bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The effect of task demands on the neural patterns generated by novel instruction encoding

Alberto Sobrado[a,b], Ana F. Palenciano[c], Carlos González-García[a,b], María Ruz[a,b]

[a]Mind, Brain and Behavior Research Center, University of Granada, Spain
[b]Department of Experimental Psychology, University of Granada, Spain
[c]Department of Experimental Clinical and Health Psychology, Ghent University, Belgium

*Corresponding author. Mind, Brain and Behavior Research Center, University of Granada, 18071, Granada, Spain.
E-mail address: mruz@ugr.es (M. Ruz).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Verbal instructions allow fast and optimal implementation of novel behaviors. Previous research has shown that different control-related variables organize neural activity in frontoparietal regions during the preparation of novel instructed task sets. Little is known, however, about how such variables organize brain activity under different task demands. In this study, we assessed the impact of implementation and memorization demands on the neural representation of novel instructions. We combined functional Magnetic Resonance Imaging (fMRI) with an instruction-following paradigm to compare the effect of three relevant control-related variables (integration of dimensions, response complexity, and stimulus category) across demands, and to explore the degree of overlap between these. Our results reveal, first, that the implementation and memorization of novel instructions share common neural patterns in several brain regions. Importantly, they also suggest that the preparation to implement instructions results in a strengthened coding of relevant control-related information in frontoparietal areas compared to their mere memorization. Overall, our study shows how the content of novel instructions proactively shapes brain activity based on multiple dimensions and how these organizational patterns are strengthened during implementation demands.

Keywords

Cognitive Control, verbal instructions, fMRI, Multivariate Pattern Analysis, Frontoparietal network

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. INTRODUCTION

Instruction following allows humans to implement novel behaviors quickly without prior practice, in contrast to other mechanisms such as trial-and-error or reinforcement learning. This enhanced efficiency frames such ability as a special instance of humans' cognitive flexibility (Cole, Laurent, & Stocco, 2013). An important yet unanswered question is how the brain rapidly reformats and organizes the symbolic information conveyed by instructions into efficient action (Brass, Liefooghe, Braem, & De Houwer, 2017). Recent findings have shown that, when preparing to execute a novel instruction, brain activation patterns are organized by the different relevant dimensions of such instruction (Palenciano, González-García, Arco, Pessoa, & Ruz, 2019). However, it remains unknown whether these activation patterns are triggered by the preparation to implement instructions, or if they reflect their mere declarative maintenance.

Prior studies have reported the pervasive effects of instruction following on behavioral and neural markers. The intention to execute a recently encoded instruction induces brain activation in areas associated with control and category-selective perceptual processing (González-García, Arco, Palenciano, Ramírez, & Ruz, 2017; González-García, Formica, Wisniewski, & Brass, 2021), and impacts the neural patterns representing the instructed content (Bourguignon, Braem, Hartstra, De Houwer, & Brass, 2018; Muhle-Karbe, Duncan, De Baene, Mitchell, & Brass, 2017; Ruge, Schäfer, Zwosta, Mohr, & Wolfensteller, 2019). In this line, a recent study (Palenciano, González-García, Arco, Pessoa, et al., 2019) showed that, during implementation demands, neural patterns in areas such as the inferior frontal gyrus (IFG) are organized by the need to integrate information from different dimensions of the instruction (e.g. color and size of the stimuli), while patterns in the intraparietal sulcus (IPS) and pre-supplementary motor area (pre-SMA) represent the complexity of instructed stimulus-response associations (Palenciano, González-García, Arco, Pessoa, et al., 2019).

Another set of neuroimaging studies from the instruction following literature has focused on how the declarative information of the instruction is transformed into a procedural format, that is, into an action-oriented proactive binding of relevant motor and perceptual information (see Brass et al. 2017 for a review). This transformation (often defined as *proceduralization*) enables a highly accessible task model containing the condition-action rules from the instruction ready to be implemented, creating an optimal preparatory and

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reflexive-like state that enables a fast and optimal execution (Brass et al., 2017; but see Liefooghe & De Houwer, 2017). Interestingly, brain areas linked to novel instruction processing (such as the dorsal premotor cortex and middle frontal gyrus) seem to be involved also when participants only need to declaratively memorize (but not execute) the content of the instruction, although they show higher levels of activation or information decodability under implementation demands (Bourguignon et al., 2018; Muhle-Karbe et al., 2017). This raises the question of how exactly declarative and procedural neural states differ.

A prominent theoretical proposal of instruction processing puts forward a 3-step model of how the information contained in verbal instructions is transformed into action plans (Brass et al., 2017). First, the instruction content has to be encoded, building the representation of the declarative information and rules that specify the proper response (Hartstra, Waszak, & Brass, 2012; Sakai, 2008). Afterwards, a preparation stage takes place before response execution (Sakai, 2008), where the task set is assembled, and its proactive maintenance induces an adjustment of the features relevant to achieve the task (González-García et al., 2017; Muhle-Karbe et al., 2017; Ruge, Jamadar, Zimmermann, & Karayanidis, 2013). Finally, the task-set is executed and the action requested by the instruction is carried out (Stocco, Lebiere, O'Reilly, & Anderson, 2012). As mentioned before, proactive control not only biases perceptual and motor systems to enhance processing of upcoming stimuli and relevant responses, but it also organizes neural activity according to control-related variables, such as the need to integrate dimensions, or response complexity (Cole, Patrick, Meiran, & Braver, 2018; Palenciano, González-García, Arco, & Ruz, 2019). However, whether the reported proactive organization in control-related areas underlies the procedural implementation that ultimately leads to execution, or alternatively, the declarative memorization of the novel task demands remains unknown.

The aim of the current study was to test the extent to which proactive pattern organization in control-related regions is specific to the proceduralization of instruction content. To that end, we adapted an instruction-following fMRI paradigm (González-García et al., 2017; Palenciano, González-García, Arco, Pessoa, et al., 2019; Palenciano, González-García, Arco, & Ruz, 2019), in which novel verbal instructions had to be either implemented (proceduralized) or memorized (non-proceduralized). Across both conditions, the instructions were manipulated according to three variables related with

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proactive preparation (integration of information dimensions, response complexity and target category). Using univariate and multivariate pattern analysis (MVPA; Haxby, Connolly, & Guntupalli, 2014) we aimed to explore the strength with which control-related variables organize patterns of brain activity during both implementation and memorization demands (Bourguignon et al., 2018; Muhle-Karbe et al., 2017; Palenciano, González-García, Arco, Pessoa, et al., 2019). We expected, first, that the content of instructions affected activation patterns during task encoding in several areas in implementation and memorization conditions. Specifically, we predicted the engagement of the IFG when instructions required the integration of different stimulus dimensions, of premotor and motor areas with increased response complexity, and of visual areas when dealing with different target categories. Second, we hypothesized higher decoding accuracy of such variables in the implementation condition than in memorization.

2. METHODS

2.1. Participants

Thirty-seven students from the University of Granada (Spain) took part in the study (29 females, 8 males, mean age = 22.97, SD = 3.42). The participants were native Spanish speakers, right-handed and with normal or corrected-to-normal vision. They received economic compensation (20-35€, depending on performance) for their participation. They all signed a consent form approved by the Ethics Committee for Human Research of the University of Granada. Two participants were excluded due to excessive head movement (> 3mm) and other three due to low performance (< 75% of correct responses), resulting in a final sample of 32 participants. The sample size (n=32) was calculated a priori with the power analysis software PANGEA (Westfall, 2015), for a power of 0.8 and an expected effect size of 0.3 (Cohen's *d*) in a two-way behavioral interaction (see behavioral results).

