TESIS DOCTORAL

Programa de Doctorado en Psicología

Bases neurales de la implementación de instrucciones

Doctoranda

Ana Francisca Palenciano Castro

Directora

María Ruz

UNIVERSIDAD DE GRANADA

Departamento de Psicología Experimental

Octubre de 2019

Editor: Universidad de Granada. Tesis Doctorales
Autor: Ana Francisca Palenciano Castro
ISBN: 978-84-1306-394-2
URI: http://hdl.handle.net/10481/58386

Contents

List of figures and tables

Chapter 1:

INTRODUCTION

1. INTRODUCTION

Many animal species display highly complex behavioral repertoires that adjust harmoniously to their ecological niche. These patterns have been crafted by countless evolution episodes of natural selection. Slow adaptations allow animals to be highly successful in their stable, well-known environment. Humans also display behaviors acquired through slow evolutionary adaptations, in addition to slow trial and error learning. However, our species excels in novel scenarios, which pose challenges that require rapid adaptations. Humans, crucially, can make use of a powerful resource in such circumstances: instructions make situational key elements explicit and allow their communication among equals and through generations (Cole, Laurent, & Stocco, 2013). The ability to act according to instructions establishes a sharp distinction among us and other apes and constitutes a key aspect of our flexible adaptation to changing environments. Consequently, understanding how this complex behavior is implemented in the brain is of high relevance for Cognitive Neuroscience. The present thesis is composed by three studies that employed functional Magnetic Resonance Imaging (fMRI) and Electroencephalographic (EEG) recordings to investigate the neural processes that sustain our capacity to implement tasks at the first try by using verbal instructions.

In the first investigation, we described how novel complex verbal instructions engage distinct neural networks with different temporal profiles. Instruction following relies on cognitive control, which refers to the set of high-level processes that allow goal-oriented behavior when automatic routines do not lead to the desired outcome (Norman & Shallice, 1986). Neuroimaging techniques, especially functional magnetic resonance imaging (fMRI), have been key for

uncovering the brain regions related to control processes. During the last decades, the increasing sophistication of both experimental design (Petersen & Dubis, 2012) and analysis techniques (Dosenbach et al., 2007) have led to the identification of two networks of frontal and parietal regions which implement control by acting at different, transient and sustained, timescales (Dosenbach et al., 2007). Nonetheless, despite the relevance of flexibility and novelty as core attributes of controlled processing (Norman & Shallice, 1986), the majority of experimental settings employed are quite rigid and involve the repetitive implementation of a few simple rules. This had left unaddressed the relevance of the two control networks in novel and variable scenarios. Consequently, one of the goals of this work was to extend this dual framework to flexible instructed behavior, studying its sustained and transient profiles of activations.

In the second and third studies, we employed fMRI and EEG recordings, together with novel pattern analyses, to investigate the neural representation of the multidimensional content of novel instructions. Further conceptualizations of control have highlighted the existence of two differentiated modes of control. On the one hand, proactive control acts in anticipation, preparing our system for upcoming demands. On the other hand, reactive mechanisms act in an online fashion during performance, enabling quick adjustments upon the appearance of unexpected changes, conflicting information or errors. Although both processes operate in a coordinated fashion during demanding situations, proactive ones are of special relevance in novel instructed behavior (Cole, Patrick, & Braver, 2018). In this sense, proactive control generates mental models (representations) of the upcoming task, containing information about target stimuli, relevant responses, and rules linking both, together with expected outcomes. These high-level

representations, also known as task-sets (Sakai, 2008), ultimately orchestrate our actions by biasing the processing in relevant perceptual and motor systems (Miller & Cohen, 2001). Crucially, the implementation of instructions requires their translation into effective task sets. However, the flexible neural mechanisms allowing task-set reconfiguration in complex, multidimensional novel contexts are uncertain (Brass, Liefooghe, Braem, & De Houwer, 2017; Cole, Laurent, et al., 2013). A second aim of this work was assessing these processes. In doing so, we focused on the nature of novel task representations, addressing how their constituent dimensions organize neural activity, and the temporal dynamics of the underlying mechanisms.

In the pages that follow, we provide a general overview of the theoretical background scaffolding the research presented in the thesis. We start by briefly describing the main current perspectives about neural implementation of cognitive control. Then, we discuss the emerging studies addressing novel instructed behavior, emphasizing its proactive component and its correspondence with general control models. In addition, we highlight the relevance of motivation on these proactive control processes.

1.1. Two brain networks supporting cognitive control

In the last two decades, multiple attempts have been made to understand the brain underpinnings of cognitive control. One of the main and most robust findings so far is the existence of a wide fronto-parietal network that is consistently recruited by many demanding task contexts: conflict resolution, task switching, error processing, problem solving, fluid intelligence tasks, and so on (Dosenbach et al., 2006; Duncan, 2010; Fedorenko, Duncan, & Kanwisher, 2013). This network, named as Multiple Demand (MDN, Duncan, 2010), includes the

dorsolateral prefrontal cortex (DLPFC), specifically, the inferior frontal gyrus (IFG), sulcus (IFS) and junction (IFJ), the rostrolateral prefrontal cortex (RLPFC), the dorsal anterior cingulate cortex (dACC), the intraparietal sulcus (IPS) and pre-supplementary motor area (preSMA).

Despite the frequent coactivation of these areas, it seems biologically implausible that all the nodes within the MDN perform the same computations. Further research collapsing across multiple datasets (Dosenbach et al., 2006) and also assessing communication dynamics among areas (Dosenbach et al., 2007), differentiated two components within the MDN (Dosenbach, Fair, Cohen, Schlaggar, & Petersen, 2008). On the one hand, a network anchored at the dACC, the frontal operculum and the RLPFC (the cingulo-opercular network, or CON) displays anticipatory (i.e., cue-locked) and sustained activations, consistent with the preparatory reconfiguration of task-sets and their maintenance, respectively. On the other hand, a network composed mainly by the DLPFC and the IPS (the fronto-parietal network, or FPN), shows transients activations especially during conflict and error processing, and thus, is potentially more related to reactive adjustments of behavior.

While this is an appealing and straightforward scenario, inconsistent evidence has also been found. For example, different subsections of the LPFC the IPS, both part of the reactive component, have been linked to anticipatory task setting (Sakai, 2008). Other studies have assigned to the CON a more general role in salience detection (Seeley et al., 2007) or tonic alertness (Sadaghiani & D'Esposito, 2015). Thus, it seems that these regions show functional patterns not always consistent with the Dual-Network Model of Control (Dosenbach et al.,

2008). In Chapter 3, we include a detailed review of this topic, together with potential interpretations of the contradictions found in the literature.

1.2. Proactive neural processes during novel instructed-behavior.

Multiple behavioral studies show that instructed-behavior relies on proactive control to a higher extent than on reactive control (Cole, Braver, & Meiran, 2017; Cole et al., 2018; Meiran, Pereg, Kessler, Cole, & Braver, 2015). To explore the brain regions supporting this proactive preparation, recent studies have analyzed brain activation patterns during instruction encoding and preparation. For this, the studies analyze the time interval between the instruction and the target, and frequently compare novel rules against practiced ones (Cole, Laurent, et al., 2013).

Congruent with the expectations based on behavioral findings, novel instructions engage MDN areas to a higher extent than practiced rules (Cole, Bagic, Kass, & Schneider, 2010; Hartstra, Waszak, & Brass, 2012; Ruge & Wolfensteller, 2010). These activations are mostly constrained within the FPN, involving different portions of the LPFC and the IPS (Demanet et al., 2016; González-García, Arco, Palenciano, Ramírez, & Ruz, 2017; Hartstra, Kühn, Verguts, & Brass, 2011; Hartstra et al., 2012; Ruge & Wolfensteller, 2010). Importantly, it is the intention to implement the instruction which triggers the involvement of these regions (Demanet et al., 2016; González-García et al., 2017; Muhle-Karbe, Duncan, De Baene, Mitchell, & Brass, 2017). This supports their proactive functional role over other processes, such as maintaining verbal content in working memory.

The peaks of encoding-locked activations across studies seem to follow a rostro-caudal gradient. Most abstract and complex instructions engage more anterior

portions of the LPFC up to the RLPFC (Cole et al., 2010), whereas more concrete rules shift the activity toward posterior prefrontal locations (Hartstra et al., 2012). Sensorio-motor preparation, on the other hand, engages the IPS (Hartstra et al., 2011, 2012). This dissociation among frontal and parietal cortices has also been shown in practiced task contexts (Brass, Cramon, Yves, & Abstract, 2004; De Baene & Brass, 2014; Muhle-Karbe, Andres, & Brass, 2014). More broadly, these results resonate with general models about brain organization, which postulate a similar abstraction gradient regarding the information being processed in the frontal lobe (Koechlin, Ody, & Kouneiher, 2003). Overall, it seems plausible that novel task preparation is a distributed process involving mainly FPN nodes, each contributing at different abstraction levels. It is important to note, finally, the disagreement between these findings and the predictions from the dual framework exposed before (Dosenbach et al., 2007), which would also assign a central role to CON regions in sustained and proactive preparation in novel contexts.

Anticipatory activity is not limited to control-related regions: it has also been found in lower level, perceptual and motor regions. That is the case of primary and premotor cortices, primary and secondary somatosensory cortex, or perceptual regions as the fusiform gyrus (e.g.: Cole et al., 2010; Hartstra et al., 2011; Hartstra, Waszak, & Brass, 2012). Notably, these preactivations happen in the absence of actual stimuli or specific motor preparation or execution. These findings could be interpreted from proposals conceptualizing cognition as continuous loops among perceptual and action-related neural systems (Fuster, 2004). Thus, iterative communication among the above-mentioned FPN nodes and perceptual and motor cortices during preparation could have, as a result, a

sharpening of these areas tuning to the incoming targets and required responses, a mechanism that could be crucial for the flexible and quick implementation of novel instructions.

The role of sustained control activations in the context of novel instructed tasks is not well understood nowadays. This could be caused, at least in part, by the difficulties inherent to the extraction of simultaneous sustained (block) and transient (event-related) brain signals. This endeavor requires the employment of hybrid or mixed designs (Petersen & Dubis, 2012), which are complex paradigms that mix block and event-related ingredients (Visscher et al., 2003). Currently, there is only one published experiment that studied sustained activations during novel instructed behavior (Dumontheil, Thompson, & Duncan, 2011). Intriguingly, this study found sustained activations in a mixture of FPN and CON nodes, involving the right IFS and left RLPFC, and again, regions not directly related to cognitive control, such as the dorsal PFC (outside de MDN) and the medial occipital cortex. This puzzling evidence, which also coincides with findings outside instructed behavior, highlights the necessity to expand Dosenbach and colleagues' framework (2007, 2008) to complex and novel task performance.

Overall, while the studies addressing the neural basis of novel instructed behavior have grown in the last decade, there is still a considerable gap between the findings obtained so far and broader cognitive control models such as the one proposed by Dosenbach and colleagues (Dosenbach et al., 2008), or Braver (Braver, 2012). This could be caused by the different nature of the processes involved in novel and practiced rule following, or also by differences at the methodological level. To shed some light upon this issue, we carried out Study 1

(Chapter 4), which was designed to test the dual model of control in the context of novel instruction following.

1.3. Exploring Novel task representation: Content and structure.

The first efforts to study instruction following were highly informative about the brain regions underlying this behavior. Nonetheless, the specific computations performed by each area remained poorly understood. For example, what is the interpretation of the increased activity in the LPFC during instruction encoding? Does it reveal the representation of specific task parameters? Or is it due to more general processes needed for implementing stimulus-response associations? (Bourguignon, Braem, Hartstra, De Houwer, & Brass, 2018). This indeterminacy is partly due to the activation-based perspective adopted by previous studies, in which brain signals are averaged across neighbor spatial locations (or voxels), and the difference between the mean activation across experimental conditions is used to extract conclusions. This univariate approach treats regions as a whole to decide whether or not they are involved in a contrast of interest. As a consequence, functional differences not as broadly distributed are lost. Also, and more importantly, it easily leads to situations (as in the example above) where the same pattern of activation can be linked to equally probable roles. In this sense, univariate strategies limit the scope of our understanding of instructed behavior.

Recent analysis techniques from machine learning and computational sciences have pushed a shift towards an information-based framework (Haxby, Connolly, & Guntupalli, 2014; Kriegeskorte, Goebel, & Bandettini, 2006), which exploits the distributed activity patterns across voxels, instead of averaging them. This perspective expands the empirical questions that can be addressed, as it provides a window into the finer-grained representations encoded in these patterns.

Hence, multivariate approaches are of high relevance for addressing the representational nature of novel task-sets.

The most popular technique in the literature is Multivoxel Pattern Analysis (MVPA; Haxby et al., 2014), where classifiers are trained to distinguish the patterns of activity of two or more conditions. Later, these trained classifiers are tested with novel data. If the prediction is performed above chance levels, it is inferred that information about the conditions is readable from these patterns and thus represented in a certain area. This analysis approach has been used to explore whether, and where in the brain, relevant novel task information is represented during instruction encoding (Bourguignon et al., 2018; González-García et al., 2017; Muhle-Karbe et al., 2017), which would indicate the presence of specific preparatory mechanisms. Indeed, the instructed target category can be decoded from multiple MDN areas during the initial encoding stage and throughout the preparation interval, before stimulus presentation (González-García et al., 2017; Muhle-Karbe et al., 2017). Importantly, a recent study showed that when the instructions were only memorized, the information in control areas faded across the preparation interval, ruling out that successful instruction decoding was based on linguistic or semantic rule representations held in working memory (Muhle-Karbe et al., 2017). Moreover, the quality of the representation, quantified by the classifier's precision, is robustly correlated with behavioral performance, with higher correct rates and faster responses associated with higher classifications accuracies (Cole, Ito, & Braver, 2016; González-García et al., 2017; Muhle-Karbe et al., 2017). All the evidence, thus, converges in that these fronto-parietal regions are activating a control task-set proactively to guide behavior at the first attempt with an instruction.

The results so far point toward the unequal participation of MDN nodes in task-set representation. Among all, the IFS and middle frontal gyrus (MFG; Bourguignon et al., 2018) seem to be key regions for this function. These areas are the one showing novel instruction encoding most consistently across different studies, representing the relevant target category and also the logical rules linking stimulus and motor responses (Cole, Etzel, Zacks, Schneider, & Braver, 2011; Cole et al., 2016). These findings converge and further characterize anticipatory activations, and coincide as well with results showing practiced rule encoding in the IFS/MFG across multiple datasets (Waskom, Kumaran, Gordon, Rissman, & Wagner, 2014; A. Woolgar, Hampshire, Thompson, & Duncan, 2011; Alexandra Woolgar, Jackson, & Duncan, 2016). They also fit with classic theoretical models such as the Guided Activation Theory (Miller & Cohen, 2001) which conceptualize the LPFC as the source of top-down bias scaffolding goal-directed behavior. This all leads to the question about the nature of *representational architecture* that allows such malleable and quick task-set reconfigurations.

Some attempts have been made at exploring the organizational principles that govern LPFC novel rule representation. Again, novel analysis approaches have helped to enlighten this issue. Using cross-classification, a technique based on MVPA which assesses if information coding is generalizable across different experimental conditions, Cole and colleagues (2011) found that novel instructions were represented in the LPFC reusing some of the same neural representations (i.e., multivoxel activation patterns) employed for practiced rules. This points to the presence of compositionality in LPFC, by which simple rule components are combined to generate novel task representations (Cole, Laurent, et al., 2013). In line with this result, compositional coding has also been found in this brain area

in other contexts (Pischedda, Görgen, Haynes, & Reverberi, 2017; Reverberi, Gorgen, & Haynes, 2012; Reverberi, Görgen, & Haynes, 2012).

Another crucial approach to study the information represented in brain patterns is Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008), a technique developed to characterize the geometry of the encoding spaces. RSA focuses on the relationships among multiple conditions or stimuli, quantifying with correlations how similar their corresponding activity patterns are. The pair-wise array of similarities constitutes an estimation of the representational space, which abstracts from specific patterns and focuses on their representational organization. Returning to the issue of compositionality described above, Cole, Reynolds, et al. (2013) combined RSA with functional connectivity data, and found that the patterns of communication of the FPN with other brain networks (visual, motor, dorsal and ventral attention, default mode) were highly structured by novel task parameters. Specifically, the connectivity patterns established during novel instruction execution resulted from a combination of the patterns corresponding to the individual semantic, motor and logical rules composing the task at hand. This finding does not only support the compositional coding account, but it also emphasizes the distributed nature of task sets. Nonetheless, both studies (Cole et al., 2011; Cole, Reynolds, et al., 2013) anchored the analyses at the implementation stage, where the corresponding logical operations were performed on the targets, leading to motor execution. In addition, these studies did not track the specific informational patterns of each novel instruction, leaving the representational codes of the areas involved rather unexplained.

Motivated by this scenario, we aimed to explore the encoding organization according to different dimensional axes for novel instructions during the

encoding phase, to ensure that proactive-related representations were being captured. Study 2 (Chapter 5) and 3 (Chapter 6) were carried out to fulfill this goal.

1.4. Cognitive control – motivation interplay.

For many decades, theoretical models have treated cognitive control as a function isolated from affective and motivational factors. Nonetheless, the experimental settings employed in the laboratory contrast with our daily life, where control adjustments operate on environments loaded with value (Botvinick & Braver, 2015; Pessoa, 2017). In this sense, motivation mobilizes and energizes our behavior towards desired outcomes, prioritizing certain goals among others. Congruently, the classical observation is that the expectation of reward improves our performance in a wide variety of task contexts (Botvinick & Braver, 2015).

Initial proposals conceptualized these improvements as a general modulation, similar to an arousal boost. However, further research showed specific effects of motivation upon task processing, improving the efficiency of cognitive control mechanisms (Pessoa, 2009). Several results indicate that incentives boost both proactive and reactive mechanisms, for example, reducing switching costs (Kleinsorge & Rinkenauer, 2012; Shen & Chun, 2011) and increasing anticipatory brain activations (Engelmann, Damaraju, Padmala, & Pessoa, 2009; Krebs, Boehler, Roberts, Song, & Woldorff, 2012), or decreasing behavioral and neural conflict effects (Padmala & Pessoa, 2011). Nonetheless, a mechanistic explanation about how the motivational value is incorporated in control adjustments is still lacking. The current consensus leans towards a proactive-focused account (Chiew & Braver, 2016; Jimura, Locke, & Braver, 2010; Rowe, Eckstein, Braver, & Owen,

2008), by which the expectation of reward critically participates in goal selection and its maintenance.

The impact of motivation upon proactive preparation could take place, however, at different levels. Some results suggest that it plays a role during task-set representation, with reward expectation increasing the quality of the representational encoding in MDN areas, especially in the LPFC (Etzel, Cole, Zacks, Kay, & Braver, 2016). Moreover, in this study, the tuning of fronto-parietal representations was tightly linked to the performance improvement exerted by motivation, supporting the behavioral relevance of this finding. This also coincides with evidence showing that regions from the FP network, specifically the inferior parietal lobe, encode rules-reward associations (Wisniewski, Reverberi, Tusche, & Haynes, 2015). Overall, it is plausible that motivational values operate on control-related regions optimizing task-set reconfiguration. This could initiate a cascade of events culminating in the sharpening of reactive control adjustments and, in general, improved task processing.

The evidence to date, although compelling, is limited to constrained experimental settings where participants alternate among few rules, which have been overly practiced. Motivation-control interactions in more complex and novel task contexts are, consequently, unknown. Because of this, we incorporated economic incentives in Studies 2-3, and addressed the influence of the motivational state on prospective novel task encoding, both in the spatial and temporal domains.

Chapter 2:

AIMS AND HYPOTHESES

2. AIMS AND HYPOTHESES

In the introduction, we stressed the core role of control processes on the guidance of behavior in novel task context. Nonetheless, the majority of the research exploring how control is implemented in the human brain is based on simple repetitive paradigms. While some pioneering neuroimaging studies are now addressing novel instructed performance, it is uncertain how the content of the control-related variables that compose the instructions scaffolds their neural representation. Taking into account the importance of instruction following for our adaptability to changing environments, a better knowledge of its neural underpinning is key for human neuroscience. Our goal in this thesis was to advance the understanding of the neural implementation of the control processes supporting novel instructed behavior. As a preliminary step, we reviewed the broader literature of control processes in the brain. This was followed by three neuroimaging studies, employing fMRI and EEG. In the first one, we studied the transient and sustained processes supporting novel instructed behavior. Afterwards, we focused on novel task-set representation during proactive preparation. In this regard, we tried to answer three questions: (1) the role of relevant task parameters in organizing these representations; (2) the temporal dynamics underlying these flexible coding schemes; and (3) the effect of motivation upon their representational structure. In what follows, we detail each core aims of this work.

2.1. Neural mechanisms of cognitive control – Review 1.

As a starting point of this thesis, we carried out an in-depth review addressing the implementation of control mechanisms in the human brain. This review provided a wide perspective, key for devising the experimental studies of the thesis. We

structured the existing evidence around Braver's (2012) model, distinguishing between proactive and reactive processes. Important, we connected this framework with the Dual-Network Model by Dosenbach and colleagues (2008). Although both are based on the temporal nature of control, Braver's approach focuses on the anticipatory or online nature of these mechanisms, whereas the Dual-Network Model puts the emphasis on their transient or sustained profiles. Hence, the two frameworks only overlap partially. In the review, we aimed at clarifying both perspectives and contrasting them with the studies conducted in the field. This work is included in Chapter 3.

2.2. Transient and Sustained Control Mechanisms Supporting Novel Instructed Behavior – Study 1.

Our first fMRI study assessed the transient and sustained involvement of the fronto-parietal and cingulo-opercular networks (Dosenbach et al., 2008, 2006) in novel instructed tasks. By doing so, we increased the understanding of instructed behavior, and further characterized the influential Dual-Network Model (Dosenbach et al., 2008), which had only been tested in rigid and practiced settings. We used a paradigm where different verbal instructions were encoded and later implemented on unique sets of stimuli (González-García et al., 2017). We manipulated the amount of experience with the instructions to compare novel and practiced ones, which were equivalent otherwise. Following a mixed-design (Petersen & Dubis, 2012), we combined events with blocks to extract phasic and tonics activations. This was possible thanks to analysis techniques based on the combination of FIR and HRF modeling of the Blood Oxygen Level Dependent (BOLD) signal (Visscher et al., 2003).

