square…), the response indicated in the instruction (e.g., press A, press L), the response required by the targets, and the font and format of the different instruction cues. That approach ensured no statistical dependencies among the experimental conditions included in the analysis and additional task manipulations of non-interest.

Additionally, to avoid that participants encoded the information of the instruction screens partially (e.g., processing only one of the instructed target categories), we included 12,5% of *catch* trials where a novel, non-instructed image was shown as target (4 trials per block). Participants were told to detect those trials by pressing both keys simultaneously ("A" and "L").

4.5 Behavioral Analyses

A 2x2x2 within-subjects repeated-measures ANOVA was performed, with the three independent variables of the task as factors: Task Demand (integration, selection), Target Category (animate, inanimate), and Target Relevant Feature (color, shape). This analysis was performed independently for behavioral Accuracy and Reaction Times using R on Rstudio (RDC Team, 2016). To investigate significant interaction terms, we conducted additional post hoc and applied a Holm correction to account for multiple comparisons.

4.6 Electrophysiological Analyses

EEG Data Acquisition and Preprocessing

Electrophysiological (EEG) activity was collected with BrainVision's actiCAP equipment with 64 active electrodes. Of these, one was set as a reference channel (FCz) and two were used as electro-ocular channels (TP9, TP10). The electrodes' impedance was maintained below 10 kΩ, as the amplifier's manufacturers indicate. The EEG signal was recorded at a 1000 Hz sampling rate, and data files were structured following EEG-BIDS (Pernet et al., 2019).

EEG was preprocessed with the EEGLAB toolbox in Matlab (Delorme & Makeig, 2004) and custom Matlab scripts (López-García et al., 2022; available at https://github.com/Human-Neuroscience/eeg-preprocessing). EEG recordings were down-sampled at 256Hz, filtered with a low-pass at 126 Hz, a high-pass at 0.1 Hz, and additional notch to remove the effect of the electrical current and its harmonics (49-51 Hz and 99-101 Hz). We generated epochs including the whole trial, covering a time window of 7200 ms (from -200 ms to 7000 ms). Epochs were locked at the onset of the first instruction and spanned until the end of the inter-trial interval after target presentation. After visual inspection, noisy channels were not included for the following steps (only one channel from one participant was removed). An Independent Component Analysis (ICA) was used to remove ocular (blink and saccades) and muscular artifacts identified through visual inspection and ICLabel (Pion-Tonachini et al., 2019). A mean

of 2.87 components were removed per participant, ranging between 1 and 5. Then, an automatic rejection process was performed by removing trials with non-stereotypical artifacts that were not excluded previously with ICA. Reasons for exclusion were: 1) extreme values of voltage, in which the amplitude exceeded a range of ±150 μV; 2) abnormal spectra, in which the spectrum differed significantly from the baseline in the 0-2 Hz and 20-40 Hz frequency windows, linked to linear drifts in the signal caused by artifacts; 3) improbable data, with voltage values more than 6 standard deviations from the mean probability distribution. Noisy channels that contaminated the signal were removed and interpolated. The average reference was computed by pooling all the channels and then re-referencing to the mean. Finally, epoched data were baseline-corrected (-199 to 0 ms). Only correct and non-catch trials were included in the analyses, with an average of 44.60 ± 5.80 trials per condition (e.g., integration-animate-shape) and participant.

Representational Similarity Analysis (RSA)

The extent to which activity patterns were structured according to the instructed task components was evaluated through multivariate Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008). By calculating the pairwise distance across all conditions, this analysis enables the abstraction of the neural data into a common representational space. A time-resolved RSA was performed to obtain a fine-grained distribution of the neural patterns throughout the entire trial. The analysis procedure was adapted from Peñalver et al. (2023) and was performed as follows:

To construct the empirical Representational Dissimilarity Matrices (RDMs), we crossed all the levels of the three independent variables (Task Demand: integration, selection; Target Category: animate, inanimate; Target Relevant Feature: color, shape). This yielded a total of 8 experimental conditions. For every condition and participant, EEG data was selected and averaged every 3 time points along the entire trial epoch (-200 to 7000 ms) to reduce processing load. Voltage values were z-score normalized across trials. To calculate the distance between each pair of conditions, a cross-validated measure of Pearson's correlation coefficient was employed. This cross-validated approach diminishes the risk of bias in the results (Grootswagers et al., 2018; Walther et al., 2016). For every participant, a 3-fold cross-validation went as follows: first, data of every person was randomly divided into 3 chunks, matching condition sizes. On each of the 3 cross-validation folds, one chunk was assigned as the test set, while the remaining 2 as the training set. On every fold, the vectorized patterns of neural activity per condition were trial averaged and centered around 0 by subtracting the mean across conditions of each electrode. Per participant and time point, Pearson's correlation was calculated for every pair of conditions between the training and test sets. The resulting matrices

were transformed into distance matrices by computing 1 – Pearson's coefficient (2 as more dissimilar and 0 as more similar). This was repeated until all chunks had been part of the test set, and the distance values were averaged across the 3 cross-validation folds. An empirical 8 x 8 Representational Dissimilarity Matrix (RDM) was thus obtained per subject and time point.

Theoretical RDMs were created to reflect the expected distances between our conditions according to the variables of interest (Figure 2A). We built three theoretical models driven by the relationship between the instructed components: (1) Task Demand, (2) Target Category, and (3) Target Relevant Feature. All matrices were fully orthogonal to each other. To estimate the share of variance that was explained by each theoretical model at every time point, we fitted them into a multiple linear regression. The theoretical RDMs were included as regressors and the empirical RDM from a given time point as the dependent variable. All matrices were vectorized before performing the regressions. To avoid illusory effects, the diagonal was removed from both empirical and theoretical RDMs (Ritchie et al., 2017). Since the cross-validated approach by its nature does not yield a fully symmetric matrix, all the remaining off-diagonal elements of the matrices were included in the regression. For each participant, we obtained a t-value of every model per time point, where every portion of the variance explained was unique to one specific model.

Statistical significance was assessed non-parametrically via cluster-based permutation (Maris & Oostenveld, 2007; López-García et al., 2022). At the individual level, the labels of the theoretical RDMs were randomly permuted and the permuted matrices were included as regressors on a linear regression. This process was repeated 100 times per participant, resulting in 100 chance level t-values per theoretical model and participant. Next, we created 10^5 group-level null t-value curves by randomly selecting individual subject permutation results and averaging across them. These permuted maps were used to estimate the above and below thresholds of the empirical chance distribution, which considered the 95-percentile of the distribution at each time point and the 95-percentile cluster size detected. To determine the smallest cluster size deemed significant for alpha = 0.05, multiple comparisons' correction was implemented through the False Discovery Rate (FDR).

Time-resolved Multivariate Pattern Analysis (MVPA)

We used time-resolved MVPA (Grootswagers et al., 2018) to infer whether and when the EEG activity patterns encoded the three task-relevant components manipulated by our variables. We trained and tested the classifiers to distinguish between the two levels of each independent variable separately: integration vs. selection, animate vs. inanimate, and color vs. shape trials. Classifications were done on epochs (0 - 7000 ms) that included both task preparation (during instructions' presentation and pre-target windows) and implementation (after target onset).

Pre-processed raw voltage values for every channel and time point were used as features. All analyses were performed with the MVPAlab Toolbox (López-García et al., 2022; available at https://github.com/dlopezg/mvpalab). To minimize computational costs, the classifications were performed every 5 time points. Activity patterns were normalized across trials based on the standard deviation (King and Dehane, 2014), computed in the cross-validation loop. Both training and testing sets were standardized within each fold in the following manner:

$$X_{train} = \frac{(X_{train} - \mu_{train})}{\sigma_{train}}$$

$$X_{test} = \frac{(X_{test} - \mu_{train})}{\sigma_{train}}$$

The *μ_train* and *σ_train* parameters reflect the mean and standard deviation of each feature (channels) of the training set. Next, we smoothed the data using a moving average filter with a sliding time window length of 5 time bins, to increase the signal-to-noise ratio (although slightly reducing the temporal resolution of the data). Afterwards, the number of trials across conditions was balanced to ensure an even distribution of the two classes.

A 5-fold cross-validation approach was followed. Briefly, the data were divided into five chunks and the algorithm used the first four as a training set while testing on the remaining one. This process was iterated five times, until all chunks were used as training and test sets. The final classifier's performance value for each timepoint consisted of the mean performance across folds. Linear Discriminant Analysis (LDA) was used as a classification algorithm since it has shown good performance with EEG data while reducing computational costs (Grootswagers et al., 2018). The area under the curve (AUC) was the performance measure, as it is a non-parametric criterion-free estimate that has been recommended when dealing with 2-class classifications (King & Dehane, 2014). AUC's interpretation is similar to accuracy's: in binary classifications, 0.5 equals chance level (equal probability of true and false positives) and 1 indicates perfect between-class segregation.