2.2. Stimuli, apparatus

The stimuli consisted of 192 different verbal instructions (taken from Palenciano, González-García, Arco, Pessoa, et al., 2019). They all had the same "If …. then …" structure and were composed by two conditional statements and two responses (e. g.: "*If there are two vegetables and one fruit, press A. If not, press L"*). *A* and *L* corresponded to left- and right-hand fingers, respectively. Participants used their middle fingers for one

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

task (e.g., implementation) and the index fingers for other (e.g., memorization). The assignation of middle/index fingers to a specific task was counterbalanced across participants. All the instructions referred to different features of human faces or food items. The different face features were: *gender* (male, female), *emotion* (happy, sad), *race* (black, white), and *size* (large, small), whereas food features were: *type* (vegetable, fruit), *color* (green, yellow), *shape* (elongated, rounded) and *size* (large, small). Every face-related instruction had a food-related instruction counterpart, which was achieved equating gender and food type, emotion and color, and skin color and shape across the two categories. Given that targets consisted of a grid of 8 stimuli (see below), instructions also specified the number of items from the grid that had to be taken into consideration to respond (*one, two, or three*). Critically, we manipulated the structure of the instructions, which differed in three key aspects: **1) Integration of stimuli dimensions** (within or between dimensions), as instructions could refer to features related to the same (e. g.: "*If there are two women and one man"*; dimension = gender) or different dimensions (*"If there are two women and one happy person"*; dimensions = gender + emotion). **2) Response complexity,** as responses required could be single (*"press A. If not press L"*) or sequential (*"press AL. If not press LA"*). **3) Target category,** reflecting the fact that half of the instructions referred to faces and the other half referred to food. These three variables were manipulated in an orthogonal fashion (the level of one variable in an instruction was independent from the levels of the other two). Note that these three variables were fully crossed within each trial, at the same time in each instruction (e.g., "*If there are two women and one happy person, press A. If not, press L.* would be classified as pertaining to integration *between dimensions*, as having a *single* response and as pertaining to *faces)*. Additionally, task demands were separated in two block types, **implementation** and **memorization**, similar to previous studies (e.g., Muhle-Karbe et al., 2017). In implementation blocks, each instruction could lead to two associated grids of target stimuli: one fulfilling the conditions and another one not. All grids consisted of combinations of 4 faces and 4 food items, drawn from a pool of 16 stimuli: 8 faces (4 men, 4 women, 4 happy, 4 sad, 4 white, 4 black, extracted from the NimStim database; Tottenham et al., 2009) and 8 food items (4 vegetables, 4 fruits, 4 yellow, 4 green, 4 elongated, 4 rounded; extracted from Palenciano, González-García, Arco, Pessoa, et al., 2019). The 16 stimuli could appear in big or small size, generating a pool of 16 items per category (32 in total). All grids were generated with a specific combination of stimuli, each appearing only once during the whole experiment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In memorization blocks, instructions had the same characteristics as in implementation, but targets contained another instruction instead of a grid of stimuli. This target instruction could be the same (50%) or different than the one encoded first. Different target instructions were created by exchanging either one of the stimulus features from the original one (e.g., "*If there are two vegetables and one yellow food item*" instead of "*If there are two vegetables and one green food item",* where the feature *color* changed from yellow to green) or the response (e.g., "*press A. If not press L"* instead of "*press LA. If not press AL"*). In the target screen, the words "different" and "same" also appeared to the right and left of the instructions, indicating the keys to press. The response mapping of these keys changed across trials to prevent any potential response preparation before target onset, similar to preceding studies (Formica, González-García, Senoussi, & Brass, 2021). In the scanner, the main task was organized in 12 blocks (6 implementation, 6 memorization), each containing 16 trials (192 in total, 96 implementation and 96 memorization trials).

The task was presented through a screen connected to a computer running Matlab with the Psychophysics Toolbox (Brainard, 1997), with a set of mirrors mounted on the head coil, allowing the participants to see the screen. Responses were given through an MRI-compatible response pad with middle and index fingers of both hands.

The independent variables (IVs) of this study were 4. The first one was the task to execute (implementation vs. memorization). The other 3 IVs matched the instruction manipulations: Integration of Dimensions (within vs. between dimensions), Response complexity (single vs. sequential), and Stimuli Category (faces vs. food).

2.3. Procedure

The day before the scanning took place, participants attended a practice session where they were instructed about the tasks to be performed, and completed at least two blocks of each task type (implementation and memorization), with a different set of instructions (equivalent in parameters to the set used in the fMRI session). If accuracy in any of the two block types was lower than 85%, they had to repeat the tasks. They performed a maximum of 3 practice repetitions, and if they did not reach the desired accuracy level, they were not invited to the scanning session. In this case, they received payment for the time spent in the practice session (at an hourly rate of 5 Euros). Six participants were not able to reach the minimum accuracy required and therefore not further tested.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the scanner, at the beginning of each block, a verbal cue indicated the task to be performed (implementation vs. memorization). All trials began with an instruction (encoding stage; 2.5 s; 25.75º), which needed to be later executed (implementation) or remembered (memorization). Then, a fixation point (0.5º) was presented during a jittered interval ranging from 4 to 7.5 s (steps of 500 ms, average of 5.75 s). This jitter was followed by a target grid (2.5 s; 21º) in implementation trials, and by another instruction (2.5 s; 28.5º) in memorization trials. The first and second instructions within each memorization trial also differed in font size and case (lower and uppercase for the first and second instruction, respectively). This manipulation sought to avoid that participants responded *same* or *different* based on perceptual invariance or physical changes between the two instructions. Participants used the middle fingers for one task (e.g., implementation) and the index fingers for other (e.g., memorization). The use of middle and index for one task or another (e.g., middle for implementation, index for memorization) was counterbalanced between subjects. In both implementation and memorization conditions, the trial ended with a jitter of the same characteristics as the previous one. In all cases, instructions were always new and were presented only once during the entire experiment. The sequence of events is depicted in Figure 1.

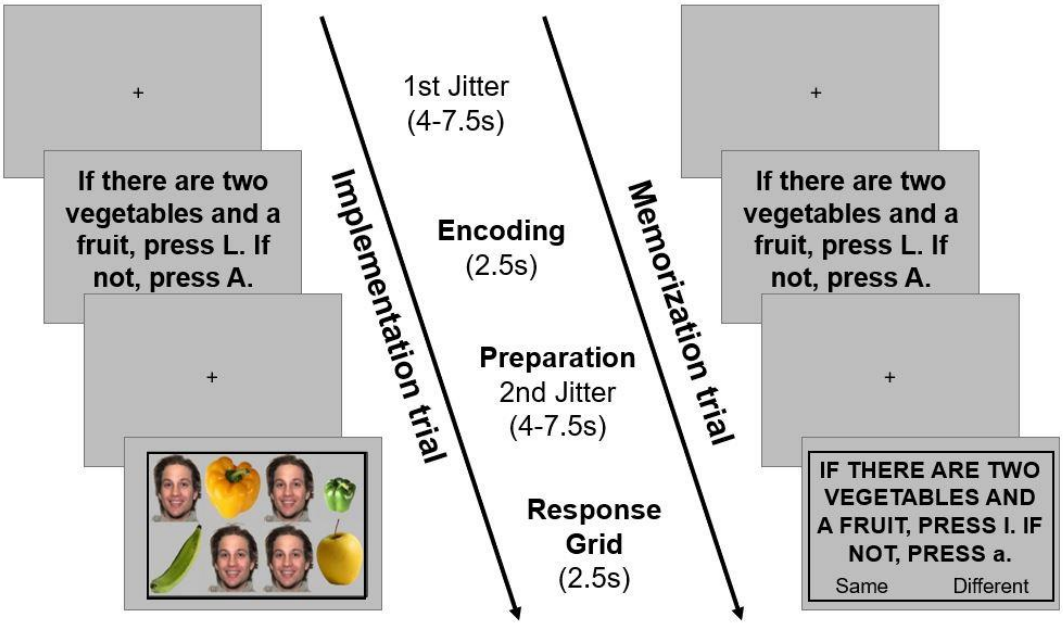

Figure 1. Behavioral paradigm.