As novelty entails demands for cognitive control (Norman & Shallice, 1986), we expected higher recruitment of both sustained and transient activations for new than practiced instructions. However, the higher load of instructed-behavior on the proactive component led us to hypothesize stronger effects at both timescales in the CON. The publication of this study is included in Chapter 4.

2.3. Representational organization of novel task sets during their proactive encoding – Study 2.

Next, we studied how novel task-sets are encoded in the neural patterns of proactive activity. Previous research shows that the content of instruction, reflected on target stimulus category, is represented with anticipation in frontoparietal regions (e.g. González-García et al., 2017). Nonetheless, the underlying organization of more complex information guiding instruction encoding is unknown. Thus, in our second fMRI study, we explored the role of relevant task parameters in structuring preparatory activations. We generated instructions by manipulating dimension integration requirements, response set complexity and target category. Using RSA (Kriegeskorte et al., 2008), we estimated the representational space in each location of the brain (Kriegeskorte et al., 2006) and compared it against models based on each of the manipulations. This approach uncovered which brain regions changed their pattern of activation according to each of these task parameters.

We expected to find distinct encoding structures across brain regions. Specifically, we anticipated that the dimension integration requirements, which was the most abstract parameter manipulated, organized task sets in LPFC. On the other hand, we expected response set complexity to have an effect on the IPS. Finally, we hypothesized that target category would affect the representational

space in prefrontal as well as in perceptual cortices. This study can be found in Chapter 5.

2.4. Temporal dynamics underlying different structures for novel task anticipatory coding – Study 3.

We also aimed to characterize the temporal unfolding of the different encoding structures found in Study 2. While recent research stressed the dynamic nature of task-sets during their preparation, no studies have been conducted in the context of novel tasks. Moreover, the employment of fMRI in Study 2 did not allow the extraction of fine-grained temporal information. In Study 3 we acquired high-density EEG data while participants followed the paradigm used in the previous study. We also replicated the analysis employed before, to allow the comparison among fMRI and EEG results. Crucially, we applied these techniques in a time-resolved fashion.

Due to the novelty of the topic, experimental approach and analyses followed, establishing clear hypotheses was risky. However, in general terms, we expected dimension integration requirements and task category to have earlier effects on task structure than response set complexity. This would be in line with previous proposals of two differentiated stages during preparation: a first one, linked to abstract goal setting, and a second, more concrete one, associated with stimulus-response link updating. The study is described in detail in its corresponding section.

2.5. Proactive control-motivation interplay in novelty contexts – Studies 2 and 3.

An additional goal of Studies 2 and 3 was to investigate if and how motivation changes representational spaces to improve behavioral performance. Previous literature shows that proactive control presents an intricate relationship with motivation (Chiew & Braver, 2016; Pessoa, 2009). Recent investigations suggest that one of these interactive mechanisms improves the fidelity of task representations when monetary rewards are expected (Etzel et al., 2016; Hall-McMaster, Muhle-Karbe, Myers, & Stokes, 2019). Whether this effect also affects novel, variable rule encoding was, however, unknown. We concurrently explored the modulation exerted by rewards on the encoding structure of complex instructions, employing fMRI in Study 2 and EEG in Study 3.

Initially, we held two alternative hypotheses, stemming from previous research. First, we expected that reward would polarize the representational structures of the relevant instruction content, making task parameters more efficient at organizing proactive activations. Alternatively, reward could just increase the overall distinguishability of instructions representations. However, against our initial hypotheses Study 2 showed that motivation increased task set similarity. Thus, in Study 3 we hypothesized that we would replicate this effect, using a different neuroimaging method.

Chapter 3:

Neural mechanisms of cognitive control – Review 1

Published as:

Palenciano, A.F.; Díaz-Gutiérrez, P.; González-García, C.; & Ruz, M. (2017). Neural mechanisms of cognitive control / Mecanismos neurales de control cognitivo. *Estudios de Psicología*, 38, 311-337

https://doi.org/10.1080/02109395.2017.1305060

Abstract

Understanding the neural basis of cognitive control is a central issue in cognitive neuroscience, given its core importance for the flexibility that characterizes human behaviour. This review integrates the main findings in the field, underscoring the role of fronto-parietal regions in both proactive (representing tasks in anticipation to prepare the system for action) and reactive (detecting and resolving conflicts in processing) control. In addition, we review the dynamics of interaction between these areas and other brain regions in the range of slow frequencies. Finally, we highlight central ques- tions in the field that have yet to be answered.

Our adaptation to different and ever-changing environments responds largely to an ability to guide our behaviour according to goals, especially in novel situations or those that trigger significant but ineffective action plans (Norman & Shallice, 1986). Underlying this ability are cognitive control processes (or control processes from here onwards), which comprise the mechanisms that articulate human executive functions. Psychology and cognitive neuroscience seek to describe and explain them, as well as to understand their neural bases.

The question of how the brain supports control has been studied at different levels of analysis ranging from the most microscopic level, which explores the operation of neural assemblies, to macroscopic neural networks and their dynamics. Diverse evidence suggests that cognitive functions such as control are implemented in sets of regions, or networks, where each area carries out specific computations (Posner & Petersen, 1990). The development of neuroima- ging techniques (such as functional magnetic resonance imaging) and sophisti- cated analysis strategies (e.g., multi-voxel pattern analysis), together with the study of the dynamics of interaction between regions (employing electro or magnetoencephalography) have advanced our knowledge in this field.

However, consensus on the definition and scientific taxonomy of cognitive control is elusive. For instance, its partial overlap with the construct of attention reflects these difficulties. While some theoretical models conceptualize control as one of three attentional functions (Posner & Petersen, 1990), others associate this construct with brain regions that maintain task goals and bias activity in relevant information processing areas, where the selection associated with attention takes place (Desimone & Duncan, 1995). Other theoretical frameworks on cognitive control propose a more complex picture by adding a temporal dimension. In this

line, Braver (2012) proposed the existence of two control modes: proactive, underlying preparatory adjustments that occur prior to a demanding situation; and reactive, related to the resolution of demands as they arise. In the current article, we review the key findings of these models. First we provide a description of the control networks in the human brain, and then we detail the structures involved in its proactive and reactive operations together with their underlying dynamics.

3.1. Brain networks of control.

In 1990, Posner and Petersen published a seminal review of early studies on the anatomical basis of control processes in the human brain. This review not only laid the theoretical basis for subsequent studies of attention and control, but made a key proposal of the network underlying control in the brain, the Anterior Attentional System or Executive Control Network. This network presented attributes associated with controlled processing according to classical theoretical models (e.g., capacity limits or access of information to consciousness; Norman & Shallice, 1986). Thus, this network was initially associated with focal attention (characterized by target detection), as opposed to the Posterior Attention Network, related to the spatial orienting of attention.

Posner and Petersen (1990) located control processes in two structures: the anterior cingulate and prefrontal cortices (ACC and PFC respectively, shown in Table 3.1, along with the acronyms for other brain regions mentioned throughout the text). Further research has extended the number of regions involved and clarified their roles, describing the control signals they process. One of the most significant contributions is the proposal by Dosenbach et al. (Dosenbach, Fair, Cohen, Schlaggar, & Petersen, 2008), which distinguishes two networks of

executive control. First, a fronto-parietal network is associated with fast, phasic processes of adjustment such as the start of tasks and the commission of errors. Second, a cingulo-opercular network would be at the base of slow or tonic control, keeping the task goals and rules active over prolonged time periods. In addition to describing the key role of other regions in control, this proposal emphasizes the importance of temporal dynamics (phasic vs. tonic) in the func- tions implemented by each system. It also allows the prediction of how these two systems interact with each other and with incoming stimuli: both exert a top-down influence on processing, but their connections with the cerebellum also allow continuous bottom-up access to information, as well as interaction between the two networks.

Tabla 3.1. Índice de las siglas en inglés empleadas para designar las distintas regiones cerebrales.

Siglas	**Región cerebral**
ACC	Corteza del cíngulo anterior
aI/fO	Ínsula anterior / Opérculo frontal
aPFC	Corteza prefrontal anterior
dACC	Corteza del cíngulo anterior dorsal
IFJ	Unión frontal inferior
IFS	Surco frontal inferior
IPS	Surco intraparietal
LPFC	Corteza prefrontal lateral
PFC	Corteza prefrontal
preSMA	Área motora presuplementaria
RLPFC	Corteza prefrontal rostrolateral
VLPFC	Corteza prefrontal ventrolateral
vmPFC	Corteza prefrontal ventormedial

The role of a fronto-parietal network in control processes together with regions such as the insula and ACC has also been described by other models such as the

Multiple Demand Network (MDN), proposed by Duncan (2010). This theory defines the computations underlying control, emphasizing its role in the assembly of subtasks through structured mental programmes (Duncan, 2010). In this sense, the control network fulfils three main objectives: (1) it represents the specific content of the current cognitive goal; (2) it quickly reorganizes resources accor- ding to changes in mental status; and (3) it separates successive subtasks in a distinctive fashion. Although in Duncan's model the dissociation at a neural level of proactive and reactive control is not as straightforward as Dosenbach suggests, the involvement of the set of areas that compose the MDN in contexts of high cognitive demand is clear (see Figure 3.1). Specific regions include the inferior frontal sulcus (IFS), the rostrolateral prefrontal cortex (RLPFC), the anterior insula/frontal operculum (aI/fO), the ACC, the pre-supplementary motor area (pre-SMA) and the intraparietal sulcus (IPS). In addition, these areas are also active during tests of fluid intelligence.

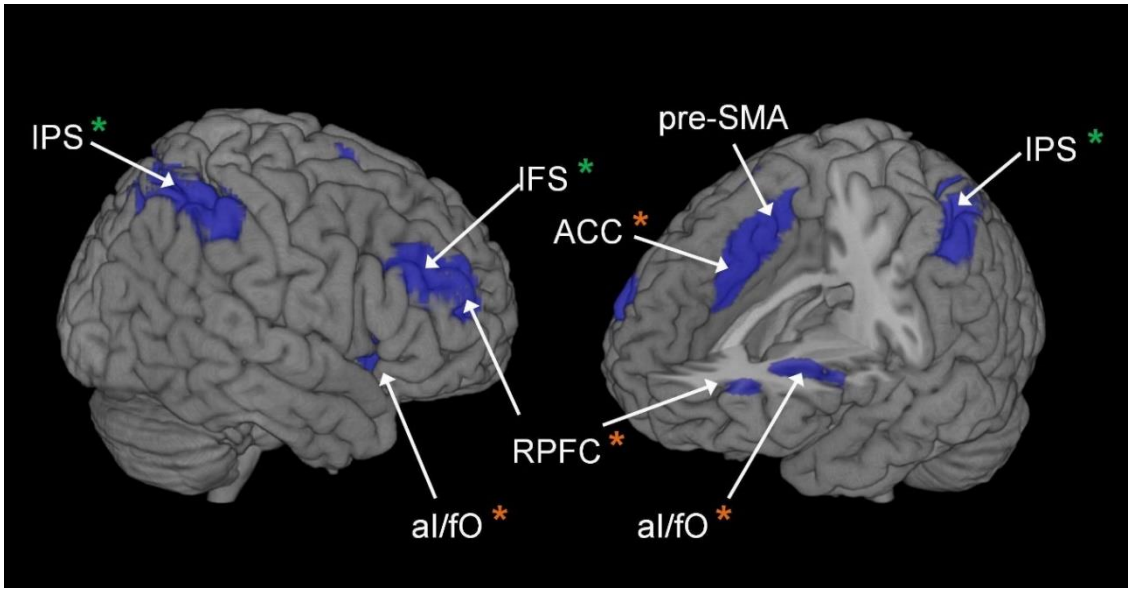

Figure 3.1. Multiple Demand Network (Duncan, 2010; in imaging.mrc-cbu.cam.ac.uk/imaging/MDsystem). The anterior cingulate cortex (ACC), rostrolateral prefrontal cortex (RLPFC) and the anterior insula/frontal operculum (aI/fO) comprise the cingulo-opercular system, and the intraparietal sulcus (IPS), and the inferior frontal sulcus (IFS) are part of the fronto-parietal network (orange and green asterisks, respectively; Dosenbach et al., 2008).

Data from the study of the brain at rest (in the absence of a task) also support the view that the areas mentioned above do indeed comprise a functional net- work. In this sense, Fox et al. (2005) proposed a Task-Positive Network. This network presents extensive overlap with the MDN and shows synchrony in the activity fluctuations of its nodes at rest, which indexes functional communica- tion. On the other hand, recent studies support its subdivision into the two components (fronto-parietal and cingulo-opercular) proposed by Dosenbach. Crittenden, Mitchell, and Duncan (2016) showed that functional connectivity is greater within each subcomponent of the network than among them, and also that the information encoded in each region differs depending on the system to which it belongs.

Finally, it is important to highlight the significance of other structures beyond the MDN. For example, the Default Mode Network (DMN; Raichle et al., 2001), anchored in the ventromedial prefrontal cortex (vmPFC) and the precuneus, is frequently deactivated during cognitive tasks. This DMN has been associated with functions that differ from external control processes, such as mind-wandering. However, its active role, along with the hippocampus, has recently been evi- denced when large changes in cognitive context are required (Crittenden, Mitchell, & Duncan, 2015). On the other hand, the continuous interaction between the basal ganglia and the PFC appears to underlie the acquisition of complex goal- directed behaviours (Buschman & Miller, 2014). Similarly, some of the models already mentioned highlight the role of areas such as the thalamus or the cere- bellum (e.g., Dosenbach et al., 2008).

In short, the evidence obtained so far agrees on the importance of a number of structures in cognitive control. Once identified, unravelling their function is

crucial. To this end, in the following paragraphs we describe the current literature on this matter, beginning with processes related to the proactive control of behaviour.

3.2. Proactive control in the brain

Humans can prepare in advance by applying control in a proactive manner. To do this, we encode the relevant aspects of the task in advance and maintain them in an active state (Sakai, 2008). In terms of processes, this relates to the representa- tion of an abstract and global task set and the activation of the specific rules that compose it (Rubinstein, Meyer, & Evans, 2001). In addition, motor preparation and inhibition of irrelevant responses also take place. All this relates to the selective activation of perceptual and motor processes, which improves subse- quent performance (Miller & Cohen, 2001).

The study of this phenomenon employs paradigms that specify certain aspects of the behaviour that will be demanded; one of the best-known examples is using cues to instruct the task to be performed on a subsequent stimulus (e.g., Monsell, 2003). The time interval often introduced between cue and target stimulus allows participants to prepare in advance. The brain activity generated by the cue and maintained throughout the interval allows us to study the role of different regions in proactive control.

Prefrontal cortex and the representation of the task set.

The PFC is crucial to the anticipatory representation of task sets, that is, the mental models of the task that include the relevant stimuli and responses, the rules that bind them, and the consequences of executing the actions (Sakai,

2008). This complex representation biases the activation in other structures related to more modular task-relevant computations (Miller & Cohen, 2001).

Specifically, the lateral prefrontal cortex (LPFC) frequently shows activity related to the coding and maintenance of rules (e.g., Brass & von Cramon, 2004), which can be decoded before implementation (Reverberi, Gorgen, & Haynes, 2012). Importantly, the accuracy of this decoding correlates with perfor- mance, and it is modified by factors such as motivation (Etzel, Cole, Zacks, Kay, & Braver, 2016). All this indicates that these processes are of a preparatory nature and exert a clear impact on behaviour.

As we move rostrally in the PFC towards the anterior prefrontal cortex (aPFC), the representations become more abstract, and strategies and intentions are also encoded (Haynes et al., 2007). Both the lateral and the medial portions of the aPFC have been related to these processes, although the latter (not part of control networks such as the MDN) has recently accumulated the most evidence in this regard (e.g., Landsiedel & Gilbert, 2015). However, the lateral portion of the aPFC participates as well (Momennejad & Haynes, 2013), and also plays an important role in coordinating the activation of LPFC regions during preparation (Sakai & Passingham, 2006). In short, these data as a whole can be interpreted in terms of a gradient along the rostro-caudal axis in the PFC (Figure 3.2): anterior regions represent more abstract content and exercise control over posterior areas (e.g., Koechlin, Ody, & Kouneiher, 2003).

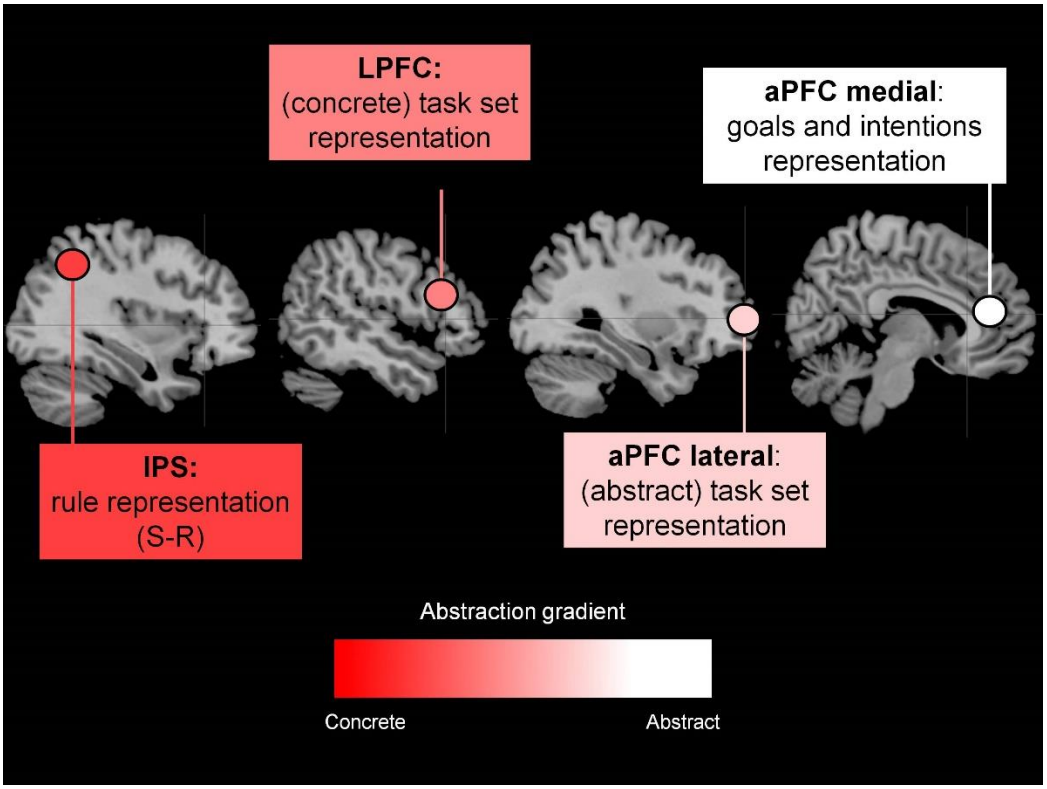

Figure 3.2. Possible interpretation of the proactive control processes in terms of a rostro- caudal gradient of abstraction (Koechlin et al., 2003). The gradient moves from the left (more specific, in red) to the right (more abstract, in white), indicating both the underlying representations and the associated brain regions. IPS: Intraparietal sulcus; LPFC: Lateral prefrontal cortex; aPFC: Anterior prefrontal cortex.

It is also important to emphasize the role of a functionally distinguishable area in the caudal part of the PFC: the inferior frontal junction (IFJ). Several meta-analyses indicate that it is responsible for updating the task set when demands change (e.g., Brass, Derrfuss, Forstmann, & von Cramon, 2005), which is a central aspect of cognitive flexibility.

Role of parietal cortex during preparation

The intraparietal sulcus (IPS) is frequently involved in proactive control. Its role in sensorimotor integration has fostered its association with the representation of specific rules linking stimuli and responses (Brass & von Cramon, 2004). If we extend the rostro-caudal gradient to the parietal cortex, the IPS would specify the more abstract task set represented in anterior regions of the frontal cortex (see

Figure 3.2). In support of this idea, the disruption of the IPS with transcranial magnetic stimulation specifically affects the reconfiguration of action rules (Muhle-Karbe, Andres, & Brass, 2014).

However, the proactive role of the parietal cortex is not yet clear. Its involvement in task-cueing experiments is often very similar to that of the LPFC, which makes it difficult to dissociate both areas (Crone, Wendelken, Donohue, & Bunge, 2006). There is also evidence that the key to prepare for highly complex tasks is precisely the synergy between these two regions (that is, the operation of the frontoparietal network as a unit; Cole et al., 2013). On the other hand, other results show that in certain situations the IPS is the only region that encodes the task set, even when this includes rather abstract contextual information (Wisniewski, Reverberi, Momennejad, Kahnt, & Haynes, 2015). All this points to the need to refine experimental designs and to interpret with caution the data obtained from this region.

The presupplementary motor area

The presupplementary motor area (pre-SMA) has been associated, in this context, with two different functions. On the one hand, it may support unspecific pre- paratory processes: for example, when two consecutive signals are presented before the stimulus, the pre-SMA activates to encode them both, regardless of whether they indicate the same or different tasks (Brass & von Cramon, 2004). On the other hand, it has also been involved in the inhibition of responses or previously relevant contingencies (Crone et al., 2006). However, it is also possible to decode specific aspects of the task set using multivoxel activity patterns from this area (Crittenden et al., 2016). Hence, it is possible that, as occurs with the parietal cortex, this region performs various preparatory processes, depending on

the characteristics of the situation. Its precise role in each situation could be determined by the abstract representations of context that orchestrate the networks involved in the task and the dynamics of their interactions (see below).

The cingulo-opercular network: a comprehensive system for proactive control.