Statistical inference to detect above-chance classification was based on a non-parametric cluster-based permutation protocol (Maris & Oostenveld, 2007; López-García et al., 2022). The labels of trials were randomly permuted 100 times for each participant, obtaining 100 AUCs results that represented empirical chance level distributions. Afterwards, we created 10^5 group-level null AUC curves by randomly selecting individual subject permutation results and averaging across them. Then, we employed these permuted maps to threshold our results considering both the 95-percentile null AUC value at each time point, and the 95-percentile cluster size detected in the permuted, null distribution. The final distribution was centered around an AUC value of 0.5 (chance level). Next, all the time points of the group's permuted AUC maps that exceeded the estimated threshold were collected, yielding the normalized null

distribution of cluster sizes. A False Discovery Rate (FDR) was implemented at a cluster level as a correction for multiple comparisons, allowing to achieve the smallest cluster size deemed significant (López-García et al., 2022). This process was done independently for every variable analyzed.

Temporal Generalization Analysis

To examine whether the activity patterns coding the instruction components were stable or changed over time, and also to address if the classifiers generalized from preparation to task implementation, we performed a Temporal Generalization Analysis (King & Dehane, 2014) for each of the independent variables. Following the parameters described above, we trained a classifier on a given time point and tested it on all the time points of the window. We iterated this protocol across the whole epoch, obtaining 3 temporal generalization matrices showing the AUC estimates for every train-test combination. To obtain statistical significance, the procedure indicated for time-resolved analysis was repeated (see *Time-resolved MVPA* in Methods), with the exception that individual-level permutations were reduced to 10 and group-level permutations to 10^4, to decrease computational load.

Cross-Condition Classification Performance (CCGP)

The CCGP allows quantifying the extent to which the neural codes of the three instruction components support generalization, by examining through cross-classification all possible ways of choosing training and testing data (Bernardi et al., 2020). A high average CCGP suggests that the underlying neural codes support task-specific low-dimensional geometries of the neural space. In this case, the classifiers are generalizing across different train-test combinations of the conditions, and not only over noise related to trial-to-trial variations (Bernardi et al., 2020), but they do so at the cost of separability (Badre et al., 2021).

First, the train and test data were split according to the condition's label, with the training and testing sets always consisting of different conditions. Considering that we had 3 main instruction components, the full-crossing of the variables resulted in 8 conditions. For the CCGP analysis of every instruction component (Task Demand, Target Category, and Target Relevant Feature), there were 4 conditions of one level of the variable and 4 of the other, creating a balanced dichotomy of the data (4 conditions of class A vs. 4 conditions of class B). We performed a cross-classification loop by setting 2 of the 8 conditions as the testing set (1 condition of class A vs. 1 condition of class B), while the remaining conditions were kept as the training set (3 conditions of class A vs. 3 conditions of class B), since larger training sets lead to better generalization performance (Bernardi et al., 2020). For example, to compute the CCGP of Task Demand, in each iteration of the analysis we trained a classifier to discriminate

844 the neural patterns of integration and selection from 3/4 of the conditions (e.g., [Selection-
845 Animate-Color], [Selection-Animate-Shape], [Selection-Inanimate-Shape] vs. [Integration-
846 Animate-Shape], [Integration-Inanimate-Color], [Integration-Inanimate-Shape]), and then
847 tested it against the remaining 1/4 of the conditions (in this example, [Selection-Inanimate-
848 Color] vs. [Integration-Animate-Color]). Then, this process was iterated for all unique
849 combinations of the conditions into training and testing sets, yielding a total of 16 repetitions
850 (Bernardi et al., 2020, see *Figure 4 - Figure Supplement 1*). Hence, classifiers were always
851 tested on neural patterns of different conditions from the ones they were trained on. Given that
852 the analysis was repeated along the 16 possible combinations, a 16-step cross-validation loop
853 was implemented along the conditions. For this reason, CCGP analyses usually result in lower,
854 although informative, performance scores than regular MVPA (Bernadi et al., 2020).