Implementation and memorization blocks were presented in alternation, with order counterbalanced across participants. The other conditions were equally distributed in all blocks and appeared in random order within each block. Jitters were organized so their

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

distribution and the average duration were equated across blocks. Participants spent ~90 minutes in the scanner, with the main task lasting ~75 minutes.

2.4. fMRI acquisition, preprocessing and analysis

Participants' MRI data were acquired with a 3T Siemens Magneton Trio scanner located at the Mind, Brain, and Behavior Research Center (University of Granada, Spain). Functional images were collected using a T2*-weighed echo-planar imaging (EPI) sequence (TR = 2000 ms, TE = 24 ms, flip angle = 70°). Each volume consisted of 34 slices, obtained in descending order, with 3.0 mm thickness (gap 20%, voxel size = 3 mm^3). For each participant, a total of 1740 volumes were obtained, in 12 runs of 145 volumes each. Additionally, we acquired a structural image with a high-resolution anatomical T1-weighted sequence (192 slices of 1 mm, TR = 2500 ms, TE = 3.69 ms, flip angle = 7°, voxel size = 1 mm^3).

We used SPM12 to preprocess and analyze the data. The first 4 volumes of each run were discarded to allow stabilization of the signal. The remaining volumes were spatially realigned, unwarped and slice-time corrected. Then, the anatomical T1 was coregistered to the realigned functional images and segmented into different brain tissues. The deformation fields thus obtained were used to normalize the functional data to the MNI space (3 mm^3 voxel size). Last, the images were smoothed using an 8 mm Gaussian kernel. Multivariate analyses at the individual level (see below) were conducted with non-normalized and non-smoothed images, and results were then normalized and smoothed before whole-brain group level analysis. All analysis, uni- and multivariate, were done focusing on the encoding stage of the trial. This was done to avoid potential confounds from differences at the target stage in visual (grid of stimuli vs. instruction) and motor response levels (single or sequential responses vs. fixed single response) between the implementation and memorization conditions.

2.4.1. Univariate analysis

For the first level of analysis, a General Linear Model was estimated with separate regressors for all the combinations of conditions per run (16 in total, resulting from the crossing 2x2x2 of the IVs Integration of Dimensions, Response complexity, and Stimuli Category). We defined two events per trial: the instruction and grid, both modeled with their duration (2.5s) and convolved with the canonical hemodynamic response function.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Every regressor was modelled as a combination of the two trials of the same condition per run. Jitters were not modeled and contributed to the implicit baseline. Additionally, we included 6 motion parameters and errors (error trials were modelled with the full duration of instruction, grid and both jitters) as nuisance regressors. At the second level, we performed 2 separated contrasts (t-tests) at the individual level (implementation >/< memorization) with the instruction regressor. The individual maps of activation were then entered into a second-level (group) analysis. The rationale was to find the brain areas that showed increased activation during the encoding of instructions to be implemented vs. memorized, and vice versa. Additionally, we used those areas as ROIs for subsequent MVPA analysis to study whether they also showed differential strength of patterns related to the content of the instructions (see *2.4.3 MVPA: ROI-based decoding*). We report the results surviving an FWE cluster-level correction of $p < 0.05$ for multiple comparisons (from an initial uncorrected threshold of $p < 0.001$).

2.4.2. MVPA: Whole-brain cross-classification

As we aimed to study how task demands impact the way instructed content is represented, we performed multivariate pattern cross-classification (Kaplan, Man, & Greening, 2015) between implementation and memorization to find the brain areas that showed similar patterns across these conditions. We used these results to obtain a set of ROIs using a Leave-One-Subject-Out (LOSO) approach, to avoid circularity (Esterman et al., 2010), complementing those resulting from the univariate analysis, to assess the effect of task on the patterns generated by the encoding of the instructions. Whereas with the univariate results we tried to find brain areas that diverged in activation levels between implementation and memorization, with the cross-classification analysis we aimed to find regions that coded for both tasks in a similar way. To perform this analysis, first, we ran another GLM with the same structure as the one used for univariate analyses but on non-normalized and unsmoothed data. We estimated trial-wise BOLD responses using a Least-Square Separate approach (LSS; Arco, González-García, Díaz-Gutiérrez, Ramírez, & Ruz, 2018; Mumford, Davis, & Poldrack, 2014), to gain sensitivity and to reduce collinearity between regressors. After that, we performed a cross-classification analysis (Kaplan et al., 2015) between implementation and memorization encoding screens using a searchlight approach across the brain (sphere's radius = 4 voxels). In this analysis, if a classifier algorithm trained with activation patterns of one task (e.g., implementation) is capable of performing successful classifications in another task (e.g., memorization), this

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is interpreted as evidence of generalizability between tasks, that is, a common encoding neural code. For cross-classification, we trained three separated Support Vector Machines, one for each instruction variable (integration of stimuli dimensions, response complexity and target category), and used them separately to decode between the two levels of each variable (within vs. between dimensions, single vs. sequential responses, faces vs. food instruction). Crucially, the algorithm was trained (e.g., differentiating instructions referring to faces vs. food) in one task (e.g., implementation) and then tested in the other one (e.g., memorization). Then, the opposite scheme was performed. Note that having different fingers to respond to implementation and memorization could not confound or bias the results of this analysis, given that we always decoded classes (e.g., faces from food instructions) within each of the two tasks (implementation or memorization), and not across them. To cross-validate the classifiers' performance we followed a leave-one-run out scheme, training the classifier in 5 of the 6 runs and testing it in the remaining one. This resulted in a 6-fold scheme, with data from each run used as the test set once. Then, we averaged the decoding accuracy maps across runs and across cross-classification directions and introduced them into a group-level one-sample t-test against chance (resulting in three group maps, one for each instruction variable). The resulting areas after correction (thresholded maps at $p<.001$) were used as candidate ROIs (see *2.4.3 MVPA: ROI-based decoding*) for the subsequent analysis. In the set of ROIs estimated through cross-classification not all the clusters obtained in the group analysis were observed across the LOSO iterations in all participants. Because this inconsistency could lead to unreliable or biased results, we excluded these regions from further analysis. Thus, we only included ROIs found in more than 25 participants (80% of the sample).