The network comprising the anterior cingulate and insula/frontal operculum (ACC and aI/fO, respectively) shows, in addition to transient responses to cues and errors, activity sustained over prolonged periods of time (Dosenbach et al., 2008). This could underlie the maintenance of task sets that endure over time, thus freeing demands of reactive control.

However, there is no clear evidence that these regions encode the task to be performed at these slower time scales. Although they represent the rules to implement in a transient manner, a more discernible pattern is found in fronto-parietal areas (Crittenden et al., 2016). This may indicate that the role of cingulo-opercular regions in proactive control is more general: sustained activity may establish a 'control mode', in terms of a highly abstract strategy common to demanding tasks. However, there is conflicting evidence. For example, a recent meta-analysis suggests that the functional pattern of this network is indistinguish- able from that associated with the fronto-parietal network (Anderson, Kinnison, & Pessoa, 2013). Hence, characterizing the information implemented by the cingulo- opercular system is one of the key questions in the current research scenario of proactive control.

3.3. Reactive control in the brain

As introduced above, control mechanisms not only prepare the system in advance, but also make adjustments adaptively during the execution of the task.

This happens when we are faced with events that generate conflict in information processing, usually by the simultaneous activation of incompatible action tendencies (Botvinick, Braver, Barch, Carter, & Cohen, 2001). These reactive processes have been studied using paradigms of interference, such as the Stroop (Stroop, 1935) or Flanker (Eriksen & Eriksen, 1974) tasks. In these types of paradigms, irrelevant dimensions of stimuli are associated with preponderant or automatic responses that can be incongruent and interfere with proper actions.

Thus, reactive control involves two mechanisms, one for conflict detection and another one for conflict resolution (Botvinick et al., 2001). In addition, the task set that guides preparatory processes should also be active during task execution. Nonetheless, we will focus on the mechanisms of conflict detection and resolution, as they are the hallmark of reactive control.

Conflict detection.

Regarding the first of these processes, multiple sources of evidence agree on the relevance of the ACC, especially its dorsal portion (dACC), as a region key for conflict detection (e.g. Botvinick et al., 2001; Shenhav, Botvinick, & Cohen, 2013). However, there have been different perspectives about the specific computations or mechanisms underlying this region.

One proposal was that the ACC is responsible for processing errors (Holroyd & Coles, 2002). This hypothesis was supported by the electroencephalographic potential termed Error Related Negativity (ERN), which has been located in the ACC and appears at the commission of errors. Subsequently, it was proposed that the ACC may not exactly detect errors per se, but rather estimate their probability of occurrence (Brown & Braver, 2005).

One of the theories with the largest impact on the field was proposed by Botvinick et al. (2001). This model, originating from simulations of classical tasks of interference, suggested that the dACC implements a mechanism for monitoring conflict. This brain region would be involved in situations where different sources of information interfere with each other. That is the case when there is competition between an automatic but irrelevant response and another relevant but less prominent one, or when the response is indeterminate and different alternatives compete to be selected (e.g., verb generation; Barch, Braver, Sabb, & Noll, 2000). In addition, and linking with previous ERN studies, a third source of conflict would be the commission of errors, normally produced by the co-activation of the correct and incorrect responses.

More recently, Shenhav et al., 2013 have proposed a new perspective on the dACC, reinterpreting its role in terms of a decision-making structure that seeks to optimize the implementation of control. In this sense, this area would carry out cost-benefit analyses, computing the Expected Value of Control (EVC), an index that would guide the decision of how much control to implement and in what direction. The dACC still has an important monitoring role in this model because to calculate the EVC, it must register information about the current state of the person and the consequences of implementing control, anticipating potential rewards and the costs inherent to carrying out that control. The dACC would receive information from the insula and ventromedial regions and the orbitofrontal cortex and thus serve as a centre for integrating control and motivation. Once the EVC has been computed, this structure would play its central role in the speci- fication of the control that needs to be implemented, seeking to maximize rewards and minimize costs.

Finally, Heilbronner and Hayden (2016) have offered an integrative view, according to which the dACC would be responsible for registering variables of different nature to generate action control signals. Thus, the probability of error, the occurrence of conflict and the value of outcomes would all be relevant to the task and thus would be used in the phasic implementation of control.

Conflict resolution.

Many of the models presented in the previous sections assign a central role to the LPFC in resolving the conflict detected by the dACC. More recently, it has been suggested that a set of fronto-parietal structures coincident with the MDN modulate different aspects of information processing to provide the adjustments needed to optimize task performance. Hence, depending on the demands of context these areas participate in proactive control as described in previous sections, or they operate online, in a reactive manner, to resolve conflict (Marini, Demeter, Roberts, Chelazzi, & Woldorff, 2016).

As an example, the dorsal portion of the LPFC is crucial for reactive control as it facilitates the perceptual processing of relevant stimuli that conflict with other more automatic ones (Egner & Hirsch, 2005). The ventrolateral prefrontal cortex (VLPFC) also participates by inhibiting responses that are in competition with the relevant alternative to the task, acting in collaboration with other regions of the MDN such as the anterior insula or the pre-SMA (Levy & Wagner, 2011). Similarly, the dorsal and ventral portions of the parietal cortex also participate in conflict resolution by inhibiting the processing of distracting stimuli that generate interference (Marini et al., 2016).

Finally, a more global mechanism at the base of reactive control of behaviour lies in the flexibility with which the fronto-parietal regions represent the key aspects of the task (rules, stimuli and relevant responses). The encoding of this information varies dynamically and adjusts to unexpected changes in demands. This principle, termed adaptive coding, has been observed along the entire MDN (Woolgar, Jackson, & Duncan, 2016).

3.4. Dynamics of cognitive control.

The previous sections have reviewed the brain structures associated with control and the computations underlying their activity. In recent years, research on the dynamics of these processes has also advanced significantly. Results in this field show that the synchronization and the coupling in different frequency ranges between control networks representing relevant targets, and more modular regions which process relevant information, are core neural computations in cognitive control (Fries, 2015). While local neural assembly computations would be mainly measured by synchronicity in the range of high or rapid frequencies, long range interactions between distant regions, related to cognitive control, use slower frequencies that encompass larger neural groups (Fries, 2015). Part of the litera- ture on cognitive control highlights the importance of coupling mediated by activity in the beta band (Bressler & Richter, 2015), and other results also point to the importance of theta and alpha (e.g., Capilla, Schoffelen, Paterson, Thut, & Gross, 2014).

The dynamics related to goal implementation have been studied in different animal species using different recording techniques and paradigms of various kinds. Buschman and Miller (2007), for example, showed neural synchrony between the PFC and the IPS, with influence of opposite directionality and

different ranges of frequencies, depending on whether behaviour was guided externally by the salient features of the stimuli or internally by the goals set by the individual. Related studies suggest that these types of interactions depend in part on mechanisms of synchrony generated in thalamic nuclei, such as the pulvinar (Saalmann, Pinsk, Wang, Li, & Kastner, 2012).

In another study of task-switching, where individuals had to alternate between rules of colour or orientation of stimuli, Buschman et al. (Buschman, Denovellis, Diogo, Bullock, & Miller, 2012) described the forma- tion of neural ensembles in prefrontal regions that selectively synchronized their firing in the beta frequency band according to the rule implemented. At the same time, the neural assembly representing the non-relevant rule increased its synchrony in the alpha range, associated with the deselection of irrelevant information. Recently, in the same line, Voytek et al. (2015) used a task in which participants implemented rules of increasing abstraction. In addition to involving regions more anterior in the prefrontal cortex, rule abstraction increased phase encoding in the theta range, and it also increased local populations synchronized in gamma, which was predictive of trial-by-trial differences in reaction time. These mechanisms are similar to others proposed in related research areas, such as, for example, sustained attention (Clayton, Yeung, & Kadosh, 2015).

3.5. Final remarks.

Throughout this review, we have discussed how a set of brain regions, acting jointly, implements proactive and reactive control on behaviour. The distinction between the two types of control is not related to the exclusive recruitment of one or another set of areas. On the contrary, this corresponds to a possible organizing principle of brain function, wherein the temporal profile of activations contextua-

lizes the computations carried out by a brain region to solve specific demands. This is a considerable advantage as it allows a functional specialization that depends on temporal dynamics, and not on the anatomical structure of the brain, which is ultimately limited (Braver, 2012).

This theoretical framework has large explanatory power and has an extensive body of evidence supporting it. When this research is taken together with data on brain dynamics, we obtain a comprehensive and detailed perspective of the neural implementation of cognitive control. The first source of information explains how representations that guide behaviour are established, and how they adapt to changes in demands. On the other hand, dynamics data offer a closer view of the mechanisms by which they exert a bias on the neural systems that act as an interface with the environment (interacting in different frequency bands). In short, both perspectives are complementary and allow us to resolve the questions posed by classic control models (Miller & Cohen, 2001).

However, there are several aspects that require further research. One of them refers to the anatomical specificity of the data collected to date. Despite multiple attempts to establish a comprehensive parcellation of the human brain, inconsistencies in labelling and delimitation of specific brain regions still prevail, making it difficult to compare results from considerably different studies. Fortunately, recent efforts are advancing the field in this direction (Glasser et al., 2016).

Parallel to this problem, the multiplicity of tasks employed (especially those used to measure a single variable, but with different manipulations) can lead to confusion between the conclusions drawn. Therefore, we must explain the function of each region more specifically when describing their role (and avoid simply labelling it). This is particularly relevant when there is overlap between

the functions assigned to an area: a more detailed description would assess whether we are in fact dealing with different processes, or if instead the area in question implements a more general computation that is evidenced in different experimental situations. In this regard, the continuous sophistication of paradigms and analysis techniques promises to shed light on the matter.

On the other hand, studies conducted so far have tried to answer, broadly, two questions: where and how is control implemented. This leaves a third, also central, issue unresolved: why is control implemented in this way and not another. Recent research using simulations and computational models has suggested that repre- sentations of control are a reflection of the structure of the problems we have had to face during phylogenetic evolution (Botvinick & Cohen, 2014), thus offering an explanation as a phenomenon occurring during the evolution of the brain as a biological system.

Finally, research on cognitive control would benefit from its integration with other theories of brain function. Such is the case of predictive coding (Friston, 2005). According to this theory, perception is not a passive phenomenon, guided by bottom-up information that accesses the system, but emerges from bottom-up and top-down cycles. From this perspective then, we should investigate whether proactive control mechanisms are involved in perceptual processes. This is an unexplored area which raises intriguing questions in the field.

In short, the study of the brain basis of cognitive control is a highly productive field, but we still have a long way to go. Undoubtedly, research in the coming years accompanied by technological innovations will answer many of these issues, which are of great interest to neuroscience and cognitive psychology.

Chapter 4:

Transient and sustained Control Mechanisms Supporting Novel Instructed Behavior – Study 1

Published as:

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

Abstract

The success of humans in novel environments is partially supported by our ability to implement new task procedures via instructions. This complex skill has been associated with the activity of control-related brain areas. Current models link fronto-parietal and a cingulo-opercular networks with transient and sustained modes of cognitive control, based on observations during repetitive task settings or rest (Dosenbach et al. 2008). The current study extends this dual model to novel instructed tasks. We employed a mixed design and an instruction-following task to extract phasic and tonic brain signals associated with the encoding and implementation of novel verbal rules. We also performed a representation similarity analysis to capture consistency in task-set encoding within trial epochs. Our findings show that both networks are involved while following novel instructions: transiently, during the implementation of the instruction, and in a sustained fashion, across novel trials blocks. Moreover, the multivariate results showed that task representations in the cingulo-opercular network were more stable than in the fronto-parietal one. Our data extend the dual model of cognitive control to novel demanding situations, highlighting the high flexibility of control-related regions in adopting different temporal profiles.

4.1. Introduction

Following verbal instructions could seem, at first glance, a trivial aspect of human behavior, perhaps due to the easiness that we often experiment when following commands in our daily life. However, in continuously changing environments, the ability to use instructions to guide actions is essential for fit performance. In fact, this skill defines a crucial distinction between us and non-human apes: using language to share task procedures freed us from slow trial-and-error learning (Cole et al. 2013). Despite the biological relevance of this complex, flexible skill, some important aspects of its underlying neural architecture remain unknown. In the present study, we employed functional magnetic resonance imaging (fMRI) and both univariate and multivariate approaches to describe the transient and sustained control processes that allow us to follow novel verbal instructions.

The transformation of an instruction into effective behavior involves different processes. First, rules are semantically *encoded*, and proactive control processes (Meiran 1996; Braver 2012) are deployed to build a representation of the task (the so-called *task-set*; Sakai 2008). This set can be activated in advance (Meiran 2010; Ruge et al. 2013), biasing task-relevant processing in sensorimotor regions (e.g. Sakai and Passingham 2003; Sakai and Passingham 2006; Ekman et al. 2012; González-García et al. 2016; González-García et al. 2017) and thus, allowing us to *prepare*. Once the task context has been instantiated, task-sets must be *implemented* (Stocco et al. 2012), and reactive control processes become crucial (Cole et al. 2017), as they allow the inhibition of previously relevant action plans and the selection of target stimuli among possible distractors (Botvinick et al. 2001; Braver 2012). These proactive and reactive neural mechanisms, necessary for successful task encoding and implementation, have received considerable

attention in the broader literature of cognitive control (e.g. Braver 2012; Palenciano et al. 2017).

Traditionally, the experimental approaches employed to study cognitive control use rather repetitive paradigms, which trigger proactive task-set reconfiguration with alternations between few rules (e.g., task switching; Monsell 2003) and/or reactive adjustments via conflict (e.g., the Stroop task; Stroop 1935). The evidence so far shows the involvement of a set of frontal and parietal areas during the execution of a wide spectrum of effortful, controlled tasks (Duncan 2010), including novel task execution (e.g., González-García et al. 2017). Due to the tight functional coupling of these regions (Fox et al. 2005; Seeley et al. 2007; Cole and Schneider 2007), they are often considered a unitary control brain network (namely, the Multiple Demand Network or MDN; Duncan 2010; Fedorenko et al. 2013). However, recent advances in experimental design and data analysis have led to its subdivision into at least two components -the cingulo-opercular and the fronto-parietal networks (CON and FPN, respectively)-, which seem to act at different, complementary time scales (Dosenbach et al. 2006; Dosenbach et al. 2008). The CON is comprised by regions that show both preparatory (cue-related) and sustained (across multiple trials) activations (Dosenbach et al. 2006), and has been associated with the proactive activation and maintenance of task-sets (Dosenbach et al. 2007). Conversely, FPN regions present mainly transient, cue and error-locked activity (Dosenbach et al. 2006) and their role has been described in terms of phasic, reactive adjustment of behavior (Dosenbach et al. 2007).

Support for this dual distinction comes not only from the analysis of sustained and transient neural signals while participants perform different tasks

(Dosenbach et al. 2006), but it has also been confirmed when analyzing the information encoded in multivoxel activity patterns in those regions (Crittenden et al. 2016) and in functional connectivity data (both in resting state and on task; Dosenbach et al. 2007; Crittenden et al. 2016). Nevertheless, it has also been evidenced that such dual functioning, and specially the sustained involvement of the CON, is absent in certain task contexts (for example, when stimuli contain enough perceptual information to guide the response; Dubis et al. 2016). Last, crucially to the current study, it remains unknown whether there is a differential involvement of the two systems during goal-directed behavior in contexts of novelty. As novel tasks entail higher control demands than practiced ones (Norman and Shallice 1986), it is expected that they would be associated with a greater recruitment of maintained and transient processes mediated by CON and FPN, which could highlight their distinction.

Research in recent years has explored the brain regions underlying the encoding and implementation of instructions, and the specific roles carried out by each one (Brass et al. 2017). The findings so far support the involvement of the two main nodes of the FPN, the inferior frontal (IFS) and the intraparietal sulcus (IPS; e.g. Ruge and Wolfensteller 2010; Dumontheil et al. 2011; Muhle-Karbe et al. 2017), as expected from Dosenbach and colleagues' model. The lateral prefrontal cortex (LPFC) in general, and the IFS in particular, have been linked to the encoding of new instructions (Hartstra et al. 2011; Demanet et al. 2016), showing higher activity in novel compared to practiced contexts (Cole et al. 2010; Ruge and Wolfensteller 2010). This region may be in charge, specifically, of the formation of novel stimulus-response mappings (when comparing against the formation of stimulus-stimulus associations; Hartstra et al. 2012). This supports its

involvement in proactive processes related to the creation of novel task-sets, and not in the mere declarative maintenance of instructions in working memory (Hartstra et al. 2012; Brass et al. 2017). The IPS has shown, generally, a similar pattern (Ruge and Wolfensteller 2010; Dumontheil et al. 2011), although there is also evidence of a less abstract, sensorimotor representation in this region (Hartstra et al. 2012; Muhle-Karbe et al. 2014; González-García et al. 2017). Importantly, the functional coupling of the IFS and IPS with other brain regions contains fine-grained information about the content of novel instructions (Cole, Reynolds, et al. 2013). These distributed mechanisms of task-set representation also add evidence for the joint activation of fronto-parietal regions as a coherent functional system.

On the other hand, the CON network consists of the dorsal anterior cingulate (dACC), the anterior insula/frontal operculum area (aI/fO) and the anterior prefrontal cortex (aPFC). In contrast to the FPN, evidence of its involvement during instructed behavior is scarce. The dACC has been associated, in this context, with the reactive inhibition of irrelevant actions that interfere with the proper response (Botvinick et al. 2001; Brass et al. 2009). However, existing evidence does not yield strong support for a role of the dACC or the aI/fO in the encoding and/or maintenance of new instructed rules. The aPFC, in contrast, has been highlighted as a key region in the construction of novel task-sets, but only when rules are complex or abstract (Cole et al. 2010). Thus, the CON has not shown, as a system, a consistent behavior as the one predicted from the dual model framework.

The differential support for the participation of the two networks in novel instructed behavior could be due to different reasons. On the one hand, the nature

of the behaviors explored could weight on transient mechanisms (FPN) to a higher extent than on sustained ones (CON), which besides of being more resource consuming (Braver 2012), develop in a time scale that may not be optimal in this context. In other words, the activity maintained in CON areas could be maximally beneficial when the relevant rules are stable in time (as in classic control paradigms), but not if quick task-set reconfigurations take place in a trial-by-trial fashion. In accordance with this idea, it has been proposed that reactive mechanisms are key to potentiate flexibility in novel instruction following (Cole et al. 2017). On the other hand, the evidence to date is scarce in contexts where novel instructions are embedded in designs aimed at isolating both control modes, which by definition act at different temporal scales.

When employing fMRI mixed designs (Petersen and Dubis 2012), the combination of events and blocks allows for the disambiguation of transient and sustained neural signals. To date, only one instructions study has been carried out using mixed designs (Dumontheil et al. 2011), and it employed complex practiced commands. These authors manipulated task-set complexity and studied transient activations linked to the encoding and implementation of instructions, while the sustained activations were analyzed only during implementation. Surprisingly, only two regions were involved in their sustained results: the IFS and the aPFC. Thus, the equal involvement of regions from both networks leaves open the role of the CON in instructed task execution and more importantly, whether this pattern applies to novel contexts.

We aimed to conduct an experiment which specifically tested the involvement of the dual control system proposed by Dosenbach and colleagues (Dosenbach et al. 2006; Dosenbach et al. 2008) during novel, instructed behavior. To do so, we

adapted an instruction-following paradigm (González-García et al. 2017) to an fMRI mixed design, manipulating the experience with the instructions (novel vs. practiced) in different blocks of trials. This allowed comparing novelty-related activity patterns (i.e., sustained and phasic activations) against a control practiced condition. Furthermore, we aimed to better characterize the sustained activation profile associated with the CON. As the standard univariate analyses employed in previous studies did not help to clarify the information held by these networks, other plausible hypotheses in addition to proactive control involvement have been proposed (e.g., tonic attention maintenance; Coste and Kleinschmidt 2016). To address this issue, we employed recent multivariate techniques (Haynes and Rees 2006), an approach that has been shown to be highly informative. For example, using a combination of Multi-Voxel Pattern Analysis (MVPA) and Representational Similarity Analysis (RSA; Kriegeskorte et al. 2008), Qiao and colleagues (Qiao et al. 2017) were able to characterize how FPN areas adaptively change the task-set being represented, and how this process deals with interference from previous relevant rules. The dual-network model would predict a better maintenance through time of task-sets in CON, complementing the quick adjustment of the information encoded across the FPN. Thus, we employed RSA to assess whether the spatially distributed task representations were more consistent over time in CON than in FPN areas.

4.2. Methods and materials

Participants

37 students from the University of Granada, all right-handed and with normal or corrected-to-normal vision were recruited for the experiment (20 women, mean age = 21.13, SD = 2.47). All of them signed a consent form approved by the Ethics

Committee of the University of Granada and received payment (20 to 25€, according to their performance) or course credits in exchange for their participation. Two participants were excluded from the final sample due to excess of head movement (> 3 mm). Sample size was selected according to recommendations for mixed designs (Petersen and Dubis 2012).

Apparatus and stimuli

We used a total of 120 verbal instructions similar to those employed by González-García and colleagues (González-García et al. 2017). They were all composed by a condition and the two responses associated with the condition being true or false (e.g.: "*If there are four happy faces, press L. If not, press A*"). Half of the instructions referred to faces (their *gender* -female, male-, *emotional expression* -happy, sad-, or both), whereas the remaining referred to letters (their *type* -vowel, consonant-, *color* -blue, red-, or both). The instruction could also specify the *quantity* of specific stimuli, their *size*, or the *spatial contiguity* between them. Finally, the motor responses indicated a left or right index button press ("*press A*" or "*press L*", respectively). Face and letter sets were equivalent in terms of these parameters. We conducted a pilot behavioral study to ensure that the difficulty was equivalent across the whole set. Then, to shorten task duration for the fMRI protocol, we built up six 100-instructions lists from the pool (again, equating face and letter-related elements) and assigned them to the participants, so each individual instruction was presented with the same frequency across our sample.