855 For every task variable separately, and every iteration of the possible training and testing sets,
856 we performed time-resolved MVPA with a cross-classification approach (López-García et al.,
857 2022). We implemented the same protocol as in *Time-resolved MVPA*, the only change was
858 that no additional cross-validation of the trials was implemented; since training and testing sets
859 always came from different data, there was no risk of overfitting. The mean cross-classification
860 performance after iterating through all 16 training and testing combinations of each task
861 component was the Cross-Classification Generalization Performance (CCGP) of that variable.

862 To assess statistical significance, a non-parametric cluster-based permutation protocol was
863 implemented (Maris & Oostenveld, 2007; López-García et al., 2022). Per subject and task
864 variable, the labels of the conditions of each train-test combination were randomly permuted
865 10 times, resulting in 10 AUCs maps, which represented the empirical chance level distribution
866 of each train-test combination. Then, the permuted maps were averaged across all 16 possible
867 train-test combinations, obtaining 10 permuted maps per subject and instruction component.
868 To perform group-level statistics and obtain the significance of the results, the same steps
869 were followed as in *Time-resolved MVPA*. Significant above-chance AUCs reflect that the
870 decoded task components were coded in an abstract format that favors generalization.

871 *Correlation between Time-Resolved MVPA and CCGP*

872 To test whether the temporal patterns of regular Time-resolved MVPA and CCGP were similar,
873 suggesting that when the task components were encoded in the neural patterns there was
874 matching evidence of abstract coding of information, we performed Pearson's correlations. Per
875 participant, we calculated the Pearson's correlation coefficient (r) of the MVPA and CCGP
876 results including all time-points of the trial, separately for each task component. To compute
877 the average coefficient across participants, we used Fisher's z-transformation (formula below;

Fisher, 1970) to create a normal distribution of the correlation values before averaging, and then reverted to Pearson's correlation coefficient (second formula below).

$$z = \frac{1}{2}\, ln\left(\frac{1+r}{1-r}\right)$$

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

To compute the average p-value of the correlations across participants, we used Fisher's method (formula below; Fisher, 1970). By summing the log-transformed p-values, we obtained Fisher's statistic (χ^2_{Fisher}), which follows a chi-square distribution. To revert to p-values, we calculated the Complementary Cumulative Distribution Function of Fisher's statistic, which provides the upper tail probability of the chi-square distribution.

$$\chi^2_{Fisher} = -2 \sum log\ (p_i)$$

To ensure that the correlation coefficients obtained were robust and not due to the temporal structure of the noise, we repeated this analysis comparing the regular MVPA and CCGP results of different variables pair-wise (e.g. correlation between MVPA results of Task Demand and CCGP results of Target Category), yielding 6 control comparisons.

Temporal Generalization Analysis with cross-classification

To extend the information given by the CCGP, we used Temporal Generalization Analysis with a cross-classification approach. We thus addressed whether the generalizable neural patterns of the instruction components were also stable through the entire epoch. This was performed separately for every instruction component and across the different levels of the remaining variables. In other words, the classifier had to discriminate between the two levels of a variable when trained and tested on different contexts, to establish if the neural codes were generalizable with independence of the condition of the other variables. To reduce computational load, we did not explore all possible data splits of the variables, but pairwise. This way, the analysis evaluated whether Task Demand was generalizable across different Target Categories, and across different Target Relevant Features; whether Target Category was generalizable across different Task Demands, and across different Target Relevant Features; and whether Target Relevant Feature was generalizable across different Task Demands, and across different Target Categories. The analysis protocol was the same as in *Temporal Generalization Analysis*, with the only difference being that training and testing data always came from different conditions. The analysis was done in both directions of the data, i.e. when trained on condition A and tested on condition B, and vice-versa. Afterwards, the

mean results and permutation maps across both directions were obtained, and statistical group analysis was performed for each direction and the mean separately.

Dimensionality Analysis: Shattering Dimensionality (SD)

The extent to which the neural activity generated by the novel instructions is coded in a high-dimensional space was tested through Shattering Dimensionality, a decoding-based approach measuring the decodability of all the binary dichotomies of the conditions with a linear classifier (Bernardi et al., 2020). The number of decodable dichotomies after performing all possible mixtures of the instruction conditions provides information about how many “dimensions” can be extracted from the neural data (Rigotti et al., 2013; Fusi et al., 2016). A high number of decoded dichotomies implies more separable or expanded codes, reflecting higher dimensionality, while a low number implies more compressed neural codes in a lower dimensional space (Rigotti et al., 2013; Fusi et al. 2016).