2.4.3 MVPA: ROI-based decoding

With the previous analyses (both univariate and multivariate), we obtained sets of ROIs to test whether the implementation and the memorization demands have an impact on the neural patterns generated by verbal instructions. The univariate results used to obtain the ROIs were group maps coming from activation differences between implementation and memorization. In the multivariate case, we used decoding accuracy maps coming from cross-classification between implementation and memorization. To select the ROIs for each participant, we followed a LOSO approach. Specifically, we performed group level t tests (implementation>memorization, memorization<implementation, and decoding accuracy regarding Integration of Dimensions, Response complexity, and Stimuli

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Category), but excluding that specific participant's data from the analysis to avoid circularity. The significant clusters of each LOSO iteration that matched the areas resulting at the group level were used as ROIs for that individual participant. These resulting ROIs were inverse normalized to the participants' native space. Then, in each of these ROIs, we performed decoding analysis to classify between the two levels of each instruction variable (e.g.: faces vs. food instruction), separately for implementation and memorization trials. We followed this entire procedure five times, two for implementation>/<memorization and one for each instruction IVs. The result was one decoding accuracy value per participant, ROI and task (implementation and memorization). Last, we performed Wilcoxon signed-rank test for every ROI and task, to test if the classifier showed above chance accuracy, and Wilcoxon signed-rank test between tasks for every ROI to assess whether the classification accuracy differed between implementation and memorization. All ROI results were Bonferroni-corrected with an α threshold of $p < .05$/number of ROIs to control for multiple comparisons.

3. RESULTS

3.1. Behavioral results

Accuracy and reaction times (RT) were analyzed separately with two four-factor repeated-measures ANOVAs: Task Type, Integration of Dimensions, Response Complexity and Target Category.

Behavioral results provided initial evidence of the impact of task demands on the effect of instruction dimensions. This was reflected on the three interactions between the task and the rest of the variables in RT: task x integration ($F_{1,31} = 36.231$, $p < 0.001$, $\eta_p^2 = 0.539$), task x response ($F_{1,31} = 17.019$, $p < 0.001$, $\eta_p^2 = 0.354$), and task x target ($F_{1,31} = 29.696$, $p < 0.001$, $\eta_p^2 = 0.489$). We performed post-hoc tests to better characterize these interactions and found that the interaction task x integration revealed smaller differences between within- vs. between-dimensions in memorization ($p = 0.004$, all other $p_s < 0.001$). In task x response, the interaction was driven by smaller differences between single vs. sequential responses in implementation ($p = 0.016$, all other $p_s < 0.002$). Last, the interaction task x target was caused by differences between all levels of the factors except between faces and food in memorization ($p > 0.1$, all other $p_s < 0.001$). A summary of the rest of the behavioral main effects can be seen in Table 1 (excluded ANOVA terms were all non-significant, neither in accuracy nor RT). The remaining contrasts, not

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

primary to our hypotheses, are detailed in Table 1.

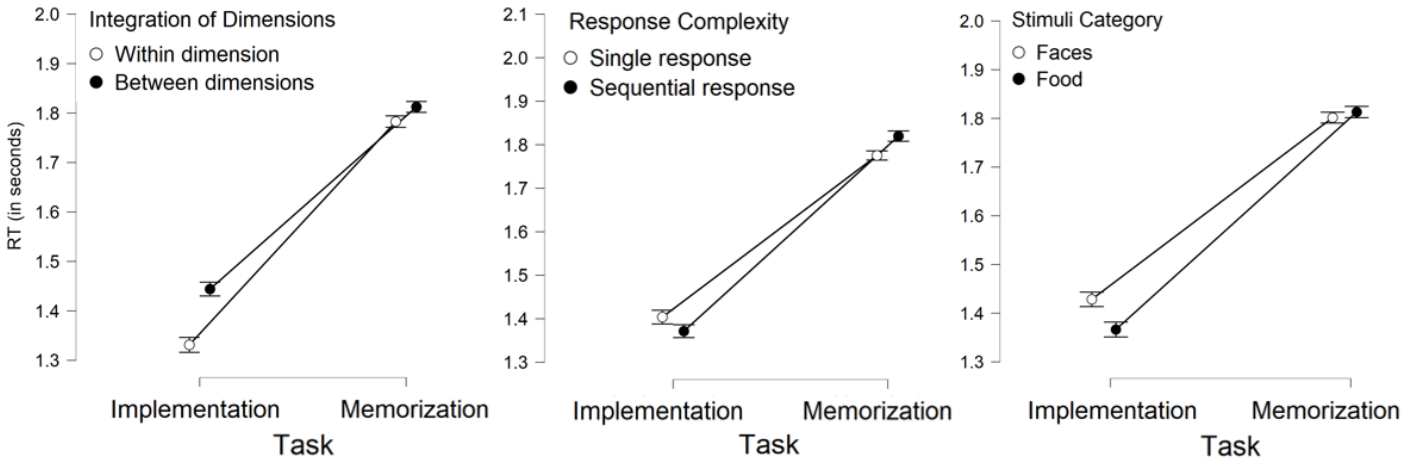

Figure 2. Reaction times as a function of implementation and memorization demands in the three remaining instruction variables (Integration of dimension, response complexity and stimuli category). Error bars represent Standard Error of the Mean.

Table 1. Main effects of variables on the behavioral results (* p < .05, ** p < .01, *** p < .001).

	F	p	η_p^2
Task (ACC)	57.203	**<.001*** **	0.649
Task (RT)	491.199	**<.001*** **	0.941
Integration of Dimensions (ACC)	16 .978	**<.001*** **	0.354
Integration of Dimensions (RT)	101 .797	**<.001*** **	0.767
Response Complexity (ACC)	9.520	**.004** **	0.235
Response Complexity (RT)	0.435	.514	0.014
Target Category (ACC)	14 .907	**<.001*** **	0.325
Target Category (RT)	9.613	**0.004** **	0.237

It is worth mentioning that responses to implementation were faster but less accurate that memorization responses, which could indicate a speed-accuracy trade-off. To rule out this option, we performed an additional analysis correlating the differences in accuracy and RT between implementation and memorization. This was to check if improvements in accuracy were consistently due to a speed response slowdown. The analysis yielded a negative non-significant correlation between both measures (r=-0.257, p=0.154, t=1.458), which argues against a trade-off explanation of the pattern of the data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2. fMRI results

3.2.1. Univariate

First, we aimed to look at mean activation differences between implementation and memorization at the instruction encoding stage. The implementation of instructions (compared to their memorization) increased activation levels in the bilateral fusiform gyrus (left: [-33, -37, -31], *k*=224; right: [33, -37, -31], *k* = 190), postcentral gyrus incurring into supramarginal and Rolandic operculum (right: [48, -19, 17], *k* = 97; left: [-54, -22, 14], *k* = 86), right SMA ([6, -4, 50], *k*=96), thalamus ([-3, -22, 11], *k* = 187) and right precentral gyrus ([36, -13, 62], *k*=186). The inverse contrast, memorization>implementation, only yielded one cluster of activation in the right angular gyrus ([24, -58, 44], *k* = 225).