For each instruction, we built two grids of target stimuli: one fulfilling the condition specified (match) and the other one not (mismatch). They all consisted of *unique* combinations of 4 faces and 4 letters, which were drawn from a pool of

16 pictures: 8 face images (2 men and 2 women, 2 with happy expression and 2 sad, each in two different sizes -big/small-) from the Karolinska Directed Emotional Faces set (Lundqvist et al. 1998) and 8 letter images (2 consonants and 2 vowels, 2 in red color and 2 in blue, each in two different sizes -large/small-). Grids from face and letter instruction sets were built in parallel (establishing an equivalence between gender-letter type and emotion-color). Across the whole sample of participants, all instruction-stimuli (matching and mismatching grids) and instruction-response combinations (press A if true, press L if false; or the opposite) were employed.

The task was created with E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). Inside the scanner, it was projected onto a screen visible through a mirror located on the head coil.

Procedure

Participants performed a task in which they implemented novel and practiced verbal instructions referring to letters or faces, inside the fMRI scanner. The timing of the whole task was adapted to match the TR of the EPI sequence (2.21s), anchoring each event to the beginning of a scan acquisition, due to requirements of the FIR analyses conducted (see *fMRI analysis section*). Each trial (Fig. 4.1) started with the presentation of a verbal instruction (25.75°; *encoding phase*) during 2.21s (i.e., one TR), followed by a jittered interval with a fixation cross (2.21-8.84s, mean =5.525s). The grid of stimuli (21°) then appeared for 2.21s, where participants had to respond (*implementation phase*) using button boxes compatible with the scanner environment. The following trial began after a second jittered delay (with the same characteristics as the previous one).

We were interested in two variables: the experience that the participants had with the trials (new vs. practiced) and the category of stimuli that the instructions referred to (faces vs. letters), having four possible conditions: Faces/New, Letters/New, Faces/Practiced, Letters/Practiced. As we employed a mixed fMRI design for our task, we manipulated those variables between blocks, for a total of 16 blocks (4 of each condition), with ten trials each. All blocks began with a cue indicating the experience and category condition (2.21s) followed by a jittered interval (2.21-8.84s, mean = 5.525s), after which the first trial began. Blocks lasted 154.7s, and were followed and preceded by pause periods of 66.3s (also indicated by pause cues of 2.21s). Importantly, pause duration was chosen to be long enough to ensure a robust baseline for block-related activity. The task was split into four runs, each composed of four blocks, one per condition. We carefully counterbalanced the order of blocks, ensuring that all of them were preceded and followed by the others the same number of times. Runs lasted 17.05 minutes, and the whole task 67.3 minutes.

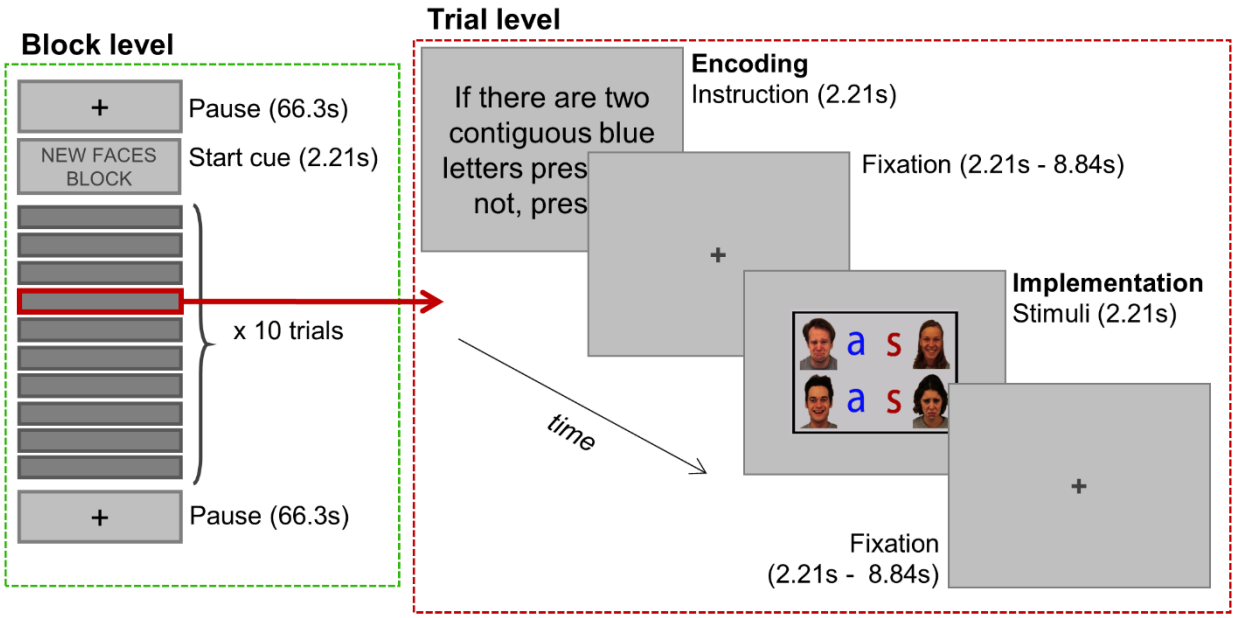

Figure 4.1: Mixed-design behavioral paradigm.

Participants came to the laboratory approximately 24 hours before the fMRI session, and performed 10 repetitions of two blocks of ten instruction-grid pairings each (i.e., Faces/Practiced and Letter/Practiced blocks), which conformed the practiced instructions. Feedback was administered after each trial in this practice session, and learning was assessed in a pre-scanner test, with a requirement of at least 85% correct responses to continue the experiment. Across participants, all materials were equally employed in new and practiced conditions.

FMRI: acquisition and analysis.

MRI data was collected using a 3-Tesla Siemens Trio scanner at the Mind, Brain, and Behavior Research Center (CIMCYC, University of Granada, Spain). We used a T2*-weighted Echo Planar Imaging (EPI) sequence (TR = 2210ms, TE = 23ms, flip angle = 70º) to obtain the functional volumes. These consisted of 40 slices, obtained in descending order, with 2.3mm of thickness (gap = 20%, voxel size = $3mm^2$). The 4 runs consisted of 468 volumes each. We also acquired a high-resolution anatomical T1-weighted image (192 slices of 1mm, TR = 2500ms, TE = 3.69ms, flip angle = 7º, voxel size = $1mm^3$). Participants spent approximately 90 minutes inside the MRI scanner.

We used SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) to preprocess and analyze the data. The first four volumes of each run were excluded to allow for stabilization of the signal. The remaining images were spatially realigned, time-corrected and normalized to the MNI space (transformation matrices were estimated from EPI images, and applied to them in the same step). Finally, they were smoothed using an 8mm FWHM Gaussian kernel. We built our experimental task on the basis of a mixed design (Petersen and Dubis 2012).

Therefore, for each subject, we created a GLM including, simultaneously, events (separately, encoding and execution phases) and block regressors for each of the four conditions, to perform the main univariate analysis of this data. Events were modeled using a Finite Impulse Response (FIR) basis set (9 stick functions, encompassing 19.89s -9 TRs- following the onset of the events), while blocks were convolved with the canonical hemodynamic response (HRF) function (Visscher et al. 2003). We also modeled the pause periods (HRF convolved) and the block/pauses starting cues (FIR modeled), and included the errors (boxcar functions with same duration as the full trials, convolved with the HRF) and six movement parameters as nuisance regressors. A 756s high pass filter was set, taking into account block duration and the maximum time elapsed between events of the same condition.

At the within-subject level of analysis, we conducted *t*-tests comparing event regressors against the implicit baseline, time bin by time bin, separately for each condition. *T*-tests were also conducted to contrast blocks with pause periods (both collapsing across conditions, and separately), and also to compare between blocks of different conditions. At the group level, separate analyses were carried out for the sustained and transient components, in both cases correcting for multiple comparisons using a $P < 0.05$ FWE cluster-wise criterion (from an initial uncorrected $P < .001$). In the first case, we used one sample *t*-tests with the subjects' block contrast images obtained from the first level analyses. For the transient activity, we included the statistical maps obtained from the event contrasts into two ANOVA (encoding and implementation), performed as a full factorial design in SPM12 (Hartstra et al. 2011; Hartstra et al. 2012) and including Experience (novel, practiced), stimulus Category (faces, letters) and Time (9 time

bins) as factors. This SPM design was chosen because it facilitates contrast specification, especially in complex models such as the one employed here. Nonetheless, all results were replicated with a repeated measures ANOVA also including a Subject factor, following an SPM flexible factorial model (Glascher and Gitelman 2008). We assessed main effects of experience and category, and their interaction with time bin. In the interaction of experience with time bin during the implementation stage, significant clusters were too big and extended over several different areas, so we adopted a stricter cluster forming threshold (uncorrected $P < 0.001$) to obtain smaller, anatomically more constrain clusters. Finally, to establish the directionality of these effects, we extracted the beta values of the significant clusters and compared the estimated hemodynamic response across conditions, both plotting the data, and performing post-hoc pairwise comparisons (Bonferroni corrected) with the SPSS software (SPSS 20.0 for Windows, SPSS, Armonk, NY).

We additionally performed non-parametric inference (based on 10.000 permutations and cluster-forming threshold of $P < .001$) on sustained activity data, using the software SnPM (http://www.sph.umich.edu/ni-stat/SnPM). We could not follow this strategy with the transient activity analysis, as the repeated-measures ANOVA design was too complex to implement with the software available. Nonetheless, it is noteworthy that the block non-parametric results successfully replicated the output from the parametric approach.

To further characterize these findings, we carried out three additional analyses. First, we performed a conjunction test (Nichols et al. 2005) to assess the overlap between areas showing sustained and transient (encoding and implementation) activity. To do so, we thresholded ($P < .05$ FWE cluster-wise criterion) the

statistical maps obtained from the following contrasts of interest: (1) *t*-test of novel vs. practiced blocks, (2) main effect of Experience during the encoding of instructions; (3) and interaction of Experience*Time during the implementation stage. These three statistical maps were post-hoc selected based on the findings obtained from the analyses described above and our hypothesis regarding the roles of the CON and the FPN. These images, after being binarized, were used to assess the intersection of the contrasts. As a result, we obtained voxels significantly activated in all three situations simultaneously.

Next, we evaluated the congruency of our results with the proposal of Dosenbach and colleagues (Dosenbach et al. 2008) of two subnetworks for cognitive control. Specifically, we assessed the extent of overlap of the regions showing sustained and transient activations in our experiment with the CON and the FPN, respectively (Dosenbach et al. 2008). For this, we built spherical 10mm radius ROIs centered on the nodes of the CON (dACC [0, 31, 24], aPFC [-21, 43, -10; 21, 43, -10], aI/fO [-35, 18, 3; 35, 18, 3]), and FPN (IFS [-41, 23, 29; 41, 23, 29], IPS [-37, -56, 41; 37, -56, 41]), as published in Fedorenko and cols. (Fedorenko et al. 2013). ROI definition, including sphere size selection, was conducted following the parameters in the study of Dumontheil and colleagues (Dumontheil et al. 2011), in order to facilitate comparisons. The network templates were then overlaid against the thresholded statistical maps that we obtained in our results (using the same contrast images as in the conjunction analysis), after which we assessed which ROIs were present in each map and the percentage of voxels of each subnetwork involved in the different contrasts (Woolgar et al. 2016). It is important to note, however, the descriptive nature of our approach, as it did not involve the computation of inference statistics. This was due to the complexity of

the mixed design analysis (which did not allow to obtain equivalent homogeneous statistics from both event and block-related signals). Nevertheless, the chosen procedure provided an informative comparison of the dual model (Dosenbach et al. 2008) and the sustained and transient activations estimated in our study.

Finally, we conducted a multivariate analysis to study the fine-grained distributed representation of instructions and their consistency along trial epochs (i.e., from the encoding to implementation stages). Specifically, we aimed to test differences in representation persistence between the two networks, and how novelty modulated this effect. To that end, we entered the non-normalized and unsmoothed functional images into a GLM similar to the specified above, with the exception that blocks were not defined and event regressors were convolved with the HRF. This modeling approach was selected because at this point there was no risk of misattributing the signal from transient and sustained components, and more importantly, because it provided a single parameter image for each event condition (instead of nine). The beta coefficient maps extracted (32 in total, corresponding to the encoding and implementation phases of each condition and run) were used to build a 32x32 Representational Dissimilarity Matrix (RDM; using The Decoding Toolbox; Hebart et al. 2014) for each FPN and CON ROI (as defined above), which had previously been inverse-normalized and coregistered to the participants' native space. In the RDMs, each column and row corresponded to a different regressor, and each $cell_{i,j}$ to the distance (computed as 1 - Pearson correlation) between the multivariate activity pattern associated with regressors i and j. Pearson correlation values were first normalized using Fisher's z-transformation. We focused on the quadrant of the RDMs capturing the dissimilarities between encoding and implementation of

instructions, in which the diagonal represented distances within different stages of same condition trials, and the off-diagonal represented values of different condition trials (Fig. 4.2). We computed the average difference between off and on-diagonal values for each ROI (González-García et al. 2018), as an index of representational consistency along time. Concretely, this index showed how similar the patterns of activations at the implementation and encoding stages of same condition were, in comparison with different condition trials. An index of 0 means that the information encoded in multivariate patterns was independent between encoding and implementation, while higher values reflect greater correspondence between the information encoded in both phases. We first checked that the index was significantly above 0 across regions using one-sample *t*-tests. As the aim of this analysis was to assess whether the consistency index varied between the FPN and the CON, we averaged the values of ROIs pertaining to each system and performed a paired *t*-test between them. Even when our main hypothesis-driven approach for this analysis was to group the regions into two segregated control networks (Crittenden et al. 2016), we also wanted to explore differences that could arise among areas of the same component -as there is no reason to assume that they all perform identical computations. To assess this possibility, we conducted a repeated-measures ANOVA within each network, with ROI as factor, which was later qualified with planned comparisons, Bonferroni-corrected. Finally, we obtained the consistency indexes separately for novel and practiced trials, and explored this effect with a repeated-measures ANOVA with Network and Experience as factors.

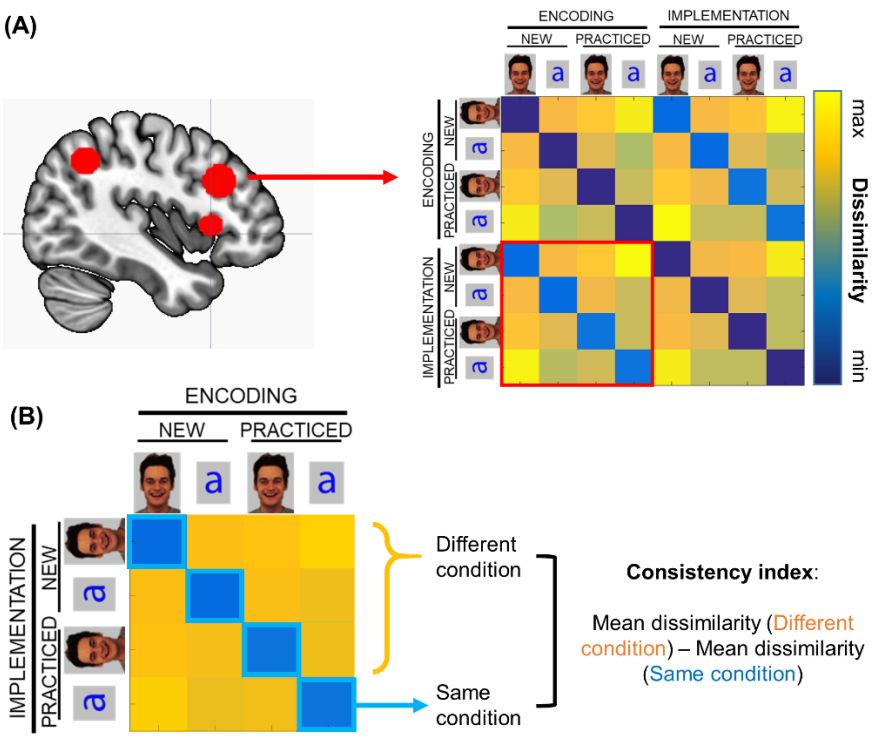

Figure 4.2: Representational Similarity Analysis. (A) First, a representational dissimilarity matrix (RDM) was built using the data of each cingulo-opercular and fronto-parietal region of interest. Each cell of the matrix indicates the dissimilarity between the representation of each pair of trial conditions at encoding and implementation stages. (B) The left lower quadrant was selected in each RDM. Within this quadrant, the diagonal (cells in blue) show dissimilarities between the encoding and the implementation of same-condition trials, and the off-diagonal values (cells in orange) refer to different-condition trials. Those values were averaged separately and subtracted to compute the persistence index employed in the analysis.

4.3. Results

Behavior

We analyzed the behavioral performance during the scanning session using two-way repeated-measures ANOVAs, with Experience (new vs. practiced) and Category (faces vs. letters) as factors. We found a significant effect of Experience in accuracy ($F_{1,34}$ = 51.12, P < .001, η_p^2 = .601), with better performance for practiced (M = 94.7%, SD = 5.3) than for novel trials (M = 88.7%, SD = 6.8). The effect of Category was also significant ($F_{1,34}$ = 5.31, P < .027, η_p^2 = .135), with better performance for faces (M = 92.6%, SD = 6.0) than for letters (M = 90.9%;

SD = 7.4%). Finally, RT data from this session replicated the significant main effect of Experience ($F_{1,34}$ = 290.48, P < .001, η_p^2 = .895), and also showed a significant interaction between Experience and Category ($F_{1,34}$ = 32.56, P < .001, η_p^2 = .489), with faster responses to faces (M = 747.0ms, SD = 196.7ms) than letters (M = 783.6ms, SD = 188.7ms) in practiced trials, and the opposite pattern in novel ones (Faces: M = 1047.7ms, SD = 183.7ms; Letters: M = 982.7ms; SD = 172.1ms). Finally, we performed two additional ANOVAs on accuracy and RT data including Run as a factor, to rule out possible fatigue effects on our behavioral measures. Neither the main effect of Run (accuracy: $F_{3,102}$ = 1.99, P < .120, η_p^2 = .055; RT: $F_{3,102}$ = 2.11, P < .104, η_p^2 = .058) nor its interaction with Experience or Category were significant (Fs < 1.02, Ps > .100). This was further confirmed with a Bayesian repeated-measure ANOVA, in which both the main effect of Run and its interactions showed a BF_{10} < .3, strongly supporting a null effect of this variable and, thus, confirming that participants' performance was stable across the whole task.

fMRI

We first conducted a *univariate analysis* to assess sustained and transient activity, with the goal of exploring the effect of the experience with the task (new vs. practiced). As specified before, we also carried out a *multivariate analysis*, focused on the within-trial time scale, to study the consistency of multivoxel representation along phases of the task (encoding and implementation).

Univariate analysis

Transient activity

Event-locked activations were estimated using a set of FIR functions, obtaining nine parameters per regressor defined at the within-subject level. Then, they were entered into two separate ANOVAs: one to capture phasic activations associated with the encoding of instructions, and the other for their implementation. In both, we assessed the main effect of Experience, and its interaction with Time.

During the encoding of instructions (Table 4.1, Fig. 4.3), the main effect of Experience was significant bilaterally in the dorsolateral prefrontal cortex (DLPFC) -including the IFS-, and aPFC. To explore the directionality of this result, we extracted the beta estimates for each conditions and time bin (averaged across participants). Intriguingly, the hemodynamic response (HDR) was more pronounced for practiced compared to novel instructions in both DLPFC clusters (see Fig. 4.3). In the aPFC, beta values were also higher in the practiced condition, but in that case the HDR did not resemble the typical curve (see Fig. 4.3), but showed a deactivation, less pronounced for practiced rules.

In contrast, a wide array of brain areas was differently activated in novel and practiced trials during the implementation of instructions (Table 4.1), as assessed by the interaction of Experience with Time (Fig. 4.4). As clusters were very large, we used a stricter statistical threshold to explore smaller, anatomically more accurate clusters (uncorrected cluster-defining threshold of $P < .0001$; this threshold was also employed to display the results in Fig. 4.4 and Table 4.1). In contrast to the encoding stage, almost all regions showed a higher HDR for novel than for practiced instructions, including the IFS, the inferior frontal junction (IFJ), the IPL and the aI/fO (Fig. 4.4). On the other hand, the bilateral supramarginal and superior temporal gyrus were more active in practiced trials.

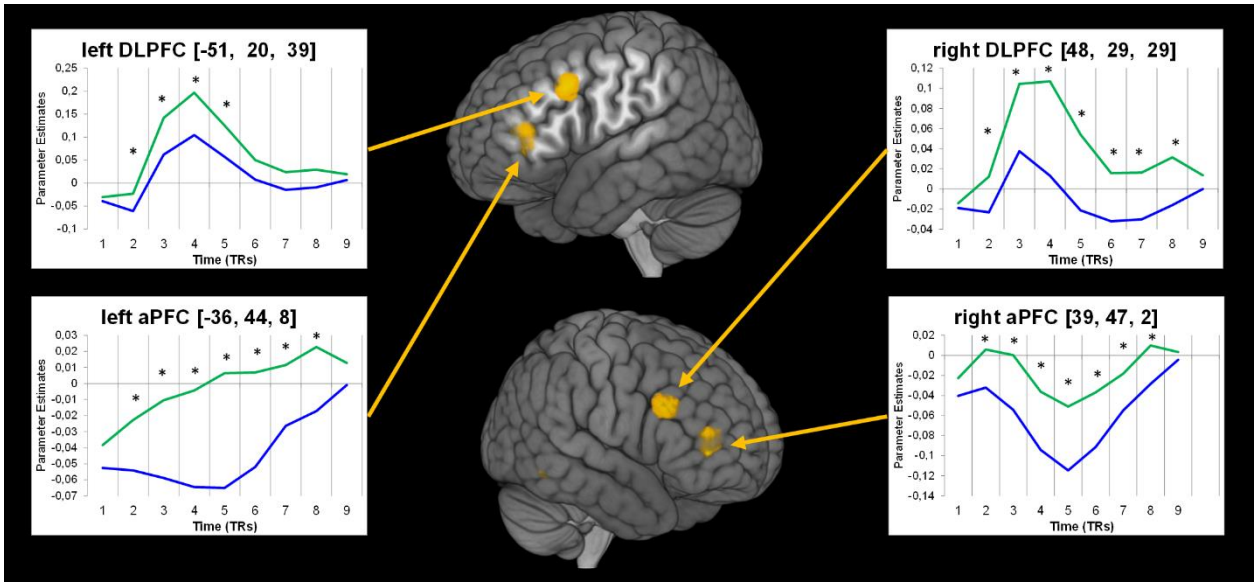

Figure 4.3: Results from the encoding stage ANOVA. Yellow clusters show regions where the main effect of Experience was significant. Insets show the hemodynamic response (beta values extraction) for novel (blue) and practiced (green) trials. Asterisks indicate that the conditions differed significantly ($P < 0.05$, Bonferroni corrected) in the corresponding time bin.