We followed an analysis protocol similar to Bernardi et al (2020). Given that crossing of our three instruction variables yields a total of 8 conditions, we obtained 35 possible dichotomies with a balanced number of conditions per side of the classification (4 vs. 4; see *Figure 5 - Figure Supplement 1*). Out of the 35 dichotomies, three correspond to our main instruction components, while the rest depict a mixture of the variables. For every dichotomy of the data, we performed a MVPA analysis as indicated in *Time-resolved MVPA*. Hence, the linear classifier was trained and tested to discriminate between the two sides of the dichotomy. The Shattering Dimensionality index was the mean decoding performance across all possible dichotomies. Our analysis yielded a value per time-point of the trial epoch.

To infer statistical significance, we implemented a non-parametric cluster-based permutation protocol (Maris & Oostenveld, 2007; López-García et al., 2022). The labels of the conditions were randomly permuted 10 times per subject and dichotomy, to obtain the empirical chance distribution of each dichotomy. After, the permuted maps were averaged across all 35 dichotomies, yielding the permuted maps per subject. After group-level statistics were implemented on the averaged decoding performance (see *Time-resolved MVPA* in Methods), significant above-chance AUCs reflected the Shattering Dimensionality index, with separable neural codes of the data.

As another measure of dimensionality, we also calculated the proportion of significantly decoded dichotomies (Rigotti et al., 2013; Bhandari et al., 2024). Per time point, we assessed how many dichotomies were significantly decoded above chance. To obtain this result, we performed a non-parametric cluster-based permutation protocol (previously explained; Maris & Oostenveld, 2007; López-García et al., 2022) at the level of each dichotomy. This way, we

obtained which instances of the epoch had been significantly decoded per dichotomy. Afterwards, we calculated the total proportion of significantly decoded dichotomies by adding up the number of significantly decoded dichotomies per time-point.

Neural Geometry - Representational Similarity Analysis (RSA)

We further tested the geometrical arrangement of patterns of brain activation across the two task demands instructed, integration vs. selection of information, to study whether their coding structure favored generalization across other instruction constituents while keeping the two task contexts disaggregated. To estimate the neural geometries between trials with different task components, we employed Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008, following the approach of Muhle-Karbe et al., 2023). We measured the proportion of variance in the neural empirical matrices explained by two geometrical models: offset and orthogonal (see Figure 6A). The theoretical models represented two hypotheses about how Task Demand, our highest-level variable, was geometrically encoded. The empirical RDMs were calculated following the same steps as in *RSA*; thus, 8x8 neural matrices per subject and time point were obtained.

The theoretical RDMs were built by estimating the pairwise distances between the conditions according to two geometry hypotheses about the spatial configuration of the Task Demand: a) the Offset Model is in line with the two levels of Task Demand (integration and selection) being represented in low-dimensional space as separable and parallel planes, which favors generalization while is prone to interference; b) the Orthogonal Model aligns with coding where integration and selection are represented in a geometry that impairs generalization but prevents cross-task interference, with the configuration between both conditions creating a 90º angle (see Figure 6A). The values determined for the generated geometry RDMs ranged from 0 (similar) to 2 (dissimilar) and the matrices included the full crossing of the 3 task components (Task Demand, Target Category, and Target Relevant Feature), resulting in 8x8 RDMs.

To investigate the variance explained by each geometry model, they were fitted into a multiple linear regression. The geometry models were included as regressors, while the neural RDMs were the dependent variables, with all matrices vectorized before conducting the regression. Thus, we obtained the t-values unique to each model. This was repeated separately per participant and time point. Finally, statistical significance was obtained via non-parametric cluster-based permutation testing (see *RSA* in Methods; Maris & Oostenveld, 2007; López-García et al., 2022).