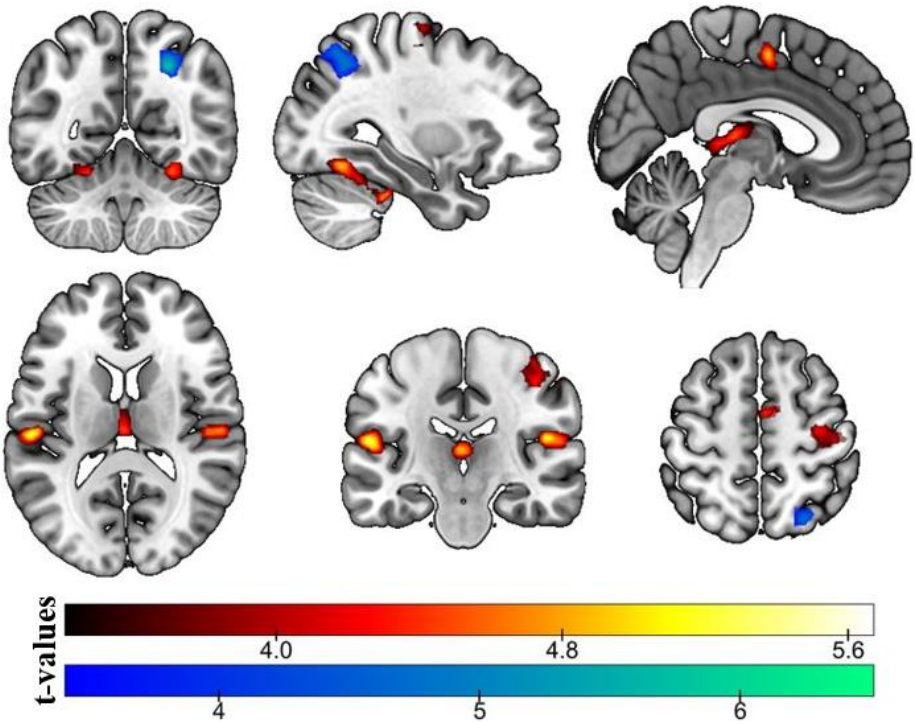

Figure 3. Graphic display of the results of the univariate analysis. Red represents the areas resulting for the implementation>memorization contrast, and blue for memorization>implementation. Color scales reflect peaks of significant t-values (p<.05, FWE corrected for multiple comparisons).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3.2.2. Multivariate results

3.2.2.1. Cross-classification results

As we aimed to explore how task demands impact the way instructed content is represented in neural patterns, we used cross-classification to find commonalities between those two tasks demands, regarding the three instruction remaining variables (Integration of Dimensions, Response Complexity and Target Category). All the results are the product of the two-way train and test scheme explained above (*2.4.2. MVPA: Whole-brain cross-classification*). The results of this analysis reveal areas that display common codes between the implementation and memorization of instructions.

For the integration of dimensions, we trained the algorithm to differentiate among within-dimension and between-dimensions instructions. This analysis revealed a generalizable decoding between tasks in the left middle temporal gyrus (MTG; [-57, -58, -7], k = 234). When using a less strict statistical threshold ($p<0.001$, uncorrected), this revealed a cluster of activation in the inferior frontal gyrus (IFG; [-57, 23, 17], k = 112).

For response complexity, the algorithm classified between instructions with single or sequential responses, yielding significant results in one larger cluster spanning along the left premotor, motor and SMA, incurring in the left parietal lobe and to right premotor areas ([-45, -70, -1], k = 4594). Two smaller clusters were found in the left middle frontal gyrus (MFG [-27, 56, 23], k = 175), and the right supramarginal gyrus (SMG, [54, -28, 38], k = 170).

Last, for target category, we trained the algorithm to differentiate between instructions referring to faces and food, with above-chance accuracy results in the left inferior frontal gyrus (IFG; [-51, 32, 14], k = 1229), the left MTG ([-42, -43, -25], k = 2069), medial superior frontal gyrus (mSFG; [-6, 53, 32], k = 455), and the superior frontal gyrus, medial orbital part ([-3, 50, -22], k = 149).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

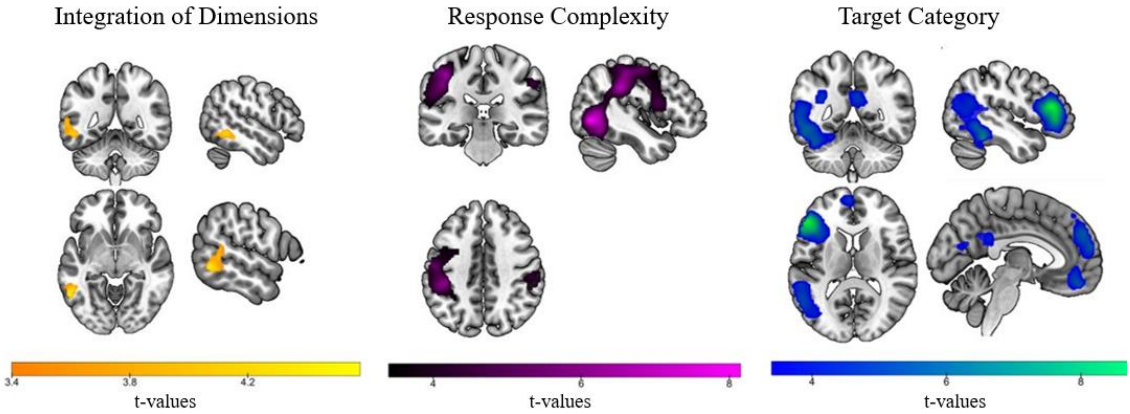

Figure 4. Results of the cross-classification for each of the three models. Color scales reflect peaks of significant t-values (p<.05, FWE corrected for multiple comparisons).

3.2.2.2.ROI-based MVPA results

To investigate if there were differences in pattern organization in the areas that we identified, we compared the decoding accuracy for implementation and memorization. A significant difference in decoding accuracy would indicate a differential impact of task demands on the extent to which the instruction content impacts the observed patterns.

We first used the set of 8 ROIs based on the univariate results (see above), 7 estimated with the contrast implementation>memorization, and 1 with the opposite, memorization>implementation. However, none of the 24 paired t-test (8 ROIs x 3 models: integration, response and target) survived a multiple-comparisons correction (all $p_s > 0.1$ after correction, only 3 $p_s < 0.05$ before correction), indicating a lack of evidence of differential coding in areas with increased univariate activation in each condition.

In contrast, we observed differential accuracy on the ROIs identified with the cross-decoding approach (see Table 2). In all the cases where a t-test was significant, differences were due to a higher decoding accuracy in implementation compared to memorization of instructions (see Figure 4).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Results of contrasting decoding accuracy between implementation and memorization on cross-classification LOSO-estimated ROIs. All *p* values are Bonferroni-corrected for multiple comparisons. *p<0.05.

Model of estimation	ROI	z	p
Integration of dimensions	Left MTG (dim)	-0.551	0.709
Response complexity	Motor cortices, left MFG	2.347	**0.028***
Target category	Left IFG	2.309	**0.021***
	Left MTG (cat)	2.739	**0.009***
	mSFG	0.514	0.304