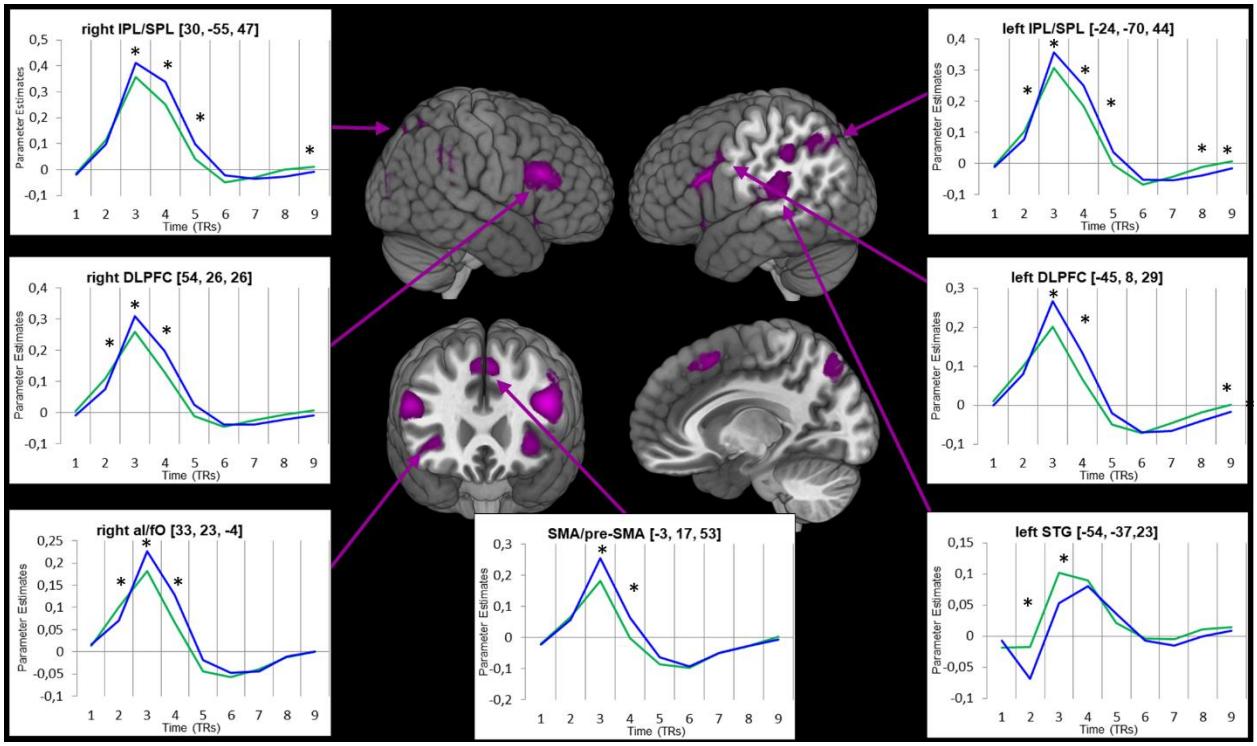

Figure 4.4: Results from the implementation stage ANOVA. Violet clusters show regions where the interaction of Experience and Time was significant. Insets show the hemodynamic response (beta values extraction) for novel (blue) and practiced (green) trials. Asterisks indicate that the conditions differed significantly ($P < 0.05$, Bonferroni corrected) in the corresponding time bin.

Table 4.1: Transient activity results.

Label	ANOVA term	Direction	Peak coordinate	Z value	k
Encoding phase					
Left aPFC	Main effect	P > N	-36, 44, 8	5.44	155
Left DLPFC	Main effect	P > N	-51, 20, 39	5.38	95
Right aPFC	Main effect	P > N	39, 47, 2	5.13	134
Right DLPFC	Main effect	P > N	48, 29, 29	4.64	91
Cerebellum (lobule VI)	Interaction	N > P	-33, -43, -22	4.14	60
Implementation phase					
Left LPFC	Interaction	N > P	-45, 8, 29	7.61	430
Right LPFC	Interaction	N > P	54, 26, 26	7.12	295
SMA/preSMA	Interaction	N > P	-3, 17, 53	6.86	177
Right SPL	Interaction	N > P	30, -55, 47	6.46	538
Left SPL/IPL	Interaction	N > P	-24, -70, 44	6.11	516
Right Fusiform gyrus	Interaction	N > P	48, -58, -13	5.86	225
Right aI/fO	Interaction	N > P	33, 23, -4	5.75	112
Left aI/fO	Interaction	N > P	-33, 23, -4	5.55	79
Left Caudate	Interaction	P > N	-21, 8, 26	5.43	57
Left SMG/STG	Interaction	P > N	-54, -37, 23	5.39	394
Left BG / posterior insula	Interaction	N > P	-33, -19, -1	5.35	104
Left fusiform gyrus	Interaction	N > P	-39, -46, -22	5.12	110
Right SMG/STG	Interaction	P > N	60, -34, 32	5.08	178
Right BG / posterior insula	Interaction	-	30, -19, 5	4.89	113
Right MTG	Interaction	-	48, -34, -10	4.73	48
Left MTG	Interaction	-	-48, -22, -4	4.68	27
Bilateral Caudate	Main effect	N > P	3, 8, -4	4.47	106
Right fusiform / PHG	Main effect	N > P	27, -31, -22	4.11	75

Note: The ANOVA terms refer to the main effect of Experience and the interaction of Experience with Time (see *Methods* sections). The direction indicates whether the activity was higher in novel (N) or in practiced (P) conditions, while hyphens designate regions with no clear directionality (because the significant interaction term is driven not by heightened activation but by different timing of the response). Abbreviations stand for anterior prefrontal cortex (aPFC), dorsolateral prefrontal cortex (DLPFC), lateral prefrontal cortex (LPFC), supplementary motor area (SMA), presupplementary motor area (preSMA), superior parietal lobe (SPL), inferior parietal lobe (IPL), anterior insula/frontal operculum (aI/fO), supramarginal gyrus (SMG), superior temporal gyrus (STG), basal ganglia (BG), middle temporal gyrus (MTG), parahipocampal gyrus (PHG).

Sustained activity.

We first aimed to detect areas showing sustained activity through long task blocks in comparison with rest, collapsing across all conditions. We did not observe any significant results in this analysis, nor when we did compare just practiced blocks

against baseline. On the other hand, sustained activity in novel blocks (vs. baseline) was found in the right aI/fO and bilaterally in the inferior parietal lobe (IPL), aPFC and DLPFC- also involving the IFS (Fig. 4.5A and Table 4.2). DLPFC and IPL were also significant when novel blocks were contrasted against practiced ones (Fig. 4.5B), providing support for their role for sustained control in new situations. Conversely, practiced blocks elicited higher sustained activity than novel ones in the ventromedial prefrontal cortex (vmPFC).

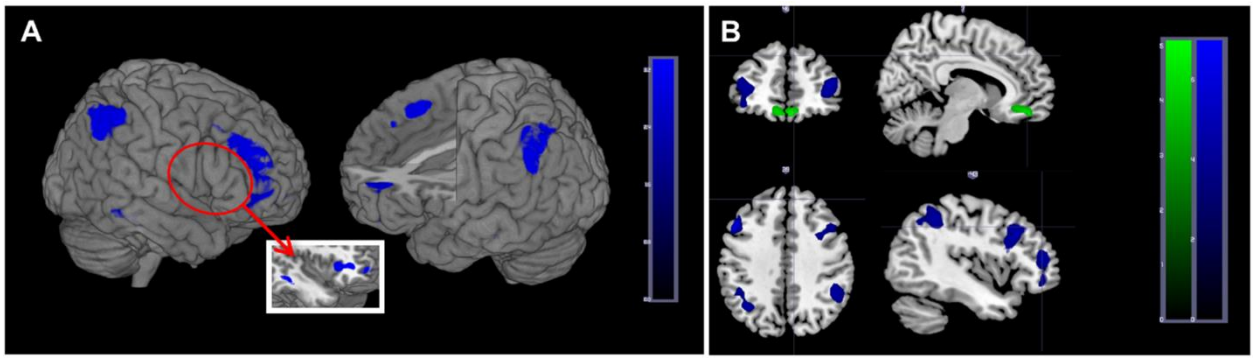

Figure 4.5: Sustained activity results. (A) Areas found in the *t*-test of Novel blocks against baseline. (B) Results from the contrast of novel versus practiced blocks. Clusters in blue show higher sustained activation in novel compared to practiced blocks, while the reverse is shown in green.

Conjunction analysis.

Results from our previous analyses suggested an overlap between regions with stronger sustained activity during novel blocks, and those with larger transient activity for the encoding of practiced instructions, and the implementation of novel ones. To quantify this observation, we performed an ad-hoc conjunction analysis with the corresponding three statistical maps obtained at the subject level (see *fMRI analysis* section). This test allowed us to confirm that one region, the left IFS, was involved across the three situations (Fig. 4.6).

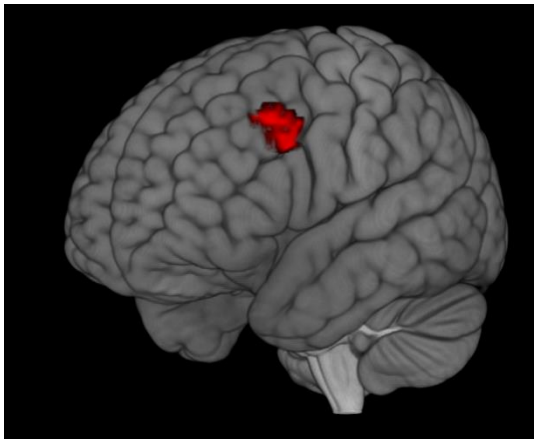

Figure 4.6: Results from the conjunction analysis. In red are voxels surviving to the conjunction test of (1) transient activity locked to practiced instructions encoding; (2) transient activity locked to novel instructions implementation and (3) sustained activity maintained through novel blocks. Peak coordinates: [-48, 20, 32], k = 63.

Table 4.2. Sustained activity results.

Label	**Block labels**	**Peak coordinate**	**Z value**	***k***
right IPL	Novel > Baseline	45, -52, 41	6.03	340
left IPL	Novel > Baseline	-42, -58, 47	5.27	330
left MTG	Novel > Baseline	-54, -31, -10	5.48	122
left aPFC/DLPFC	Novel > Baseline	-39, 47, 5	4.85	182
right aPFC/DLPFC	Novel > Baseline	39 53 -4	4.65	507
bilateral SMA/preSMA	Novel > Baseline	-9 17 53	4.59	213
right IFG/MTG	Novel > Baseline	57, -25, -19	4.49	136
right Cingulate gyrus	Novel > Baseline	9, -28, 26	4.18	262
left DLPFC/VLPFC	Novel > Practiced	-51, 20, 38	5.46	234
right aPFC	Novel > Practiced	39, 50, 5	4.37	142
right IFJ	Novel > Practiced	30, 11, 35	4.27	81
left IPL	Novel > Practiced	-36, -61, 41	4.26	155
right IPL	Novel > Practiced	48, -49, 44	4.2	134
left aPFC	Novel > Practiced	-45, 44, 14	4.16	143
bilateral vmPFC	Practiced > Novel	6, 47, -19	4.37	234

Note: Abbreviations stand for inferior parietal lobe (IPL), medial temporal gyrus (MTG), anterior prefrontal cortex (aPFC), dorsolateral prefrontal cortex (DLPFC), supplementary motor area (SMA), presupplementary motor area (preSMA), inferior frontal gyrus (IFG), middle frontal gyrus (MFG), lateral prefrontal cortex (LFPC), ventrolateral prefrontal cortex (VLPFC), inferior frontal junction (IFJ) and ventromedial prefrontal cortex (vmPFC).

Network comparison.

We also assessed the extent to which our principal sustained and transient results replicated previous findings regarding the involvement of two differentiable networks for cognitive control (Table 4.3): the CO and FP networks. Contrary to the framework put forward by Dosenbach and colleagues (Dosenbach et al. 2008), only the right aI/fO showed sustained activity throughout novel blocks, which just constituted 3.18% of the voxels of the CON template. Moreover, areas included in the FPN (bilateral IPS and the right IFS, involving a 42.92% of voxels of this network) were also present in the sustained activity maps.

At a transient time scale, the right aPFC, from the CON (4.69% of voxels), and the bilateral IFS and left IPS, from the FPN (18.61% of voxels), were involved during encoding of practiced instructions. During the implementation of novel ones, all ROIs of the FPN coincided with active clusters (although in an extent of just the 16.77% of the voxels), but were also accompanied by bilateral aI/fO from the CON (being, in this case, a 27.40% of CON voxels). Overall, the picture emerging from these comparisons is a mixture of CON and FPN involvement across both temporal modes of functioning.

Table 4.3. Transient and sustained signals at cingulo-opercular and fronto-parietal regions.

Region	Transient: encoding (practiced > novel)	Transient: implementation (novel > practiced)	Sustained (novel > practiced)
Cingulo-Opercular Network			
dACC	-	-	-
left aPFC	-	-	-
right aPFC	**X**	-	-
left aI/fO	-	**X**	-
right aI/fO	-	**X**	**X**
Fronto-Parietal Network			
left IFS	**X**	**X**	-
right IFS	**X**	**X**	**X**
left IPS	**X**	**X**	**X**
right IPS	-	**X**	**X**

Note: Crosses indicate the existence of overlap between the regions of interest of CO and FP networks (Dumontheil et al. 2011; Fedorenko et al. 2013) and results obtained for contrasts in the current whole-brain analysis. Abbreviations stand for dorsal anterior cingulate cortex (dACC), anterior prefrontal cortex (aPFC), anterior insula/frontal operculum (aI/fO), inferior frontal sulcus (IFS) and intraparietal sulcus (IPS).

Representational similarity analysis

In addition to the temporal profiles (transient vs. sustained) described above, differences between the CON and FPN may arise at a shorter time scale, within trial epochs. We explored this using RSA focused on the CON and FPN ROIs. We computed a consistency index associated with the maintenance of multivoxel representation of instructions from encoding to implementation stages (Qiao et al. 2017; see Fig. 4.2), in which larger values indicated a higher consistency along time (see *fMRI analysis* section). As expected, in all the regions examined, this index was significantly above 0 (all *P*s < .001 in one-sample *t*-tests) showing a correspondence between the information represented during novel instruction encoding and implementation. However, due to the temporal proximity of the source signal (consecutive events) this result could merely reflect the sluggish nature of BOLD response, although the jittered interval added between the

encoding and the implementation should prevent or minimize this problem. In any case, this potential confound does not affect our analysis as we only focused in the relative differences in the index between both networks.

We first collapsed across novel and practiced trials, and observed that the CON's consistency index was higher than the FPN's one ($T_{34} = 9.34$, $P < .001$), suggesting more persistent task-set representations in the former network. We then explored variations within ROIs of both subnetworks, with two additional repeated-measures ANOVAs. In both systems, the effect of ROI was significant (CON: $F_{4,136} = 91.84$, $P < .001$, $\eta_p^2 = .730$; FPN: $F_{3,102} = 30.64$, $P < .001$, $\eta_p^2 = .474$) and planned comparisons showed that the differences were statistically significant between each pair of regions, except when they involved left and right portions of the same area. Within the CO subnetwork, the region showing the highest consistency over time was the bilateral aPFC (left: M = 1.028, SD = .207; right: M = 1.017, SD = .219), followed by the dACC (M = .850, SD = .207) and, finally, the aI/fO (left: M = .669, SD = .204; right: M = .672, SD = .171). On the other hand, the bilateral IFS (left: M = .821, SD = .231; right: M = .776, SD = .187) showed larger consistency than the IPS (left: M = .623, SD = .177; right: M = .583, SD = .151) in the FPN.

Finally, to assess whether this pattern was modulated by instruction novelty, we conducted an ANOVA with this variable and Network as factors. As expected, the main effect of Network was significant ($F_{1,34} = 52.28$, $P < .001$, $\eta_p^2 = .606$), and importantly, so was the main effect of experience with the task ($F_{1,34} = 12.60$, $P = .001$, $\eta_p^2 = .270$). Specifically, practiced instructions showed a higher consistency index than novel ones (novel: M = 0.745, SD = .195; practiced: M = 0.836, SD = .191), indicating that the experience facilitated a more efficient task-set

maintenance within trials. The interaction term with Network was not significant, which suggests that the increase in similarity along the trial epochs with practice did not differ across CON and FPN regions.

4.4. Discussion

In this study we investigated which brain networks underpin instruction following, and their fit within the dual control model (Dosenbach et al. 2006; Dosenbach et al. 2007; Dosenbach et al. 2008). To do so, we adapted a mixed design to a paradigm in which different novel and practiced instructions had to be encoded and implemented, and extracted the underlying transient and sustained brain signals. Our hypothesis was that novel instructions would recruit the CON and the FPN to a higher extent than practiced ones: the former proactively -transiently during instruction encoding, and in a sustained fashion across trials-, and the latter reactively -linked to the implementation stage-. Our results showed that the transient involvement of different regions varied depending on practice and the information stage (encoding vs. implementation) of instructions. Moreover, regions from both FPN and CON were involved both in the sustained maintenance of activity during novel blocks and during transient rule implementation. Multivariate patterns of activation in both networks showed a consistent differentiation between CON and FPN in how the information was maintained across the encoding and implementation stages, as the former network seems to hold instruction representations more consistently along time, an effect that increases with practice.

The analysis of transient activations by means of FIR models allowed to study how novelty influenced the regions engaged in a phasic mode during complex verbal instruction processing. In line with previous research (Ruge and

Wolfensteller 2010; Dumontheil et al. 2011; Muhle-Karbe et al. 2017) we found that the IFS and the IPS, the main nodes of the FPN, were relevant at this time scale. Phasic activity was also found in the CON, concretely, in the aI/fO. In this sense, the whole pattern of regions presenting transient activity fits with our predictions based on Dosenbach's model (Dosenbach et al. 2006). However, to better understand these findings, it is important to consider the two different processes that unfold along the trial epoch. We studied the encoding of instructions, more related with proactive preparation, and the subsequent implementation phase, where rules were applied to concrete stimuli, closely linked to reactive adjustments. During the initial encoding, no regions were transiently more active for novel than for practiced instructions. Conversely, the bilateral IFS was more active for practiced instructions than for novel ones. Later on, during the implementation, the IFS was again recruited, together with the IPS, the aI/fO and the preSMA. Importantly, here these regions showed larger activity for novel than practiced instructions, replicating previous findings (Ruge and Wolfensteller 2010; González-García et al. 2017).

The increased recruitment of the IFS in practiced compared novel instructions encoding may seem at odds with previous literature and our own predictions. Nonetheless, this finding may reflect the difficulty of fully preparing novel complex instructions during the encoding stage -in opposition with overly practiced ones, which could automatically retrieve the proceduralized task-set during this initial stage. In agreement with this, it has been previously proposed that novel rule preparation culminates when they are first implemented in behavior (Brass et al. 2009; Cole et al. 2013), an effect that may have been potentiated by the increased complexity and abstraction of our instructions in

comparison with those used in previous research (e.g. Cole et al. 2010; Ruge and Wolfensteller 2010). As a result, the IFS activity may mediate practiced task-sets instauration and, as such, underlie a better proactive preparation in this condition. This is supported by the fact that this region has a relevant role in the preparation to implement instructions, in comparison with mere memorization demands (Demanet et al. 2016; Muhle-Karbe et al. 2017; Bourguignon et al. 2018).

Importantly, our conjunction results confirmed that the same left IFS cluster was present during the encoding of practiced instructions and the implementation of novel ones. Hence this region may underpin a preparatory process that can take place at different moments: earlier when the instruction is known (practiced) and its pragmatic representation can be retrieved, and later (i.e., when the stimuli are available) when we face a novel task, and this representation must be created from scratch. Nonetheless, which specific computations the IFS implements during this process is an open question. Different proposals have been made in the literature: binding of relevant stimuli and response parameters (Hartstra et al. 2012), mediating the transformation of semantic information into a pragmatic, action-oriented task representation (Ruge and Wolfensteller 2010), or maintaining the task-set in an active mode (Demanet et al. 2016), making it available for other lower-level regions. However, whereas novel instructions preparation seems to require the deployment of these three processes, practiced ones do not, as they do not need to be rebuilt but rather retrieved and updated. In light of our findings, therefore, task-set maintenance seems to be the most suitable common role underlying this region in both novel and practiced conditions. This is further supported by studies recording single and multiunit

activity in monkeys' LPFC (e.g., Freedman et al. 2001), which reveal the role of this area in the maintenance of different task-relevant information during delay periods.