Acknowledgments

This research was supported by grant PID2022-138940NB-100 awarded to MR and grant PID2023-151911NA-I00 awarded to AFP, funded by MCIN/EI/10.13039/501100011033/ and FEDER, UE . AFP was supported by Grant PAIDI21_00207 of the Andalusian Autonomic Government. CCGP was supported by Project PID2020-116342GA-I00 funded by MCIN/AEI/10.13039/501100011033, and Grant RYC2021-033536-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGeneration EU/PRTR. The Mind, Brain and Behavior Research Center receives funding from grants CEX2023-001312-M by MCIN/AEI /10.13039/501100011033 and UCE-PP2023-11 by the University of Granada. We are grateful to Yulya Bedritskaya and Marta Becerra Losada for assistance with data collection, and Chiara Avancini for assistance with the EEG equipment setup.

CRediT author contribution

PP: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing - Original Draft Preparation, Writing - Review & Editing, Visualization. AFP: Conceptualization, Methodology, Software, Investigation, Resources, Data Curation, Writing - Review & Editing, Supervision. CGG: Conceptualization, Writing - Review & Editing, Supervision. MR: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Data and code availability

Raw behavioral and EEG data will be made available upon final publication. Preprocessing and analysis codes can be found at https://github.com/pena-p/inst-comp-eeg.git

References

Aoi, M. C., Mante, V., & Pillow, J. W. (2020). Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature Neuroscience*, *23*(11), 1410-1420. https://doi.org/10.1038/s41593-020-0696-5

Badre, D., Bhandari, A., Keglovits, H., & Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, *38*, 20-28. https://doi.org/10.1016/j.cobeha.2020.07.002

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in cognitive sciences*, *22*(2), 170-188.

Bellman, R.E. (1957). *Dynamic Programming* (Princeton University Press).

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., & Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, *183*(4), 954-967.e21. https://doi.org/10.1016/j.cell.2020.09.031

Bhandari, A., Keglovits, H., Chicklis, E., & Badre, D. (2024). *Task structure tailors the geometry of neural representations in human lateral prefrontal cortex* [Preprint]. Neuroscience. https://doi.org/10.1101/2024.03.06.583429

Bourguignon, N. J., Braem, S., Hartstra, E., De Houwer, J., & Brass, M. (2018). Encoding of novel verbal instructions for prospective action in the lateral prefrontal cortex: evidence from univariate and multivariate functional magnetic resonance imaging analysis. *Journal of cognitive neuroscience*, 30(8), 1170-1184. https://doi.org/10.1162/jocn_a_01270

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. *Variation in working memory*, 75, 106. https://doi.org/10.1093/acprof:oso/9780195168648.003.0004

Cellier, D., Petersen, I. T., & Hwang, K. (2022). Dynamics of Hierarchical Task Representations. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *42*(38), 7276-7284. https://doi.org/10.1523/JNEUROSCI.0233-22.2022

Cichy, R. M., & Oliva, A. (2020). AM/EEG-fMRI fusion primer: resolving human brain responses in space and time. *Neuron*, *107*(5), 772-781. https://doi.org/10.1016/j.neuron.2020.07.001

Cole, M. W., Etzel, J. A., Zacks, J. M., Schneider, W., & Braver, T. S. (2011). Rapid transfer of abstract rules to novel contexts in human lateral prefrontal cortex. *Frontiers in human neuroscience*, 5, 142. https://doi.org/10.3389/fnhum.2011.00142

Cole, M. W., Laurent, P., & Stocco, A. (2013). Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, 13, 1-22. https://doi.org/10.3758/s13415-012-0125-7

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dumontheil, I., Thompson, R., & Duncan, J. (2011). Assembly and use of new task rules in fronto-parietal cortex. *Journal of Cognitive Neuroscience*, 23(1), 168-182. https://doi.org/10.1162/jocn.2010.21439

Duncan, J. (2010). The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in Cognitive Science*, 14(4), 172-179. https://doi.org/10.1016/j.tics.2010.01.004

Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., & Shea-Brown, E. (2022). Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion. *Nature Machine Intelligence*, *4*(6), 564-573. https://doi.org/10.1038/s42256-022-00498-0

Fisher, R. A. (1970). Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution* (pp. 66-70). New York, NY: Springer New York.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., & Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, *110*(7), 1258-1270.e11. https://doi.org/10.1016/j.neuron.2022.01.005

Formica, S., González-García, C., Senoussi, M., & Brass, M. (2021). Neural oscillations track the maintenance and proceduralization of novel instructions. *NeuroImage*, 232, 117870. https://doi.org/10.1016/j.neuroimage.2021.117870