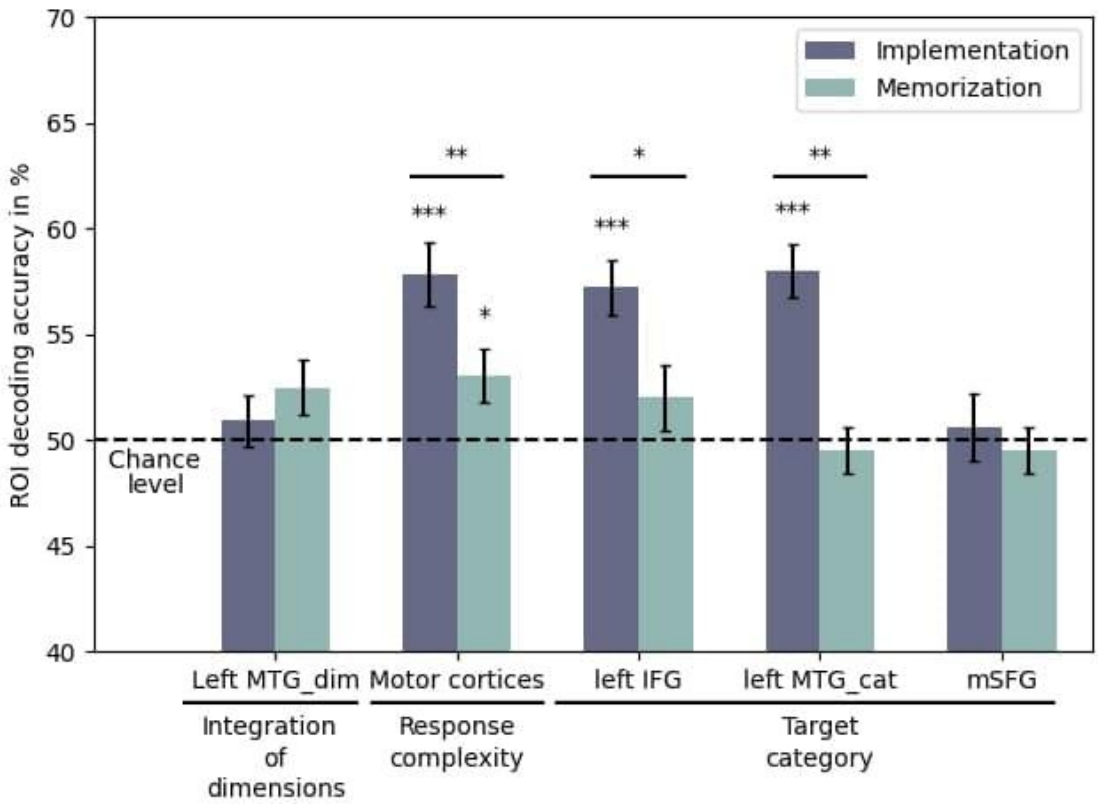

Figure 4. Decoding accuracy for each ROI for each task. Black dashed line represents chance accuracy and the asterisks represent significant *p* values for the paired t test between implementation and memorization tasks (* p < .05, ** p < .01, *** p < .001). Error bars indicate Standard Error of the Mean.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. DISCUSSION

In the present study we examined how task goals impact the neural patterns related to different complex instruction variables. We tested if the strength of neural coding for attributes related to proactive preparation was modulated by specific demands to implement or memorize the instructed content. Using univariate and multivariate analyses we assessed how different subsets of prefrontal and parietal areas are engaged by different task demands, as well as the influence of those tasks on the neural patterns referring to instruction content in those same areas. In addition, with cross-classification analysis we found evidence of brain regions that share a common coding across tasks. Critically, our results showed that, in regions involved in both conditions, the need to implement an instruction induced more decodable patterns of relevant variables compared to memorization trials. Taken together, the current results reveal a complex picture where a subset of specialized frontoparietal areas set and manage novel instruction content with a shared format for implementation and memorization demands, but with different coding strengths.

Strong evidence towards the impact of task demands on instructed performance was already found at the behavioral level. Accuracy and RT scores revealed that executing an instruction is faster but less accurate than memorizing it. This effect could reflect a better preparatory state (faster) in terms of automatic task-set formation and intention-based reflexivity (Liefooghe, Wenke, & De Houwer, 2012) for the implementation task, whereas such preparatory reflexive responses might take place to a lesser extent in the memorization task. A compatible explanation could be that implementation allows to stablish precise task-sets, based on specific action plans to apply in the response grid, whereas in memorization, action plans are less specific and difficult to prepare in an anticipatory manner (for instance, given that response mappings changed on every trial). One could argue that our behavioral result could also be explained by a speed-accuracy trade-off between the two tasks, but we overruled this possibility with a complementary analysis showing that improved accuracy in memorization was not explained by the response slowdown. More importantly, our behavioral results evidenced that task demands modulated the effect of the three proactive-control related instructions manipulations. In line with our predictions, we found a larger effect of these variables on trials where the instructions were implemented. It is important to note that even if participants did not have to explicitly integrate characteristics pertaining to different

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dimensions into one single representation, the task-set had to include those characteristics for participants to respond correctly to the instructions. The term *integration* hence refers not to an integration of different stimuli per se, but to the combined consideration of characteristics coming from different dimensions of stimuli in the service of the same task-set. Overall, the behavioral pattern of results suggests that the three variables manipulated across our instructions pool were behaviorally relevant for the instructions' execution –to a greater extent than declarative maintenance-, which is also consistent with some of the results obtained from neuroimaging data.

The univariate fMRI results showed a clear distinction between the implementation and memorization of a given instruction during the encoding stage, in line with previous studies addressing these two processes (Bourguignon et al., 2018; Muhle-Karbe et al., 2017) and complex instruction following (González-García et al., 2017). Pre-activations during encoding of to-be-implemented instructions involved a set of different areas related to different dimensions of proactive preparation, such as stimulus category (fusiform gyrus), or response preparation (premotor and motor areas). The explanation to such preparation could be the greater necessity, compared to the memorization condition, to organize in an anticipated manner the task-set to carry out the action described by the instruction. These preactivations and pattern organization during implementation demands have been at least partially described in previous investigations (González-García et al., 2017; Palenciano, González-García, Arco, Pessoa, et al., 2019). However, compared to previous investigations, our study offers new insight in this regard, given the direct comparison to memorization demands. As such, the activity found in the thalamus could reflect different processes, like a greater need for error-related control (Ide & Li, 2011) in a more demanding task like implementation, or the intent to minimize mistakes while responding. It could also reflect a finer motor control in implementation, necessary for differential preparation for an upcoming single or a sequential response (Prevosto & Sommer, 2013), which would not be needed in memorization (in this case, the response required was always of single type). Another plausible explanation for the involvement of the thalamus during implementation could be the necessity of maintenance of the current task-set representations and integration between different information (motor, perceptual and abstract) present in different brain regions (Wolff & Vann, 2019), which would be less crucial for the memorization task.

On the other hand, the increased activity found in the memorization condition was linked

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to the angular gyrus. Previous research in the field of memory has shown the involvement of this region in declarative memory (Noonan, Jefferies, Visser, & Lambon Ralph, 2013; Seghier, 2013). In the case of instructions, it could indicate recollection of information in a declarative format (Liefooghe, De Houwer, & Wenke, 2013) a previous necessary step to construct a task-set, with this recollection being produced in both tasks but with the subsequent task-set construction being absent in memorization. The lack of an incoming task-set formation would make the recollection the final step in the memorization encoding process, yielding the observed differences when compared with a full task-set formation process in implementation. In turn, the decreased activity in the angular gyrus during implementation could be explained by smaller relevance of the declarative format in implementation, which is consistent with recent neuroimaging studies (González-García et al., 2021). The angular gyrus has also been related to semantic processing and verbal working memory (Seghier, 2013), and, interestingly, it has been specifically linked to retrieval of rule-based schemas and their constituting components (Wagner et al., 2015). Tentatively, our results could reflect a complex instruction building from its simpler constituents in a prior state to execution, prior to the task-set formation, which is in line with previous studies of rule compositionality (Reverberi, Görgen, & Haynes, 2012) and co-activation of simpler rules to perform a complex one (Deraeve, Vassena, & Alexander, 2019).

To identify regions with generalizable neural codes between implementation and memorization conditions, we performed cross-classification analysis. We found above chance pattern decodability under implementation and memorization contexts in a broad extension of frontoparietal regions. In general terms, when cross-classification is successfully performed between tasks or cognitive contexts this suggests that the resulting areas share a common information code, reflected in neural patterns (Kaplan et al., 2015). To ensure avoiding a potential bias driven by response differences between implementation and memorization (given that participants used different fingers for each one), we discarded the use of an MVPA approach to identify regions with differential activity patterns for the two task demands, and employed cross-classification analysis instead. We also chose to only use cross-classification given that implementation and memorization demands could lead to differential information encoding in neural patterns due to the proceduralization of the instructed content in the implementation condition (Muhle-Karbe et al., 2017), which would give an overrepresentation of regions in one

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

condition above the other.