Another remarkable set of results in the current study is the involvement of other regions during instruction implementation, such as the IPS and the preSMA. As implementation seems to rely to a high extent on reactive mechanisms, these regions may be implementing online control adjustments upon target presentation in novel trials, compensating for the less efficient proactive preparation during the encoding stage. From this perspective, the whole pattern of transient activations could be interpreted in terms of an interplay between proactive and reactive processes, which would depend on the novelty of the instructions that govern behavior. This interpretation fits with the balanced nature of proactive and reactive control modes: situations that weight proactive mechanisms to a higher extent trigger less reactive control, and vice-versa (Braver 2012). Nonetheless, it is also important to note that the temporal profile of activation of these brain areas is highly flexible. Whereas they have been linked to reactive functions (e.g., the preSMA seems to mediate the inhibition of irrelevant stimulus-response mappings in this context; Brass et al. 2009), patterns of activation consistent with proactive preparation have also been observed, such as increases of activity during encoding and preparation intervals (e.g. Hartstra et al. 2011; Dumontheil et al. 2011; Hartstra et al. 2012; Muhle-Karbe et al. 2014; Muhle-Karbe et al. 2017).

An additional core goal of our study was to extract sustained, block-wise activations to investigate whether a stable pattern of activation was maintained in CON and FPN areas during the execution of novel, demanding tasks, as it has

been shown previously in more repetitive experimental settings (Dosenbach et al. 2006). In accordance with our expectations, blocks of new instructions were associated with a larger sustained recruitment of frontal and parietal regions, when comparing against both pause periods and practiced blocks. Nonetheless, the regions involved were more consistent with the main nodes of the FPN: the bilateral IFS and the IPS. Only the right aI/fO region and part of the aPFC, from the CON, showed sustained activation in novel blocks. Accordingly, when we explicitly tested the percentage of overlapping voxels between two networks and our results, we found higher coherence with the FPN. Our results aid to qualify the dual model of control, showing that sustained activation patterns are not the exclusive fingerprint of CON regions. In contexts of novelty, when higher flexibility is needed, nodes of the FPN are also recruited at this timescale, while sustained activity is restricted to certain nodes of the CON. This result may seem at odds with previous evidence. However, the nature of the behavior analyzed in our research departs considerably from the one captured by most of previous mixed-designed studies (Dosenbach et al. 2006), as our experiment required the continuous building and updating of novel complex task-sets. It has been argued that the sustained activation across the CON underlies the maintenance of relevant rules as long as they are needed (Dosenbach et al. 2008). While this mechanism may be efficient when the task remains the same, it may not be beneficial in long blocks where rules change in a trial-by-trial fashion. Here, the FPN may implement sustained control processes independent of the specific task-set adopted on each trial. Due to the role of this network in establishing the widest and most flexible pattern of connectivity with other brain regions (Cole et al. 2013), one possibility is that sustained activity across FPN regions implements some kind of tonic state of high efficiency in information routing between

domain-specific regions. This view is supported by two different sources of evidence. First, task-dependent variability in the sustained engagement of CON has been previously reported, as in the case of perceptually driven tasks (Dubis et al. 2016). Second, sustained activity in lateral prefrontal and parietal cortices has also been found in studies which also relied on task-set updating: during blocks in which task switching was required (Marini et al. 2016), and while executing distinct instructions (Dumontheil et al. 2011). Overall, our findings highlight that both control networks, especially FPN areas, display a rather general ability to switch between phasic and tonic temporal modes depending on the nature of the tasks to be accomplished.

The result of our conjunction test, in which we identified common clusters at both phasic and tonic timescales, gains again relevance at this point. The same left IFS cluster involved transiently during the encoding of practiced instructions and the implementation of novel ones, which we propose underlie the maintenance of instructed task-sets, is also recruited in a sustained fashion through novel blocks. The relationship between the functions carried out at the two timescales is not straightforward; nonetheless, it is unlikely that they coincide, as this may result in an unnecessary redundancy across both timescales. It could well be the case that this and other regions perform distinct computations depending on temporal parameters, as previous neuroimaging data show that the LPFC, in general, can adopt different temporal dynamics (Jimura et al. 2010; Braver 2012). Results of the current investigation indicate that a demanding and rich task environment can recruit both temporal modes of functioning of this area, and moreover, that this profile is sensitive to the novelty of the situation. On the one hand, this evidence highlights the flexible nature of this brain region. On the other hand,

such results could reflect an organizational principle by which different cognitive computations are multiplexed in distinct temporal dynamics within brain areas.

Finally, we also explored multivoxel activity patterns in both networks' nodes, obtaining results consistent with the classic dual network model (Dosenbach et al. 2008). Areas within the CON represented task-sets more consistently over trial epochs, i.e., from encoding to implementation stages. This result strongly supports the proposal that these regions are in charge of maintaining information in a sustained, proactive fashion even in the absence of maintained univariate activation. Moreover, we found that this effect was affected by the experience with the trial: when the instructions were practiced, the consistency of the representation was higher, suggesting a possible mechanism by which the task representation gains in fidelity as it is repeatedly used. Interestingly, a recent study showed that task rule representation is more stable across the pre-target epoch when the instruction must be memorized in comparison with novel to-be-implemented ones (Muhle-Karbe et al. 2017). Overall, these results agree with the idea that novel trials require the semantic information of the instruction to be transformed into an action-related representation, a process that needs time to unfold and evolves up to target presentation. Moreover, this could explain why less reactive adjustments may be deployed when practiced instructions are translated into actions, as our results of transient activity during the implementation show.

Further research is needed to connect the scarce findings provided from this and other mixed design studies, and the broader cognitive control literature. For example, a recent study showed, employing MVPA, that task-sets were better encoded (i.e., decoded with higher accuracy) in FPN than in CON regions

(Crittenden et al. 2016). These findings are not incompatible with ours, as we used RSA and our analysis was focused in the transference of rule representation between two temporal time points -and not in classification accuracies at concrete time points of the task. Nonetheless, due to the decision of using a mixed design to extract transient and sustained activations, our experiment was not optimized for performing MVPA on our data. Previous research (González-García et al. 2017) has shown that regions consistent with both CON and FPN encode the relevant stimuli category of the instructions, before its implementation. Future studies will help to characterize, from this approach, which information is contained in transient and sustained activation patterns -and whether this is segregated between the two control networks. Finally, it is important to highlight that the extent of novelty entailed by each instruction was limited, given that the global task structure remained the same throughout the experiment. To study control mechanisms acting in novel contexts, we generated a large amount of trials including unique task rules and complex and also unique target combinations (Cole et al. 2010; Hartstra et al. 2011; González-García et al. 2017). However, target categories (faces and letters) and motor responses (employing the two index fingers) remained the same across the whole task. While fixing these parameters allowed us to exert experimental control, the complexity of novel situations that humans face daily is far richer and more variable. Future studies should aim for increasingly more ecological paradigms, where the general task structure also varies in a trial-wise fashion.

4.5. Conclusions

The current study provides insights about the dual network perspective of cognitive control, expanding this model to novel complex task contexts. Crucially,

results indicate that even when the two networks are functionally differentiated, both seem act at both tonic and phasic timescales during novel instruction processing. Furthermore, the division between proactive and reactive control does not seem to be mapped in a straightforward way into these two networks. Future studies must be conducted to further detail their contributions. Specifically, the computations and information held at the sustained time scale remain unknown, as also their relationship with mechanisms that develop at a faster, transient scale. The expansion of multivariate decoding techniques could help to better disentangle between the computational roles of both neural networks.

Chapter 5:

Representational organization of novel task sets during proactive encoding – Study 2

Published as:

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 Cognitive Neuroscience, 39(42), 8386-8397

https://doi.org/10.1523/JNEUROSCI.0725-19.2019

Abstract

Recent multivariate analyses of brain data have boosted our understanding of the organizational principles that shape neural coding. However, most of this progress has focused on perceptual visual regions (Connolly et al., 2012), whereas far less is known about the organization of more abstract, action-oriented representations. In this study, we focused on humans' remarkable ability to turn novel instructions into actions. While previous research shows that instruction encoding is tightly linked to proactive activations in fronto-parietal brain regions, little is known about the structure that orchestrates such anticipatory representation. We collected fMRI data while participants (both males and females) followed novel complex verbal rules that varied across control-related variables (integrating within/across stimuli dimensions, response complexity, target category) and reward expectations. Using Representational Similarity Analysis (Kriegeskorte et al., 2008) we explored where in the brain these variables explained the organization of novel task encoding, and whether motivation modulated these representational spaces. Instruction representations in the lateral prefrontal cortex were structured by the three control-related variables, while intraparietal sulcus encoded response complexity and the fusiform gyrus and precuneus organized its activity according to the relevant stimulus category. Reward exerted a general effect, increasing the representational similarity among different instructions, which was robustly correlated with behavioral improvements. Overall, our results highlight the flexibility of proactive task encoding, governed by distinct representational organizations in specific brain regions. They also stress the variability of

motivation-control interactions, which appear to be highly dependent on task attributes such as complexity or novelty.

Significance Statement

In comparison with other primates, humans display a remarkable success in novel task contexts thanks to our ability to transform instructions into effective actions. This skill is associated with proactive task-set reconfigurations in fronto-parietal cortices. It remains yet unknown, however, *how* the brain encodes in anticipation the flexible, rich repertoire of novel tasks that we can achieve. Here we explored cognitive control and motivation-related variables that might orchestrate the representational space for novel instructions. Our results showed that different dimensions become relevant for task prospective encoding depending on the brain region, and that the lateral prefrontal cortex simultaneously organized task representations following different control-related variables. Motivation exerted a general modulation upon this process, diminishing rather than increasing distances among instruction representations.

5.1. Introduction

Humans quickly learn from instructions which elements are relevant in a context and their respective appropriate actions. These parameters are encoded proactively in our brain in an action-based code (Brass, Liefooghe, Braem, & De Houwer, 2017; Cole, Braver, & Meiran, 2017), preparing our perceptual and motor systems in advance (Cole, Laurent, & Stocco, 2013) and facilitating success in novel environments. Instructed behavior is thus critical to avoid less effective and slow trial-and-error learning, and also enables the social transmission of task procedures. There is scarce knowledge, however, about how the informational and motivational content of novel instructions organizes neural activity in a proactive manner.

Behavioral results support the role of proactive control (Braver, 2012) on instructed action (e.g. Liefooghe, Wenke, & De Houwer, 2012; see also Cole, Patrick, & Braver, 2018; Duncan et al., 2008; Luria, 1966). Recently, neuroimaging studies have revealed a link between novel instruction preparation and the fronto-parietal (FP) network (e.g. Cole, Bagic, Kass, & Schneider, 2010; Hartstra, Kühn, Verguts, & Brass, 2011; Palenciano, González-García, Arco, & Ruz, 2019). The middle (MFG) and inferior (IFG) frontal gyri, and the inferior frontal sulcus (IFS), together with the intraparietal sulcus (IPS), encode novel instruction content both in multivoxel activity patterns (Bourguignon, Braem, Hartstra, De Houwer, & Brass, 2018; González-García, Arco, Palenciano, Ramírez, & Ruz, 2017; Muhle-Karbe, Duncan, De Baene, Mitchell, & Brass, 2017) and distributed functional connectivity (Cole, Laurent, et al., 2013). Crucially, the fidelity of information encoding is linked to the intention to implement the instruction (versus mere memorization demands; Bourguignon et al., 2018;

Muhle-Karbe et al., 2017) and it is also closely related to the efficiency of behavior (Cole, Ito, & Braver, 2016; González-García et al., 2017). Nonetheless, while current studies have mainly focused on decoding the upcoming target category (González-García et al., 2017; Muhle-Karbe et al., 2017), the wider organizational structure that shapes anticipatory task representation remains unknown. To study the relevant dimensions organizing novel instruction encoding, we selected three variables known to be relevant for proactive control.

Task preparation consists of a two-step process (Rubinstein et al., 2001), composed first by an abstract goal reconfiguration and second by the activation of specific stimulus-response contingencies (De Baene & Brass, 2014; Muhle-Karbe, Andres, & Brass, 2014). Our study exploited these two phases. First, in relation to the high-level task goal setting, we manipulated the integration of information within or across feature dimensions of stimuli (Rigotti et al., 2013), a variable traditionally linked to task complexity and top-down attention (e.g. Treisman & Gelade, 1980). Second, the stimulus-response reconfiguration process was manipulated by the response set complexity, requiring single or sequential motor responses. Moreover, to explore stimuli-specific preparatory mechanisms previously documented (e.g. González-García, Mas-Herrero, de Diego-Balaguer, & Ruz, 2016; Sakai & Passingham, 2003, 2006), we also manipulated the relevant target category.

Finally, cognitive control and motivation maintain an intricate relationship during task preparation (Pessoa, 2009, 2017). Reward expectation boosts cue-locked activity across the FP network (Parro, Dixon, & Christoff, 2017), and it has been recently linked to stronger anticipatory rule encoding (Etzel, Cole, Zacks, Kay, & Braver, 2016). Nonetheless, contradictory findings have also been found

(Wisniewski, Forstmann, & Brass, 2018), and a comprehensive characterization of this interaction in complex, novel scenarios is still pending. Consequently, we included economic incentives in our paradigm and assessed the nature of their effect on instruction preparation. By varying these four variables (dimension integration, response-set complexity, target category, and reward), we built a set of novel, verbal instructions that were followed by healthy participants while functional magnetic imaging (fMRI) data were collected. Using Representation Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008), we assessed the extent to which each of our control-related variables organized instruction encoding, as well as the effect of motivation upon this organization.

5.2. Materials and methods

Participants

Thirty-six students from the University of Granada completed the experimental paradigm inside an MRI scanner (16 women, mean age = 22.97 years, SD = 3.32 years). All of them were right-handed, with normal or corrected-to-normal vision, and native Spanish speakers. In exchange for their participation, they received between 20 and 40€, depending on their performance on the rewarded trials (see below). They all signed a consent form approved by the Ethics Committee of the University of Granada. Four participants were later excluded due to excess of head movement (> 3mm) or poor performance (<70% of correct responses).

Apparatus, stimuli, and procedure

For the experiment, we built a set of 192 different novel verbal instructions. Each instruction referred to two independent conditions about faces or food items that could be met or not by the upcoming grids, and their associated responses (e.g.:

"*If there are two women and an additional sad person, press A; if not, press L*"). The conditions in the instructions referred to several dimensions of the stimuli: gender (*woman, man*), race (*black, white*), emotion (*happy, sad*) and size (*big, small*) of faces, or kind (*fruit, vegetable*), color (*green, yellow*), form (*round, elongated*) and size (*big, small*) of food items.

Instructions were created by manipulating in an orthogonal manner (1) the ***Integration of*** *stimuli* ***dimensions*** **(**within vs. across dimensions**)**, (2) the ***Response set*** required (single vs. sequential) and (3) the ***Category*** of the relevant stimuli that they referred to (faces vs. food). For example, the instruction "*If there is a woman and there is a man, press A; if not, press L*" involves within-dimension integration (i.e., gender), requires a single response (a left –"A"– or a right –"L"– index button press) and is face-related. On the other hand, "*If there is a fruit and a small food item, press AL; if not, press LA*" requires across-dimension integration (the type of food and its size), demands a sequence of two button presses to respond and is food-related. Instructions referred to either 2, 3 or 4 stimuli of the target grid. Equivalent trials were created for the different levels of these three variables.

In addition, we included ***Motivation*** as another variable: half of the instructions were associated with the possibility of receiving an economic reward if responses were fast and accurate while the other half were non-rewarded. To do so, we split our 192 instructions into two equivalent sets in terms of the manipulations of the other independent variables, and also regarding the specific attributes specified (e.g., the same number of instructions referring to happy faces in both groups). We counterbalanced across participants the assignment of these two halves to the rewarded and non-rewarded conditions. The reward status of each trial was

indicated by a cue consisting on either a plus (+) or a cross (x) sign, in either silhouette or filled in black. We counterbalanced across participants whether they should attend to the shape (plus vs. cross) or the appearance (contour vs. filled sign) to obtain the reward information. This way, each participant had two different cues indicating each motivation condition, preventing a one-to-one mapping between reward expectation and visual cue identity, which otherwise could generate spurious confounds in further analysis.

For each instruction, we created two grids of stimuli, one that fulfilled the conditions instructed, and another one that did not. We counterbalanced them so that individual participants saw only one of the two instruction-grid pairings. All grids were unique combinations of images of 4 faces and 4 food items, which were pseudo-randomly selected from a pool of 32 pictures, composed by 16 faces pictures (8 different identities, half of them women and half men, half with happy expression and half with sad ones, half white and half black, appearing each of them in large and small sizes), extracted from the NimStim database (Tottenham et al., 2009), and 16 food pictures (8 different items, half of them vegetables and half fruits, half in green color and half in yellow, half with a round shape and half elongated, appearing each of them in large and small sizes) obtained from available sources on the internet (all of them with Creative Commons license). Upon target presentation, the responses required were always one or two sequential button presses, performed with the left ("A") and/or right ("L") index. The sequence of trial events is depicted in Figure 5.1. Each trial started with a jittered fixation point (0.5º), with a duration that ranged from 4500 to 7500ms, in steps of 500ms (mean = 5750ms). Then, a reward cue was presented (1.5 º; 2000ms), followed by the instruction (25.75º; 2500ms). Next a second jittered

fixation appeared (with the same characteristics as the previous one), and the target grid (21º) was presented for 2500ms, where participants were required to respond. Afterward, a feedback symbol was presented (1.65º; 500ms), indicating whether the participant had earned money in that trial (with a Euro symbol), whether the response was correct but no money was achieved (tick symbol) or whether the response was incorrect (cross symbol).

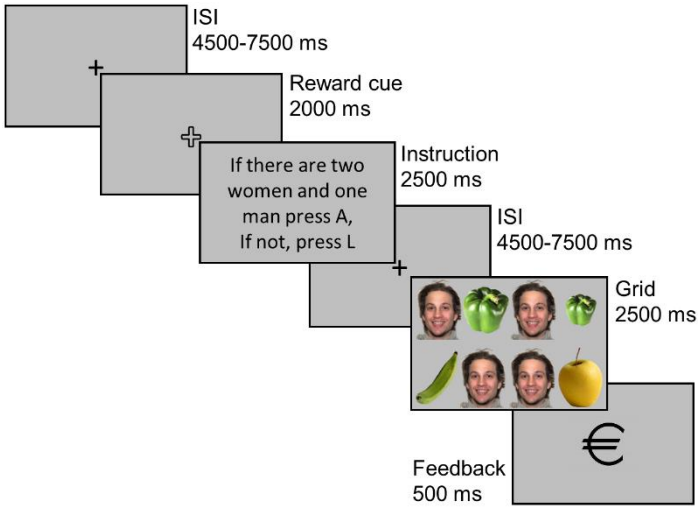

Figure 5.1: Sequence of events in a single trial.

Before being scanned, participants completed a behavioral practice session. They received indications about how to perform the task, as well as details on how rewards would be administered, emphasizing that both accurate and fast responses were needed to accumulate money for a maximum of 40€. Specifically, they were informed that they would receive 20€ for their time and that the rest of the compensation would depend on their performance on rewarded trials: the initial extra increases would be easier to earn while approaching the upper limit of the payment would require a higher accuracy rate. Then, they performed a simple discrimination task with the different reward cues, and after that, they practiced the instruction-following task, completing one block of 32 trials. Practice instructions were drawn from a separate set (which was equivalent in all the parameters specified above) and were not employed in the MRI experiment, to maintain trial novelty. Participants repeated the practice block as many times as needed to obtain an accuracy rate above 75% (on average, participants

performed the practice block 1.75 times). Once this phase was completed, the experimental paradigm was performed inside the scanner. This was composed by the full 192 instructions set, presented in six different runs (32 trials each). All runs included an equal number of face and food-related, single and sequential responses, within and across-dimension integration and rewarded and non-rewarded instructions. Overall, participants spent 90 minutes approximately inside the MRI scanner.

Experimental Design and behavioral statistical analysis

Our task was built following a 4-way factorial design, in which the following within-subjects independent variables were orthogonally manipulated: (1) Dimension integration; (2) Response set complexity; (3) Target category and (4) Reward.

We conducted an a priori power analysis to compute sample size. Using the PANGEA software (https://jakewestfall.shinyapps.io/pangea/), we calculated the minimum number of participants to detect a behavioral two-way interaction term (i.e., between reward and any other proactive control-related variable), assuming a medium effect size (Cohen's d = .3).

We used IBM SPSS Statistics v20 software to analyze accuracy and reaction time data. We conducted two repeated-measures ANOVAs, specifying four factors corresponding to our independent variables. To explore significant interaction terms, we carried out further post hoc tests, using a Bonferroni correction for multiple comparisons.

fMRI preprocessing

MRI data were acquired using a 3-Tesla Siemens Trio scanner located at the Mind, Brain, and Behavior Research Center (CIMCYC, University of Granada, Spain). Functional images were collected employing a T2* Echo Planar Imaging (EPI) sequence (TR = 2210ms, TE = 23ms, flip angle = 70º). Each volume consisted of 40 slices, obtained in descending order, with 2.3mm of thickness (gap = 20%, voxel size = $3mm^3$). A total of 1716 volumes were obtained, in 6 runs of 286 volumes each. We also acquired a high-resolution anatomical T1-weighted image (192 slices of 1mm, TR = 2500ms, TE = 3.69ms, flip angle = 7º, voxel size = $1mm^3$).

The functional images were preprocessed and analyzed with SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/), with the exception of single-trial parameter estimation (see *RSA section*), which was conducted on AFNI. After discarding the first four volumes of each run to allow for stabilization of the signal, the images were spatially realigned and slice-time corrected. Then, the participants' structural T1 image, which had been coregistered with the EPI volumes, was segmented to obtain the transformation matrices needed to normalize the functional images to the MNI space. Finally, they were smoothed with an 8mm FWHM Gaussian kernel. The full preprocessing pipeline was completed before conducting the univariate analysis, while only realigned and slice-timing corrected images were employed for the multivariate tests (see next section). In the latter, normalization and smoothing were performed after the individual-level analysis, following the same strategy as above.

fMRI statistical analysis

Control univariate analysis

We first conducted a univariate standard GLM, modelling each of the sixteen combinations of our variables (for example: within-dimension integration/simple response required/faces-related/ rewarded) and specifying two regressors per trial: one for the encoding phase (from the reward cue until the end of the instruction), and another for the implementation stage (encompassing the target grid presentation and until the end of the feedback cue). All regressors were convolved with the canonical hemodynamic response function. We also added error trials and six motion parameters as nuisance regressors, and a high-pass filter of 128s to avoid low-frequency noise.