Formica, S., González-García, C., Senoussi, M., Marinazzo, D., & Brass, M. (2022). Theta-phase connectivity between medial prefrontal and posterior areas underlies novel instructions implementation. *Eneuro*, *9*(4). https://doi.org/10.1523/ENEURO.0225-22.2022

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66-74. https://doi.org/10.1016/j.conb.2016.01.010

González-García, C., Arco, J. E., Palenciano, A. F., Ramírez, J., & Ruz, M. (2017). Encoding, preparation and implementation of novel complex verbal instructions. *NeuroImage*, 148, 264-273. https://doi.org/10.1016/j.neuroimage.2017.01.037

Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, *179*, 252-262. https://doi.org/10.1016/j.neuroimage.2018.06.022

Hartstra, E., Kühn, S., Verguts, T., & Brass, M. (2011). The implementation of verbal instructions: an fMRI study. *Human brain mapping*, 32(11), 1811-1824. https://doi.org/10.1002/hbm.21152

Hartstra, E., Waszak, F., & Brass, M. (2012). The implementation of verbal instructions: dissociating motor preparation from the formation of stimulus–response associations. *Neuroimage*, 63(3), 1143-1153. https://doi.org/10.1016/j.neuroimage.2012.08.003

Kadohisa, M., Kusunoki, M., Mitchell, D. J., Bhatia, C., Buckley, M. J., & Duncan, J. (2023). Frontal and temporal coding dynamics in successive steps of complex behavior. *Neuron*, *111*(3), 430-443.e3. https://doi.org/10.1016/j.neuron.2022.11.004

Kikumoto, A., & Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(19), 10603-10608. https://doi.org/10.1073/pnas.1922166117

Kikumoto, A., Bhandari, A., Shibata, K., & Badre, D. (2023). *A Transient High-dimensional Geometry Affords Stable Conjunctive Subspaces for Efficient Action Selection*. *bioRxiv*. https://doi.org/10.1101/2023.06.09.544428

Kikumoto, A., Shibata, K., Nishio, T., & Badre, D. (2024). Practice Reshapes the Geometry and Dynamics of Task-tailored Representations. *bioRxiv*, 2024-09.

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

Kriegeskorte, N., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci*. 2, 1–28. https://doi.org/10.3389/neuro.06.001.2008

López-García, D., Peñalver, J. M., Górriz, J. M., & Ruz, M. (2022). MVPAlab: A Machine Learning decoding toolbox for multidimensional electroencephalography data. *Computer Methods and Programs in Biomedicine*, 214, 106549. https://doi.org/10.1016/j.cmpb.2021.106549

Manelis, A., & Reder, L. M. (2015). He Who is well prepared has half won the battle: an fMRI Study of task preparation. *Cerebral cortex*, 25(3), 726-735. https://doi.org/10.1093/cercor/bht262

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, 164(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202. https://doi.org/10.1146/annurev.neuro.24.1.167

Muhle-Karbe, P. S., Duncan, J., De Baene, W., Mitchell, D. J., & Brass, M. (2017). Neural coding for instruction-based task sets in human frontoparietal and visual cortex. *Cerebral Cortex*, 27(3), 1891-1905. https://doi.org/10.1093/cercor/bhw032

Muhle-Karbe, P. S., Sheahan, H., Pezzulo, G., Spiers, H. J., Chien, S., Schuck, N. W., & Summerfield, C. (2023). Goal-seeking compresses neural codes for space in the human hippocampus and orbitofrontal cortex. *Neuron*, *111*(23), 3885-3899.e6. https://doi.org/10.1016/j.neuron.2023.08.021

Musslick, S., & Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, *25*(9), 757-775. https://doi.org/10.1016/j.tics.2021.06.001

Palenciano, A. F., González-García, C., Arco, J. E., Pessoa, L., & Ruz, M. (2019a). Representational Organization of Novel Task Sets during Proactive Encoding. *Journal of Neuroscience*, *39*(42), 8386-8397. https://doi.org/10.1523/JNEUROSCI.0725-19.2019

Palenciano, A. F., González-García, C., Arco, J. E., & Ruz, M. (2019b). Transient and sustained control mechanisms supporting novel instructed behavior. *Cerebral Cortex,* 29(9), 3948-3960. https://doi.org/10.1093/cercor/bhy273