When concurrent deployment of prefrontal and parietal areas is found during task-set or rule representation, we refer to the implication of the Multiple-Demand Network (MDN; Duncan, 2010), broadly related to task-set implementation (Dosenbach et al., 2006) and found in tasks where different rules, stimuli and responses must be represented (Woolgar, Jackson, & Duncan, 2016). Our data provide evidence for the existence of a common code between implementation and memorization of instructions in the MDN, which may seem incompatible with previous research (Bourguignon et al., 2018; Demanet et al., 2016). However, we argue that these results are not contradictory but complementary to ours. Overall, our results suggest that both processes share a common code, but also that the coding strength in both situations was sufficiently different to tell them apart. Similar pattern commonalities in brain areas between in implementation and memorization have been also described in previous studies, revealing a certain degree of similarity in frontoparietal activity in implementation and memorization conditions (Muhle-Karbe et al., 2017). One central aspect in the instruction-following literature is the transformation of the declarative information carried in the instruction into a proceduralized action task-set (Demanet et al., 2016; Hartstra, Kühn, Verguts, & Brass, 2011; Muhle-Karbe et al., 2017), which could lead to expect low cross-classification between implementation and memorization. Nonetheless, this transformation is thought to primarily take place in the so-called preparation stage (Brass et al., 2017; Liefooghe et al., 2013). Because our window of analysis focused on the preceding instruction encoding stage to avoid decoding confounds due to perceptual stimulation between conditions (see below for further discussion), it is reasonable to think that this information transformation is milder and both codes are more similar than if we had focused the analysis during this preparation stage. However, this approach allowed us to define ROI to test that coding difference avoiding undesired confounds.

One of our hypotheses, based on a previous study (Palenciano, González-García, Arco, Pessoa, et al., 2019), predicted neural patterns decodability in the IFG when integration of information coming from the same or between different dimensions is needed. Contrary to our expectations, this hypothesis was not supported when using a strict statistical correction (nonetheless, it is worth mentioning that this pattern was noticeable in group results with a lesser strict statistical threshold). Moreover, we found differences in pattern strength coding in the MTG, when people integrated information about the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

same or different information dimensions. Interestingly, prior studies focusing on retrieval control of semantic content had described a relationship between the left IFG and the MTG where both areas work coordinately in semantic demanding tasks (Davey et al., 2016). In our case, this could in fact reflect a controlled retrieval of semantic information and its integration to correctly perform the action specified by the instruction. It could also be the case that the absence of the IFG in information integration could be partially driven by having dimensions and categories that may have not generated enough separation between classes to require the involvement of this region in the integration of their characteristics. Unfortunately, given that the result of the IFG activation did not survive statistical correction, the results from such analysis should therefore be treated with caution.

The variable of response complexity, on the other hand, could be decoded from pre/post central gyrus, SMA, and left MFG, including a noticeable part of the IFG. This result is relevant due the fact that the IFG, in association with motor and premotor regions, has been shown to support actions sequences (Koechlin & Jubault, 2006; Van Opstal, Fias, Peigneux, & Verguts, 2009). Our manipulation of response complexity also involves sequentiality, as a complex response is sequential by definition. Such sequentiality could be a potential source for this pattern decodability in the IFG. Future studies could introduce a manipulation to separate complexity from sequentiality and thus tease apart their specific effects in the preparatory patterns generated by complex verbal instructions.

Finally, the ROI analyses complemented the aforementioned ones, to search if, even if implementation and memorization share a common code in some regions, there are differences in the strength of the coding between conditions. These results yielded a complex picture where implementation and memorization demands had differential coding strength in those regions where different instruction content (information integration, response complexity and stimuli category) is processed. Of the five ROIs tested, we found statistical differences in three of them. This pattern of results partially supports the hypothesis that the need to implement an instruction influences the neural representation of that instruction. One possible explanation for this pattern of results is the difference in nature between implemented (proceduralized) and memorized (non-proceduralized) instructions. The transformation of the information into a task-set in implementation could lead to clearly distinguishable representations in this condition, while in memorization, the semantic format would lead to less distinct representations,

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which would be more difficult to differentiate. More subtle differences in memorization would need more statistical power to be detected, and it could be that the statistical power achieved in our design was enough to detect the larger effects in implementation but insufficient for the smaller ones in the memorization condition. Nonetheless, with our results we can conclude that both instruction-following and memorization depend on a broad set of frontoparietal regions corresponding to the MDN, and both share an early common information code but with meaningful differences in their neural patterns.

It is worth noting that in this study, both univariate and multivariate analysis were performed at the encoding stage, while participants read the instruction, thus, minimizing possible confounds yielded by differences during the response stage. This could be one key point in explaining some of the observed differences with prior studies characterizing the procedural (implementation) and declarative (memorization) processing of instructions. Most of these studies focused on the preparation stage, right after instruction coding and before the motor response (Bourguignon et al., 2018; González-García et al., 2017; Muhle-Karbe et al., 2017), given that differences between both tasks are expected to be maximum in this stage (Liefooghe et al., 2013). Our paradigm, however, did not allow to explore this preparation stage, because of the major differences in visual stimulation at the target stage, which probably encouraged different preparation strategies. Nonetheless, it would be interesting to explore this possibility in a follow-up study, with a paradigm optimized to extrapolate our findings to such preparation stage.

In sum, the present work examined the extent to which implementing and memorizing novel instructions share brain areas and how these two task goals influence the neural patterns referring to novel instructions. We show for the first time how these two processes can have differential impact on the neural representation of the instruction content in a broad set of frontoparietal areas. We also demonstrate that, during the encoding of novel instructions, these areas share to certain extant a common code between the two tasks. Moreover, we show how flexible coding organizes brain activity based on different relevant instruction dimensions, and how this organization might be encoded to a greater extent during implementation demands. Altogether, these findings help to further clarify the neural mechanisms underlying instruction processing and represents a step forward to better characterize the implementation process.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Funding

This work was supported by the Spanish Ministry of Science and Innovation (PID2019-111187GB-I00 to M.R., and IJC2019-040208-I to C.G.G), the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (Ref. 835767, to C.G.G.). and the Spanish Ministry of Education, Culture and Sports (FPU17/01627 to A.S.)

Acknowledgments

This research is part of A.S.'s activities for the Psychology Graduate Program of the University of Granada.

CRediT author statement

Alberto Sobrado: Methodology, Formal analysis, Investigation, Visualization, Writing - original draft, Writing - review & editing.

Ana F. Palenciano: Conceptualization, Methodology, Software, Resources.

Carlos Gonzalez-García: Conceptualization, Methodology, Writing - review & editing.

María Ruz: Conceptualization, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare no competing financial interests.