The rationale of this analysis was to check the effect of motivation during the encoding of novel instructions with the aim of ensuring that our manipulation successfully generated typical reward-related patterns of activation (Parro et al., 2017). This was done by performing *t*-tests at the individual (first) level, contrasting rewarded versus non-rewarded encoding regressors, and carrying these statistical maps to a group one-sample *t*-test. The result was cluster-wise FWE-corrected for multiple comparison at $P < .05$ (from an initial threshold of $P < .001$ and k = 10). With this approach, we obtained one large cluster that extended across multiple brain regions. To obtain smaller, anatomically coherent clusters, we employed a stricter threshold (uncorrected cluster-forming threshold of $P < .0001$, with the corresponding FWE correction at $P < .05$), as done previously (e.g. Dumontheil et al., 2011; Palenciano et al., 2019).

Representational Similarity Analyses

We conducted a series of multivariate RSAs, following a two-step approach. First, we analyzed whole-brain data, using a searchlight approach, to find regions encoding novel instructions according to each of our three control-related variables. Second, we used the significant areas as Regions Of Interest (ROIs) and focused on them to explore the effect of reward on their representational geometry.

Whole-brain model-based RSA. We first studied whether the representational structure of novel instructions was explained by three variables related to cognitive control preparation: dimension integration, response set complexity and target category. Importantly, we specifically wanted to explore this during the initial encoding stage, where proactive task-set reconfiguration takes place. To do so, we first obtained trial-by-trial estimations of our signal, following a Least-Square-Sum approach (LSS; Turner, 2010) to ensure the smallest possible collinearity among regressors (Arco, González-García, Díaz-Gutiérrez, Ramírez, & Ruz, 2018). We generated and estimated one separate model per trial, in which we defined: (1) a regressor isolating the encoding phase of the individual trial of interest; (2) a second regressor containing the rest of trials (encoding phase) of the same condition; (3) thirty-one additional regressors encompassing the rest of conditions at the encoding and implementation phases (as in the GLM specified above), and (4) nuisance regressors (movement, errors). To do so, we employed AFNI's function 3dLSS (https://afni.nimh.nih.gov/pub/dist/doc/program_help/3dLSS.html). Once the trial-wise parameter images were obtained, the rest of the RSA was performed with The Decoding Toolbox (Hebart, Görgen, & Haynes, 2014).

In our analysis, we compared three theoretical models of representational organization (one per preparation-related independent variable) with the empirical one, built from spatially distributed activity patterns. To do so, we employed a spherical searchlight (radius: 4 voxels) and applied it to the whole brain (Kriegeskorte, Goebel, & Bandettini, 2006). First, we built three theoretical representational dissimilarity matrices (RDM, Fig. 5.2.a), which captured the expected dissimilarity (represented with 0s and 1s) between pairs of trials, according to the corresponding variables of interest. For example, in the Category RDM, dissimilarity is expected to be minimal within pairs of trials that refer either to faces or to food, while maximal between pairs of trials referring to different target categories. Then, in each iteration of the searchlight, we generated a neural RDM, using a measure of distance based on Pearson correlation. Specifically, we extracted the corresponding single-trial beta values of the voxels involved, correlated each pair of the trials' activity patterns, and subtracted that value from 1. Afterwards, this neural RDM was Spearman-correlated with the theoretical ones (Fig. 5.2.c), and the coefficients were normalized with Fisher's z transformation and assigned to the central voxel of the searchlight sphere. Importantly, both theoretical and neural matrices were built trial-wise (i.e., not averaging within conditions), and thus, were fully symmetrical with a diagonal of 0s. Consequently, only the lower triangle of the matrices, excluding the diagonal, was included in the correlation to avoid inflated positive results (Ritchie, Bracci, & Op de Beeck, 2017). After iterating the searchlight across the whole brain, we obtained three maps per participant representing how well the representational geometry in different regions matched the one expected by each of our three theoretical models.

Statistical significance was assessed non-parametrically via permutation testing, as proposed by Stelzer, Chen, & Turner (Stelzer, Chen, & Turner, 2013). We first performed 100 permutations at the individual level, where trial labels were randomly shifted and the whole analysis was repeated. Then, at the group level, we resampled 50,000 times one of the permuted maps of each subject and averaged them. The resulting bootstrapped group maps were used to build a voxel-wise null distribution of correlation values, which was used to extract the correlation coefficient coinciding with a probability of 0.001 of the right-tailed area of the distribution (i.e., linked to a p <= .001) of each individual voxel. The group map of the results was then thresholded using these values. From the bootstrapped maps we also built a null distribution of cluster sizes (Stelzer, Chen, & Turner, 2013), which determined the probability of each cluster extent under the null distribution. We used this to assign the corresponding *P* value to the surviving clusters of the group results map, and FWE-corrected (P < .05) them to control for multiple comparisons.

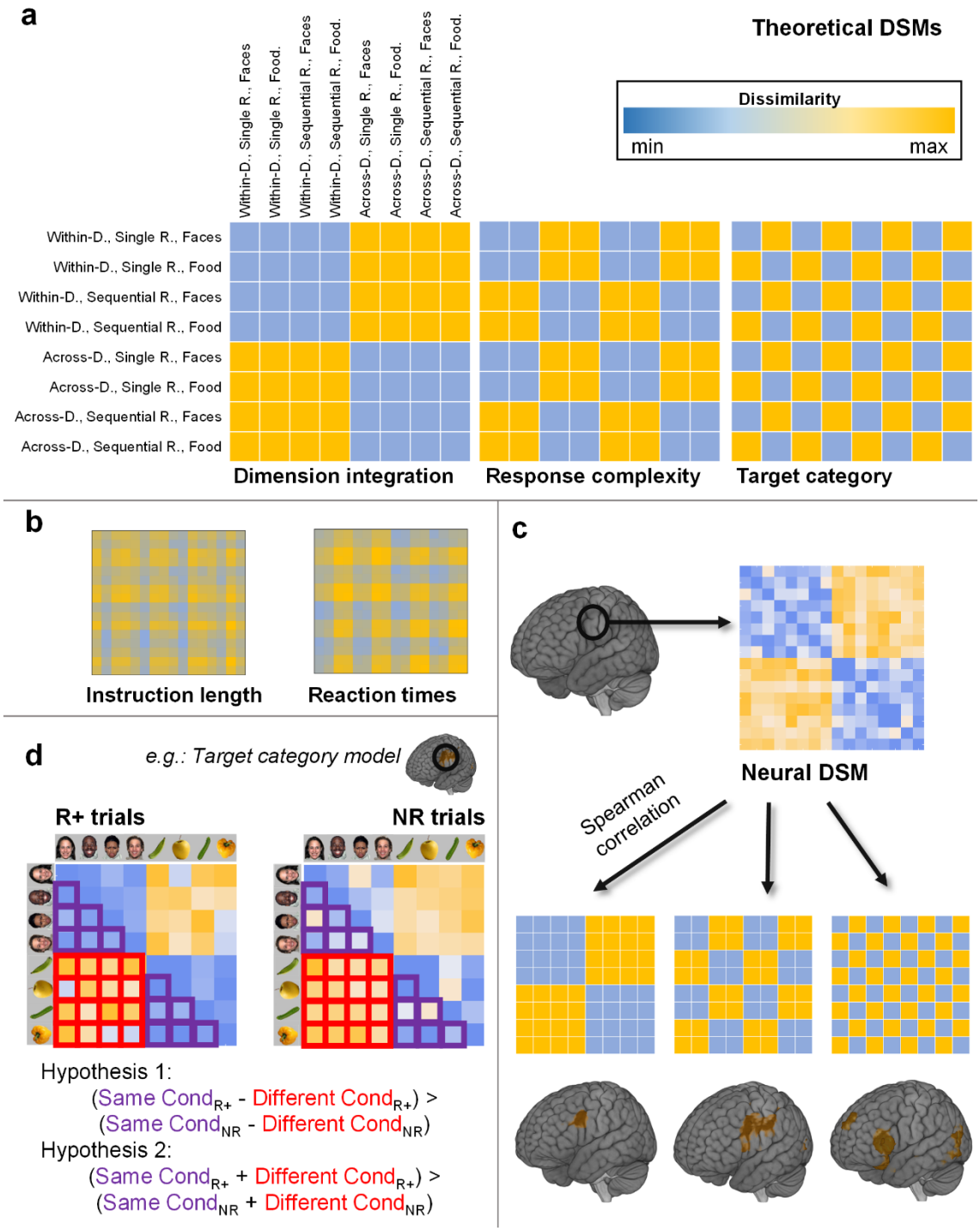

Figure 5.2: Main analysis procedure. (a) Theoretical Representational Dissimilarity Matrices (RDMs) employed in the Representational Similarity Analysis (RSA). Within/Across-D. stands for within-dimension and across-dimension integration, while Single/Sequential R. stands for single response and sequential response. (b) RDMs capturing differences in instruction length (number of letters) and reaction time, included in a multiple regression analysis together with matrices shown in (a) to control for the effect of these two variables. (c) Following a searchlight approach, we extracted the neural RDM at each brain location and compared it – via Spearman correlation – with our three theoretical RDMs. As a result, we obtained three whole-brain correlation maps, one per model. (d) To assess the effect of motivation, for each region significant in (c) we extracted the neural RDMs from rewarded (R+) and non-rewarded (NR) trials. To study potential interactions of reward expectation and the corresponding model variable (Hypothesis 1), we averaged the dissimilarity values among same-condition and different-condition trials and tested if the subtraction among these two values was higher in the rewarded condition (using Wilcoxon signed-rank test). We also checked for a general increase in dissimilarities associated to reward (Hypothesis 2). *Note*: All matrices in the figure were simplified for visualization purposes by averaging cells within conditions. The matrices shown in (b) were further averaged across the sample. In (d), matrices display only one task variable (collapsing between the remaining two) to highlight the analysis logic. In all the analyses, however, trial-wise and single subject matrices were employed.

We performed a further conjunction test to find areas sharing the three representational organizational schemes. To do so, we thresholded ($P < .05$, FWE corrected) and binarized the three maps from the previous step, and obtained the overlapping voxels (Nichols, Brett, Andersson, Wager, & Poline, 2005).

Importantly, the RSA results could be influenced by other variables statistically related to our manipulations (Popov, Ostarek, & Tenison, 2018), such as instructions' length and speed of responses, which differed slightly between conditions. To examine their influence on the results, we performed an additional multiple regression analysis taking both variables into account. We built two different RDMs (see Fig. 5.2.b) in which each cell contained the absolute difference in the number of letters (instruction's length RDM) or reaction time (response speed RDM), respectively, between specific pairs of instructions. We then used them as regressors together with the three proactive control-related RDMs, predicting the neural pattern of dissimilarities in each iteration of a searchlight. The regressors were built vectorizing the lower triangle of the RDM, excluding the diagonal values. It is important to note that there were small but still significant correlations among some of the regressors included in the analysis. Specifically, dimension integration correlated with instruction length and RT, and target category did so with instruction length. To assess the impact of these correlations on the regression estimation, we computed Variance Inflation Factors (Mumford, Poline, & Poldrack, 2015), an index of the regressors' collinearity. For our five models, and in all the participants, VIF were always below 1.1 (being 5 a typical cutoff above which the estimation would be compromised; Mumford et al., 2015). Thus, even despite the relationship among variables, the results of our main analyses are still meaningful. The corresponding

beta weight maps obtained showed the regions where the effect of our variables of interest remained significant even when instruction's length and response speed were included.

Finally, even when the distance measure employed to build the neural RDMs (i.e., Pearson correlation) is insensitive to differences in mean signal intensity between conditions, differences in signal variance could be affecting it (Walther et al., 2016). For that reason, these analyses as well as the reward-related tests (see below), were repeated after a z-normalization of the multivoxel activity patterns, ensuring equal mean (0) and standard deviation (1) across all pairs of trials. The results thus obtained did not differ from the initial non-normalized ones, so we do not report them here.

ROI-based RSA. The previous analysis identified brain areas encoding instructions according to each one of three proactive control variables, separately. We next ran ROI analyses to further explore the role of the three variables for task coding in these regions. Specifically, we estimated the extent to which each of the manipulated control variables explained the neural organization in the ROIs identified in the previous analysis. We followed a Leave-One-Subject-Out (LOSO) cross-validation procedure (Esterman, Tamber-Rosenau, Chiu, & Yantis, 2010), using the searchlight maps obtained before. First, we identified regions sensitive to each of the three models for each participant, running a group level *t*-test with the corresponding maps from the rest of the sample, i.e., excluding their own data. Significant clusters showing consistency across all LOSO iterations were selected as ROIs, and inverse normalized to the participants' native space. In a second step, we estimated the ROIs RDMs and correlated them with the three models RDMs. Importantly, thanks to the LOSO procedure we avoided circularity

in the analysis, as independent data was employed to select the ROIs and to compute de correlations with the models. The correlation coefficients (for each participant, one per ROI and model) were then introduced in a repeated measures ANOVA, with ROI and Model as factors, and the interaction term was examined to detect heterogeneity in task encoding organization across regions (Reverberi, Gorgen, & Haynes, 2012). Interactions were further characterized by one sample *t*-tests, in order to determine which models had an effect on the different regions studied. Whenever the normality assumption was not met (assessed with the Saphiro-Wilk test), we employed Wilcoxon signed-rank tests instead. All *P* values were Bonferroni-corrected for multiple comparisons, adjusting them to the number of ROIs explored.

Additionally, we aimed to extrapolate our findings to regions consistently found in the literature during both practiced (e.g. Woolgar, Hampshire, Thompson, & Duncan, 2011) and novel (e.g. González-García et al., 2017) task preparation, and in general, when demanding cognitive processing is deployed (Duncan, 2010). This set of brain areas belong to the Multiple Demand Network (MDN; Duncan, 2010), which includes the bilateral RLPFC, MFG, IFS, anterior insula/frontal operculum (aIfO) area, IPS, anterior cingulate cortex (ACC) and pre-supplementary area (preSMA). To assess the organization of novel task encoding across this MDN, we employed functionally derived masks of its nodes (from Fedorenko, Duncan, & Kanwisher, 2013; template available at http://imaging.mrc-cbu.cam.ac.uk/imaging/MDsystem), inverse normalized them to the participants' native space, and followed the same ROI-approach as above, extracting each ROI RDM and correlating it with the models' matrices. Again, correlation coefficients were entered into a repeated measures ANOVA

with ROI and Model as factors, interactions were examined, and finally, a series of one-sample *t*-tests (or Wilcoxon signed-rank test when normality was violated) were conducted.

Analysis of reward-related effects on RSA results. A final goal of our study was to assess whether the representational space of novel instructions was affected by motivation. Our initial hypothesis was that reward would polarize the representational geometry, enhancing the effect of our control-related variables at structuring rule encoding. In other words, and taking as an example the target category variable, we assessed whether reward expectations would increase the distance between representations of instructions referring to different stimulus categories (in extension to the other variables, indicated as *different-condition dissimilarity*), while decreasing the distance among those referring to same target category (*same-condition dissimilarity*). Our second, alternative hypothesis was that reward would exert a general effect, globally increasing the distances among instruction representations, independently of the other variables manipulated. In this sense, we expected that both *different* and *same-condition dissimilarity* would be increased in rewarded trials, in comparison with non-rewarded ones. The two possibilities would be compatible with previous findings showing that reward expectancy enhances rule decodability (Etzel et al., 2016).

To test these two hypotheses, we run ROI analyses (Fig. 5.2.d) for each of our control-related variables, focusing on the regions that resulted statistically significant in the main RSA. To do so, at the individual level and for each variable, we first ran a searchlight and generated four whole-brain maps containing dissimilarity values among: (1) same-condition rewarded trials; (2) different-

conditions rewarded trials; (3) same-condition non-rewarded trials; and (4) different-conditions non-rewarded trials. These values were the result of averaging and normalizing (with the Fisher transformation) the pertinent cells of the neural RDM (see Fig. 5.2.d for an example) in each searchlight iteration. The maps thus obtained were normalized to the MNI space, so we could extract participants' mean dissimilarities for each of our ROIs using MarsBar (Brett, Anton, Valabregue, & Poline, 2002). After that, and for each ROI and variable, we conducted two Wilcoxon signed-rank tests (Nili et al., 2014). First, to assess our main hypothesis, we tested whether ($DifferentCond._{Rewarded}$ - $SameCond._{Rewarded}$) > ($DifferentCond._{NonRewarded}$ - $SameCond._{NonRewarded}$). To explore the second possible hypothesis, we collapsed across same and different conditions, and tested if ($DifferentCond._{Rewarded}$ + $SameCond._{Rewarded}$)/2 - ($DifferentCond._{NonRewarded}$ + $SameCond._{NonRewarded}$)/2 was greater than 0 (Fig 5.2.c). In both analyses, we corrected for multiple comparisons (number of ROIs being tested) with an FWE threshold of $P < .05$.

Last, to investigate the relevance for behavior of the effect of motivation on representational structure, we correlated this effect with behavioral data. Specifically, for each participant, we computed the average decrease in dissimilarity and in the inverse efficiency scores (IES; Townsend & Ashby, 1978) linked to rewarded trials (in comparison with non-rewarded ones). The IES was employed in this analysis to take into account, simultaneously, improvements in accuracy and response speed. As we performed as many correlations as ROIs assessed in this analysis, we again controlled for multiple comparisons with an FWE threshold of $P < .05$.

Additionally, to explore the possibility of motivation exerting an effect during the subsequent implementation of instructions, we also ran the analyses detailed above with beta images obtained from this stage.

MVPA-based assessment of reward effects.

Finally, to further connect our results with previous findings, we performed multivoxel pattern analysis (MVPA) to explore the effect of reward on decoding precisions (Etzel et al., 2016). We decoded the two conditions of each of our three control-related variables, training three binary classifiers: one for distinguishing between within versus across-dimension integration instructions, other for single versus sequential response requirements, and the last one for faces and food-related trials. This was done separately for rewarded and non-rewarded trials. Again, we used non-normalized and unsmoothed trial-wise beta images from the encoding stage. As we aimed to detect any region with reward-related increases in task decodability, we performed the MVPA in a whole brain fashion, using searchlight (instead of biasing the results using ROIs resulting from the RSA). In each searchlight iteration, we followed a leave one-run-out cross-validation approach, training a linear support-vector machine classifier (C=1; Pereira, Mitchell, & Botvinick, 2009) with five of our six runs, and testing it with the remaining one, in an iterative fashion. Then, for each of our variables, we subtracted the accuracy map obtained from non-rewarded trials to the map from rewarded ones, and then normalized and smoothed these images, to conduct an above zero one-sample *t*-test at the group level. This way, we assessed the benefits in classification precision associated with reward.

5.3. Results

Behavioral results

We analyzed RT and accuracy data separately, conducting two repeated measures ANOVA with four factors, corresponding to the four variables manipulated: dimension integration (within vs. across), response set complexity (single vs. sequential), category (faces vs. food items) and motivation (rewarded vs. non-rewarded). Importantly, the main effect of motivation was statistically significant on both accuracy ($F_{1, 31} = 4.97$, $P < .05$, $\eta_p^2 = .14$) and RT ($F_{1, 31} = 6.52$, $P < .05$, $\eta_p^2 = .17$) data, with more accurate (rewarded: M = 0.85, SD = 0.11; non-rewarded: M = 0.83, SD = 0.12) and faster (rewarded: M = 1.16, SD = 0.21; non-rewarded: M = 1.20, SD = 0.20) responses on the rewarded condition (see Fig. 5.3). This indicates that participants made use of reward cues and the economic incentives had the expected effect on behavior, improving its efficiency.

In addition, accuracy data showed a main effect of dimension integration ($F_{1, 31} = 9.24$, $P < .05$, $\eta_p^2 = .23$), with better performance when within-dimension integration was required (within dimension: M = .86, SD = 0.13; across dimensions: M = .83, SD = 0.12), and a significant three-way interaction of category, response set complexity and dimension integration ($F_{1, 31} = 4.46$, $P = .043$, $\eta_p^2 = .13$). Even despite the lack of hypothesis regarding an interaction at this level, we performed post hoc pair-wise comparisons, which revealed that the interaction was driven by less robust ($P > .05$) differences among within and across-dimensions trials that required a single response and was food-related (while, in the rest of combinations of independent variables, this difference was significant).