Parthasarathy, A., Herikstad, R., Bong, J. H., Medina, F. S., Libedinsky, C., & Yen, S.-C. (2017). Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience*, *20*(12), 1770-1779. https://doi.org/10.1038/s41593-017-0003-2

Peñalver, J. M., Lopez-Garcia, D., González-García, C., Aguado-Lopez, B., Górriz, J. M., & Ruz, M. (2023). Top-down specific preparatory activations for selective attention and perceptual expectations. *NeuroImage*, 271, 119960. https://doi.org/10.1016/j.neuroimage.2023.119960

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., & Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific data*, 6(1), 103. https://doi.org/10.1038/s41597-019-0104-8

Pion-Tonachini, L., Kreutz-Delgado, K., & Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198, 181-197. https://doi.org/10.1016/j.neuroimage.2019.05.026

RDC Team. (2016). R: A language and environment for statistical computing. *R Found. Stat. Comput.*, *1*, 409.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), Article 7451. https://doi.org/10.1038/nature12160

Ritchie, J. B., Bracci, S., & de Beeck, H. O. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage*, *148*, 197-200.

Riveland, R., & Pouget, A. (2024). Natural language instructions induce compositional generalization in networks of neurons. *Nature Neuroscience*, 1-12. https://doi.org/10.1038/s41593-024-01607-5

Ruge, H., Schäfer, T. A., Zwosta, K., Mohr, H., & Wolfensteller, U. (2019). Neural representation of newly instructed rule identities during early implementation trials. *Elife*, *8*, e48293. https://doi.org/10.7554/eLife.48293

Sakai, K. (2008). Task set and prefrontal cortex. *Annu. Rev. Neurosci.*, *31*, 219-245. https://doi.org/10.1146/annurev.neuro.31.060407.125642

Sakai, K., & Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nature neuroscience*, 6(1), 75-81. https://doi.org/10.1038/nn987

Sheahan, H., Luyckx, F., Nelli, S., Teupe, C., & Summerfield, C. (2021). Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron*, *109*(7), 1214-1226.

Sobrado, A., Palenciano, A. F., González-García, C., & Ruz, M. (2022). The effect of task demands on the neural patterns generated by novel instruction encoding. *Cortex*, 149, 59-72. https://doi.org/10.1016/j.cortex.2022.01.010

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78(2), 364-375. https://doi.org/10.1016/j.neuron.2013.01.039

Stokes, M. G., Muhle-Karbe, P. S., & Myers, N. E. (2020). Theoretical distinction between functional states in working memory and their corresponding neural states. *Visual Cognition*, *28*(5-8), 420-432.

Vaidya, A. R., Jones, H. M., Castillo, J., & Badre, D. (2021). Neural representation of abstract task structure during generalization. *eLife*, *10*, e63226. https://doi.org/10.7554/eLife.63226

Verbeke, P., & Verguts, T. (2022). Using top-down modulation to optimally balance shared versus separated task representations. *Neural Networks*, *146*, 256-271. https://doi.org/10.1016/j.neunet.2021.11.030

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, *137*, 188-200. https://doi.org/10.1016/j.neuroimage.2015.12.012

Westfall, J. (2015). PANGEA: Power analysis for general ANOVA designs. *Unpublished manuscript*. Available at http://jakewestfall. org/publications/pangea.pdf, 4.

Wójcik, M. J., Stroud, J. P., Wasmuht, D., Kusunoki, M., Kadohisa, M., Myers, N. E., Hunt, L. T., Duncan, J., & Stokes, M. G. (2023). *Learning shapes neural geometry in the prefrontal cortex* (p. 2023.04.24.538054). bioRxiv. https://doi.org/10.1101/2023.04.24.538054

Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011a). Adaptive coding of task-relevant information in human frontoparietal cortex. *Journal of Neuroscience*, 31(41), 14592-14599. https://doi.org/10.1523/JNEUROSCI.2616-11.2011

Woolgar, A., Thompson, R., Bor, D., & Duncan, J. (2011b). Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. *Neuroimage*, 56(2), 744-752. https://doi.org/10.1016/j.neuroimage.2010.04.035

Woolgar, A., Williams, M. A., & Rich, A. N. (2015). Attention enhances multi-voxel representation of novel objects in frontal, parietal and visual cortices. *NeuroImage*, *109*, 429-437. https://doi.org/10.1016/j.neuroimage.2014.12.083