Data availability

Public archiving of anonymized individual data is not permitted by the ethics approval. Readers seeking access to the data should contact the corresponding author via email. Data will be released to interested researchers with approval from the local research ethics committee. Unthresholded maps corresponding to group crossclassification results are available at https://neurovault.org/collections/10326/. Digital study materials are available at https://osf.io/5yt4h. The analysis were carried out with SPM, The Decoding Toolbox (Hebart, Görgen, & Haynes, 2015) and JASP, with no custom code. No part of the study procedures was preregistered. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Arco, J. E., González-García, C., Díaz-Gutiérrez, P., Ramírez, J., & Ruz, M. (2018). Influence of activation pattern estimates and statistical significance tests in fMRI decoding analysis. *Journal of Neuroscience Methods*, *308*, 248–260. https://doi.org/10.1016/j.jneumeth.2018.06.017

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

Brainard, D. H. (1997). The Psychophsycis Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1017/CBO9781107415324.004

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience and Biobehavioral Reviews*, *81*, 16–28. https://doi.org/10.1016/j.neubiorev.2017.02.012

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 and Behavioral Neuroscience*, *13*(1), 1–22. https://doi.org/10.3758/s13415-012-0125-7

Cole, M. W., Patrick, L. M., Meiran, N., & Braver, T. S. (2018). A role for proactive control in rapid instructed task learning. *Acta Psychologica*, *184*, 20–30. https://doi.org/10.1016/j.actpsy.2017.06.004

Davey, J., Thompson, H. E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., … Jefferies, E. (2016). Exploring the role of the posterior middle temporal gyrus in semantic cognition: Integration of anterior temporal lobe with executive processes. *NeuroImage*, *137*, 165–177. https://doi.org/10.1016/j.neuroimage.2016.05.051

Demanet, J., Liefooghe, B., Hartstra, E., Wenke, D., De Houwer, J., & Brass, M. (2016). There is more into 'doing' than 'knowing': The function of the right inferior frontal sulcus is specific for implementing versus memorising verbal instructions. *NeuroImage*, *141*, 350–356.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

https://doi.org/10.1016/J.NEUROIMAGE.2016.07.059

Deraeve, J., Vassena, E., & Alexander, W. (2019). Conjunction or co-activation? A multi-level MVPA approach to task set representations. *BioRxiv*, 521385. https://doi.org/10.1101/521385

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., … Petersen, S. E. (2006). A Core System for the Implementation of Task Sets. *Neuron*, *50*(5), 799–812. https://doi.org/10.1016/j.neuron.2006.04.031

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

Esterman, M., Tamber-Rosenau, B. J., Chiu, Y. C., & Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage*, *50*(2), 572–576. https://doi.org/10.1016/j.neuroimage.2009.10.092

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

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

González-García, C., Formica, S., Wisniewski, D., & Brass, M. (2021). Frontoparietal action-oriented codes support novel instruction implementation. *NeuroImage*, *226*, 117608. https://doi.org/10.1016/j.neuroimage.2020.117608

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

Haxby, J. V, Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of*

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Neuroscience, *37*, 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*, 88. https://doi.org/10.3389/fninf.2014.00088

Ide, J. S., & Li, C. R. (2011). A cerebellar thalamic cortical circuit for error-related cognitive control. *NeuroImage*, *54*(1), 455–464. https://doi.org/10.1016/j.neuroimage.2010.07.042

Kaplan, J. T., Man, K., & Greening, S. G. (2015). Multivariate cross-classification: Applying machine learning techniques to characterize abstraction in neural representations. *Frontiers in Human Neuroscience*, *9*, 1–12. https://doi.org/10.3389/fnhum.2015.00151

Koechlin, E., & Jubault, T. (2006). Broca's Area and the Hierarchical Organization of Human Behavior. *Neuron*, *50*(6), 963–974. https://doi.org/10.1016/j.neuron.2006.05.017

Liefooghe, B., & De Houwer, J. (2017). Automatic effects of instructions do not require the intention to execute these instructions. *Journal of Cognitive Psychology*, *30*(1), 1–14. https://doi.org/10.1080/20445911.2017.1365871

Liefooghe, B., De Houwer, J., & Wenke, D. (2013). Instruction-based response activation depends on task preparation. *Psychonomic Bulletin and Review*, *20*(3), 481–487. https://doi.org/10.3758/s13423-013-0374-7

Liefooghe, B., Wenke, D., & De Houwer, J. (2012). Instruction-based task-rule congruency effects. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(5), 1325–1335. https://doi.org/10.1037/a0028148

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

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, *103*, 130–138. https://doi.org/10.1016/j.neuroimage.2014.09.026

Noonan, K. A., Jefferies, E., Visser, M., & Lambon Ralph, M. A. (2013). Going beyond

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

inferior prefrontal involvement in semantic control: evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *Journal of Cognitive Neuroscience*, *25*(11), 1824–1850. https://doi.org/10.1162/jocn_a_00442

Palenciano, A. F., González-García, C., Arco, J. E., Pessoa, L., & Ruz, M. (2019). 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. (2019). Transient and Sustained Control Mechanisms Supporting Novel Instructed Behavior. *Cerebral Cortex*, *29*(9), 3948–3960. https://doi.org/10.1093/cercor/bhy273

Prevosto, V., & Sommer, M. A. (2013). Cognitive control of movement via the cerebellar-recipient thalamus. *Frontiers in Systems Neuroscience*, *7*, 56. https://doi.org/10.3389/fnsys.2013.00056

Reverberi, C., Görgen, K., & Haynes, J. D. (2012). Compositionality of rule representations in human prefrontal cortex. *Cerebral Cortex*, *22*(6), 1237–1246. https://doi.org/10.1093/cercor/bhr200

Ruge, H., Jamadar, S., Zimmermann, U., & Karayanidis, F. (2013). The many faces of preparatory control in task switching: Reviewing a decade of fMRI research. *Human Brain Mapping*, *34*(1), 12–35. https://doi.org/10.1002/hbm.21420

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

Sakai, K. (2008). Task Set and Prefrontal Cortex. *Annual Review of Neuroscience*, *31*(1), 219–245. https://doi.org/10.1146/annurev.neuro.31.060407.125642

Seghier, M. L. (2013). The angular gyrus: Multiple functions and multiple subdivisions. *Neuroscientist*, *19*(1), 43–61. https://doi.org/10.1177/1073858412440596

Stocco, A., Lebiere, C., O'Reilly, R. C., & Anderson, J. R. (2012). Distinct contributions of the caudate nucleus, rostral prefrontal cortex, and parietal cortex to the execution of instructed tasks. *Cognitive, Affective and Behavioral Neuroscience*, *12*(4), 611–628. https://doi.org/10.3758/s13415-012-0117-7

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.08.434338; this version posted June 16, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., … Nelson, C. (2009). The NimStim set of facial expressions: Judgements from untrained research participants. *Psychiatry Research*, *168*(3), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006.The

Van Opstal, F., Fias, W., Peigneux, P., & Verguts, T. (2009). The neural representation of extensively trained ordered sequences. *NeuroImage*, *47*(1), 367–375. https://doi.org/10.1016/j.neuroimage.2009.04.035

Wagner, I. C., van Buuren, M., Kroes, M. C. W., Gutteling, T. P., van der Linden, M., Morris, R. G., & Fernández, G. (2015). Schematic memory components converge within angular gyrus during retrieval. *ELife*, *4*. https://doi.org/10.7554/eLife.09668.001

Westfall, J. (2015). PANGEA: Power ANalysis for GEneral Anova designs. https://doi.org/10.1017/CBO9781107415324.004

Wolff, M., & Vann, S. D. (2019). The cognitive thalamus as a gateway to mental representations. *Journal of Neuroscience*, *39*(1), 3–14. https://doi.org/10.1523/JNEUROSCI.0479-18.2018

Woolgar, A., Jackson, J., & Duncan, J. (2016). Coding of visual, auditory, rule, and response information in the brain: 10 years of multivoxel pattern analysis. *Journal of Cognitive Neuroscience*, *28*(10), 1433–1454. https://doi.org/10.1162/jocn_a_00981