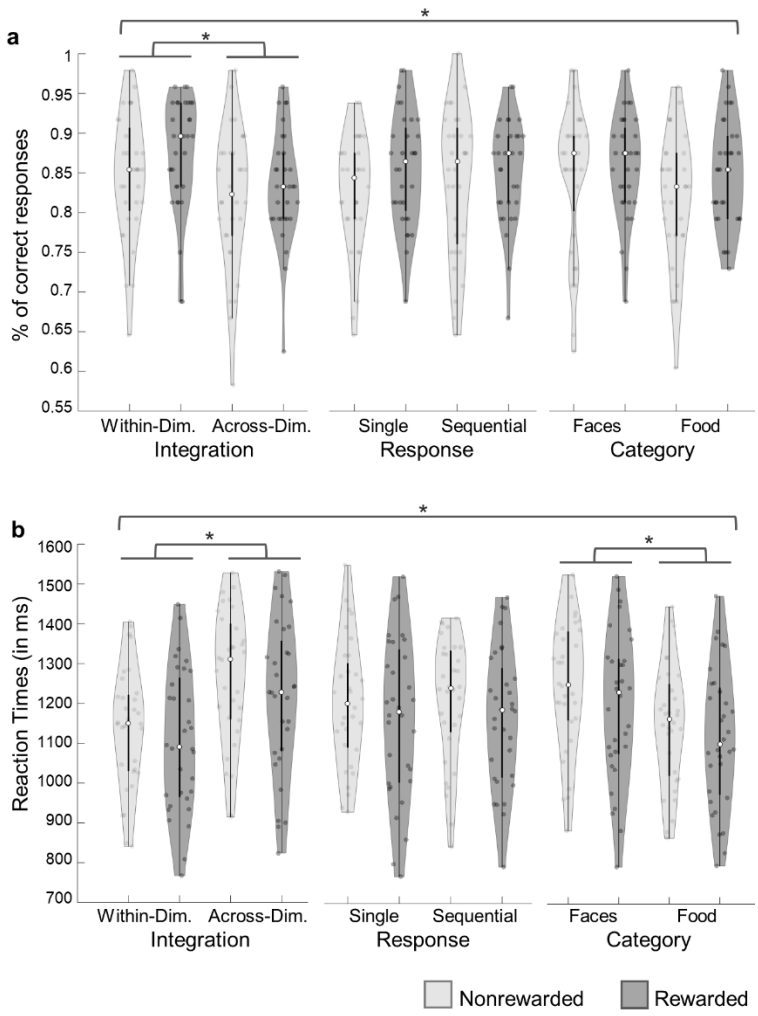

Figure 5.3: Behavioral data. Violin plots showing correct responses (a) and Reaction Time (b) data for each condition, in rewarded and non-rewarded trials.

On the other hand, RT results also showed a main effect of dimension integration ($F_{1, 31}$ = 61.81, P < .001, η_p^2 = .67) in the same direction as above (within-dimension: M = 1.12, SD = 0.17; across-dimensions: M = 1.24, SD = 0.2), and a main effect of category ($F_{1, 31}$ = 74.89, P < .001, η_p^2 = .71), with faster responses to food-related instructions (faces: M = 1.23, SD = 0.21; food items: M = 1.14, SD = 0.19). Neither the effect of response set complexity (accuracy: $F_{1, 31}$ = 0.31, P = .579, η_p^2 = .01; reaction time: $F_{1, 31}$ = 0.21, P = .653, η_p^2 = .01) nor any other ANOVA term resulted significant in the behavioral measures (main effect of

Category on accuracy: $F_{1, 31} = 3.23$, $P = .082$, $\eta_p^2 = .094$; all interactions terms, except the ones stated above, $P > .100$).

Univariate results: reward-related activations during instruction encoding.

We first assessed mean activity during novel instruction encoding, comparing rewarded against non-rewarded trials. To do so, we performed a univariate GLM, defining regressors for each combination of variables (e.g.: within-dimension integration, single response, face-related rewarded trials), separately for the encoding and the implementation stages. A group level *t*-test showed that, in accordance with our expectations and previous literature (Parro et al., 2017), the basal ganglia and fronto-parietal cortices were more active for rewarded than non-rewarded instruction encoding. We observed peaks of activation (see Fig. 5.4) in the bilateral inferior frontal junction (IFJ), premotor and supplementary motor areas (left: [-33, 5, 26], z = 5.07, k = 489; right: [33, 2, 59], z = 4.79, k = 572), cingulate cortex ([-9, 5, 32], z = 5.48, k = 20), bilateral IPS extending into the precuneus (left:[-18, -64, 35], z = 4.77, k = 357; right: [33, -52, 53], z = 4.36, k = 324), the accumbens, ventral portion of the caudate and thalamus ([12, -22, 20], z = 5.13, k = 1176), inferior temporal gyrus ([48, -58, -13], z = 4.48, k = 52), occipital cortex ([30, -61, -25], z = 5, k = 1364) and midbrain ([0, -31, -4], z = 5.19, k = 255). Thus, regions involved in reward processing (Haber & Knutson, 2009), as well as in cognitive control paradigms with monetary incentive manipulations (e.g. Engelmann, 2009), were engaged by our task, indicating the success of the reward manipulation.

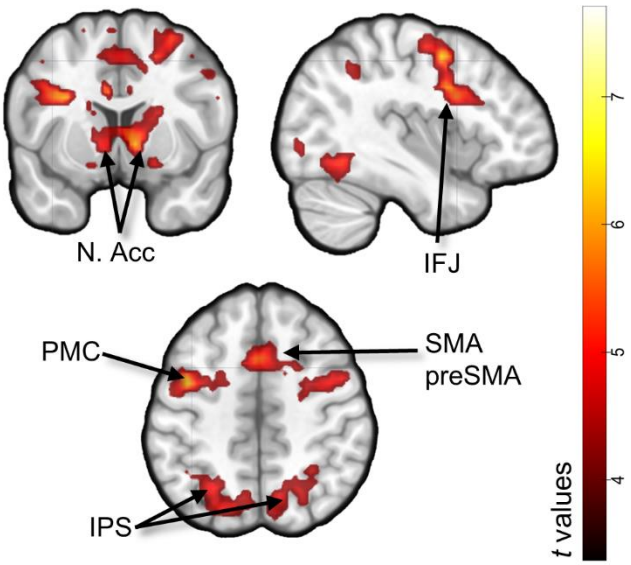

Figure 5.4: Regions showing greater activity during the encoding of rewarded compared to non-rewarded instructions. Abbreviations stand for Nucleus Accumbens (N. Acc), inferior frontal junction (IFJ), premotor cortex (PMC), supplementary motor cortex (SMA), pre-supplementary motor cortex (preSMA) and intraparietal sulcus (IPS).

Model-based RSA results: instruction encoding structured by proactive-control variables.

We aimed to identify regions whose organization during task encoding was explained by dimension integration, response set complexity and target category. With that purpose, we employed an RSA (Kriegeskorte, Mur, & Bandettini, 2008) to compare the representational dissimilarity matrices (RDMs) found in neural data during the encoding stage with theoretical RDMs corresponding to the three proactive control-related variables (see Fig. 5.2). In neural RDMs, each cell contained the dissimilarity (1 – Pearson correlation) between the multivariate patterns of activation of two trials. In the theoretical RDMs, cells contained dissimilarities (1: maximal, 0: minimal) that we would expect if a certain variable organized encoding (i.e.: for target category, all faces-related trials would be minimally dissimilar, while face and food-related trials would be maximally

dissimilar). Using searchlight (Kriegeskorte et al., 2006), we Spearman-correlated neural and theoretical RDMs across the brain and obtained maps showing how well these three variables captured the representational space of different areas. The modality of **dimension integration** (Fig. 5.5.a) only had a significant effect on rule encoding at the left MFG and IFG, incurring into the IFS ([-51, 20, 26], k = 642). **Response set complexity** (Fig. 5.5.b), on the other hand, organized task representations on a wide cluster including the bilateral IFG, premotor, supplementary and primary motor cortices, somatosensory area, middle temporal gyrus and superior and inferior parietal lobe extending along the IPS ([-42, -31, 44], k = 8583) and in the left parahippocampal cortex ([-18, -40, -1], k = 301). Finally, in the case of the **target category** RSA (Fig. 5.5.c), significant correlations were found in an extensive cluster on the left hemisphere covering the IFG incurring into the IFJ, the fusiform gyrus, the temporo-parietal junction (TPJ), the inferior and middle temporal gyrus and the precuneus ([-39, -67, 17], k = 5581). On the right hemisphere, the analysis was also significant on the right middle temporal gyrus and TPJ ([39, -58, 23], k = 442) and the IFG ([42, 26, 14, k = 295]. Finally, the medial superior frontal gyrus ([-9, 53, 26], k = 377) was also involved.

As instructions' length and speed of responses varied among some of our variables, we performed an additional multiple regression analysis, in which we included our three theoretical models, an RDM based on dissimilarities in length, and another one based on RT as regressors. Importantly, the multiple regression statistical model was examined to detect an excess of collinearity which could have impaired the interpretability of these results. We computed the VIF for all the regressors and across our whole sample of participants, and all of were under

1.1, an index of good estimability of regression weights. The beta maps (one per model) obtained after iterating the analysis in a searchlight procedure ensured that the variance linked to our RSA models was not misattributed due to differences in instruction length or speed of responses. Importantly, the results obtained this way were very similar to the ones extracted with the standard approach, identifying the same clusters than before.

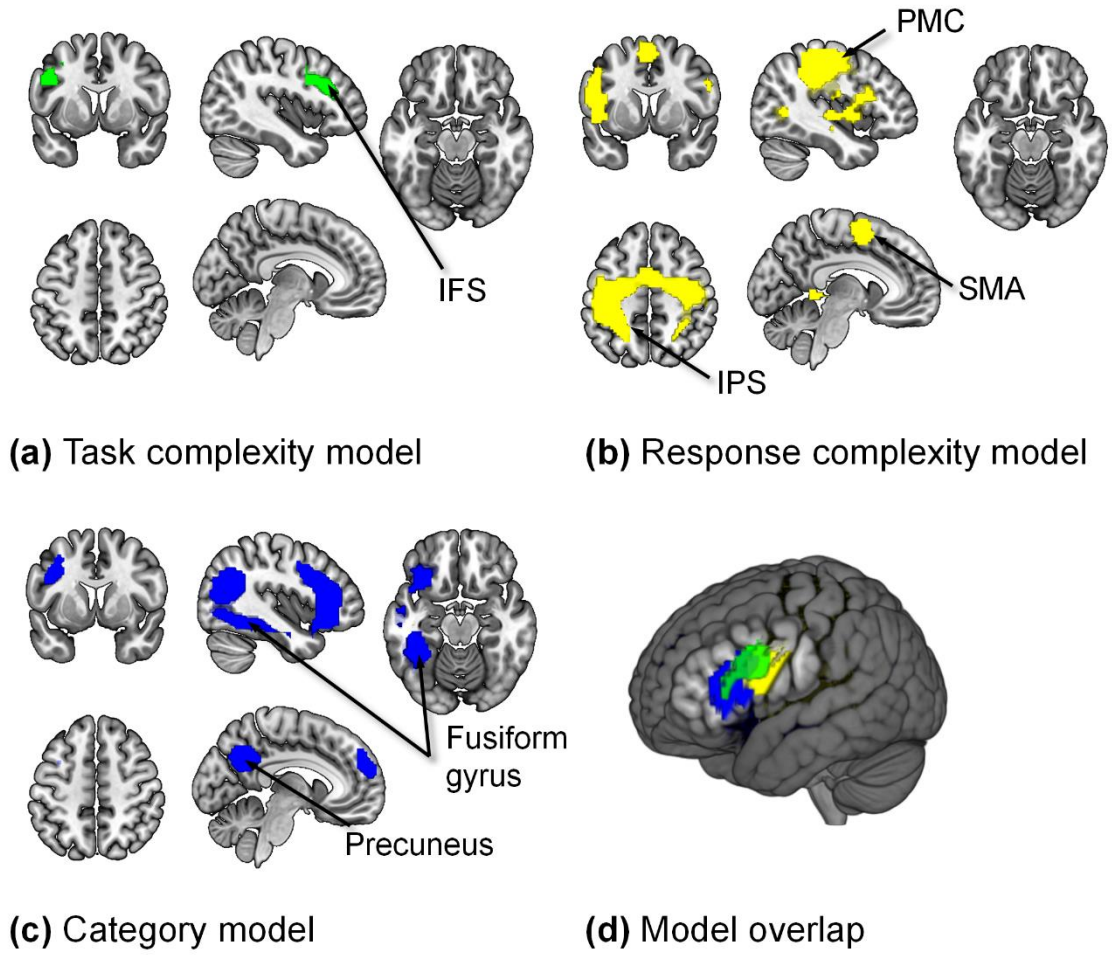

Figure 5.5: Model-based RSA searchlight results for the three models (a-c) and render image showing the overlap among them (d). Note: Identical sections were employed to display the results across models.

We also conducted a **conjunction analysis** to assess the overlap among regions common to the three organizational schemes. Only the left IFG and IFJ resulted significant in this test (Fig. 5.6).

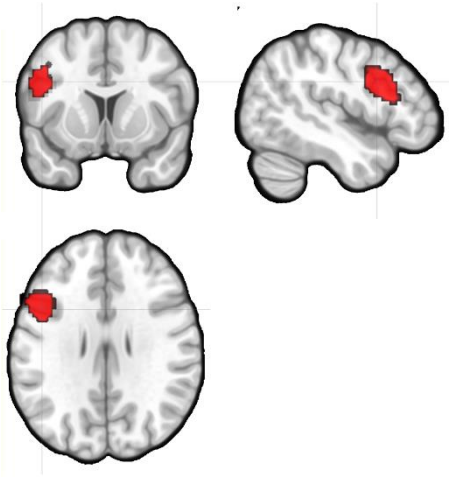

Figure 5.6: Conjunction analysis results.

LOSO-based ROI analysis: assessing confluence of models within regions.

The previous analyses left unexplained the extent to which each of the brain areas isolated by RDM analyses reflected in their organization the three manipulated variables. Furthermore, the conservative correction for multiple comparisons used in the searchlight could overshadow this effect elsewhere in the brain. To shed some light upon this issue, we employed a more sensitive ROI analysis, together with a LOSO approach to avoid double dipping when selecting regions.

All the clusters identified in the main group results (Fig. 5.5) were consistently found across all participants with the LOSO approach, with the exception of the medial superior frontal gyrus under the category model, which was absent in four subjects and thus not included in the analysis. The correlations of the ROIs' RDMs and the three models' matrices were analyzed with a repeated measures ANOVA, in which we found a significant interaction of ROI and Model ($F_{12, 348}$ = 6.050, $P < .001$, $\eta_p^2 = .173$), evidencing variability in instruction coding structure across regions. We then ran one sample *t*-tests or Wilcoxon signed-rank tests (depending on data distribution) to assess model performance in each ROI (see Table 5.1). The general pattern obtained replicated the searchlight results: the

model which originally identified each specific ROI in the searchlight was the one explaining most robustly its encoding activity. Further, in almost all the regions, we did not find enough evidence supporting the effect of the remaining variables. Converging with the previous analyses, the left IFG identified with the dimension integration model was also significantly correlated with response set complexity and category. Similarly, the left IFG cluster found in the category RSA was correlated with the dimension integration model too. In addition, this confluence of models analysis revealed that the response set model was also significant in the category-related cluster involving the left fusiform and precuneus (see Table 5.1).

Table 5.1. Effect of the three models on the LOSO-estimated ROIs.

Original model	**ROI**	**Model tested**	**T value**	**Z value**	***P* value**
Dimension integration	Left IFG	Dim.	3.354		.008
		Resp.	3.292		.009
		Cat.	3.635		.004
Response set complexity	Left IPS	Dim.	0.614		1
		Resp.	5.351		< .001
		Cat.		1.975	.163
	Motor cortices, left LPFC	Dim.	2.478		.067
		Resp.	3.647		.004
		Cat.	1.166		.886
Target category	Left fusiform gyrus and precuneus	Dim.	0.476		1
		Resp.	3.463		.006
		Cat.	5.466		< .001
	Left IFG	Dim.	2.832		.029
		Resp.		0.699	.242
		Cat.	4.930		< .001
	Right MTG	Dim.		-0.144	.557
		Resp.		-1.008	.843
		Cat.		2.859	.002
	Right IFG	Dim.		1.275	.101
		Resp.		-0.206	.582
		Cat.		3.085	.001

Note: P values displayed are Bonferroni-corrected for multiple comparisons. Abbreviations stand for inferior frontal gyrus (IFG), intraparietal sulcus (IPS), and middle temporal gyrus (MTG), Dimension integration model (Dim.), Response complexity model (Resp.) and Target Category (Cat.).

ROI analysis spanning Multiple Demand Network regions.

Following a similar strategy as above, we also examined task encoding organization across the regions comprising the MD network. We extracted each MD region's RDM and correlated it with our three models' RDM, and then entered the correlation coefficients into a repeated measures ANOVA. Again, a significant ROI*Model interaction was found ($F_{20, 620} = 2.168$, $P = .002$, $\eta_p^2 = .065$). To assess which models significantly structured activations across MD ROIs, we conducted one-sample *t*-tests or Wilcoxon signed-rank tests when data were not normally distributed (see Table 5.2).

Only a subset of MD network regions encoded instructions consistently according to any of the proactive control variables, and all of them were located on the left hemisphere and in the LPFC and parietal cortex. The findings were, however, consistent with the searchlight and ROI-related results presented so far. The three variables exerted an effect on different left lateral prefrontal sections: dimension integration and response complexity on the IFG; dimension integration and target category on the more dorsal MFG; and finally, category on the RLPFC. Response complexity was the attribute which most robustly captured representational organization in the IPS.

Table 5.2. Effect of the three models on the MD network ROIs.

ROI	Model	T val	Z val	*P value*
ACC/preSMA	Dim.		0.645	1
	Resp.		1.673	.115
	Cat.	-0.026		1
Left RLPFC	Dim.		1.019	.571
	Resp.		0.346	.365
	Cat.		2.665	.023
Left IFS	Dim.	3.644		.005
	Resp.	4.423		< .001
	Cat.		2.328	.058
Left MFG	Dim.		2.739	.014
	Resp.		0.870	.754
	Cat.	4.298		.002
Left aIfO	Dim.	0.667		1
	Resp.		1.206	.228
	Cat.		2.197	.060
Left IPS	Dim.	1.617		.638
	Resp.		2.814	.025
	Cat.	2.639		.071
Right RLPFC	Dim.		0.365	1
	Resp.	1.460		.849
	Cat.	0.861		1
Right IFS	Dim.	2.220		.186
	Resp.		1.599	.211
	Cat.		-0.626	1
Right MFG	Dim.	2.311		.152
	Resp.	1.294		1
	Cat.	2.042		.273
Right aIfO	Dim.	0.023		1
	Resp.		1.299	.280
	Cat.	1.352		1
Right IPS	Dim.		1.262	.548
	Resp.		1.842	.330
	Cat.		-0.701	1

Note: P values displayed are Bonferroni-corrected for multiple comparisons. Abbreviations stand for anterior cingulate cortex (ACC), presupplementary motor area (preSMA), rostrolateral prefrontal cortex (RLPFC), inferior frontal sulcus (IFS), middle frontal gyrus (MFG), anterior insula/frontal operculum area (aIfO), intraparietal sulcus (IPS), Dimension integration model (Dim.), Response complexity model (Resp.) and Target Category (Cat.).

Effects of reward on representational geometry.

We then explored the effects of motivation in each of the ROIs encoding different

attributes of the instructions (Fig. 5.5), assessing two possible mechanisms that could underlie the behavioral improvements linked to reward (Fig. 5.2). On the one hand, we tested whether reward made our variables more efficient in sharpening the representational space (Fig. 5.2.d, Hypothesis 1), In other words, and taking as an example the target category variable, we assessed whether reward expectations would increase the distance between representations of instructions referring to different stimulus categories (in extension to the other variables, indicated as *different-condition dissimilarity*), while decreasing the distance among those referring to same target category (*same-condition dissimilarity*). On the other, we tested the alternative possibility that dissimilarities would be, in general, greater in the rewarded trials (Fig 5.2.d, Hypothesis 2), regardless of the variables manipulated (i.e., regardless of the pair of instructions being same or different-condition). This could reflect a mechanism for making rule representations more distinguishable among each other, and also, it would be compatible with the increase in rule decoding accuracy that has been liked to motivation in previous reports (Etzel et al., 2016). With that purpose, we extracted, for each region, the average dissimilarity among pairs of instructions pertaining to the same and different conditions, separately for rewarded and non-rewarded trials. We then used Wilcoxon signed-rank tests (Nili et al., 2014) to check whether the difference between different-condition and same-condition trials was larger in the rewarded than in the non-rewarded condition, and also, whether the mean dissimilarity (collapsing across same and different-condition) was increased by motivation.

In the first case, no reward-related differences were observed for any of the instruction-related variables (all *P*s >.1). It is important to note, however, that

these results (as most of the findings presented in this study) are anchored to the instruction's encoding stage, in which proactive control configuration takes place. To explore the possibility that the hypothesized interaction shaped neural activations during the later implementation phase (more related to reactive control; Braver, 2012; Palenciano, González-García, Arco, & Ruz, 2019), we conducted a further test employing beta images from this epoch. However, and again, the expected effect was not significant for any of the ROIs examined (all *P*s >.1).

When addressing the second hypothesis, surprisingly, we found the opposite pattern: reward systematically decreased the dissimilarity values in all the ROIs evaluated (all *P*s < .05, see Table 5.3). To test the behavioral relevance of this finding we correlated, across our participants, the average decrease in dissimilarities associated with reward, with the benefit of motivation on performance (IES; Townsend & Ashby, 1978). We found that in fact, the decrease in representational distances due to reward was significantly correlated with the motivation-related improvements in behavioral performance. Furthermore, this seemed to be a quite robust effect, being present in all of the ROIs included in the analysis (see Table 5.3 for further details).

Table 5.3. Effect of reward on dissimilarity values and correlation with behavioral improvement.

ROI	Effect of reward on dissimilarity values	Correlation RSA - behavior
Task set complexity		
Left IFG/IFJ	Z = -3.005*	r = 0.515*
Response set complexity		
M1 / PM / SMA / IPS	Z = -3.712*	r = 0.565*
Left PHC	Z = -3.712*	r = 0.558*
Target category		
Left fusiform gyrus/ precuneus / IFG/IFJ	Z = -3.712*	r = 0.543*
Right MTG/TPJ	Z = -4.419*	r = 0.495*
Right IFG	Z = -3.712*	r = 0.533*
Medial SFG	Z = -2.652*	r = 0.482*

Note: The asterisks indicate significance at P < .05 on the Wilcoxon paired-sample signed-rank test (middle column) or Pearson correlation coefficient (left column). In the last case, multiple comparisons were controlled with an FWE criterion. Abbreviations stand for inferior frontal gyrus (IFG), inferior frontal junction (IFJ), primary motor cortex (M1), premotor cortex (PM) supplementary motor area (SMA), parahippocampal cortex (PHC), middle temporal gyrus (MTG), temporoparietal junction (TPJ) and superior frontal gyrus (SFG).

MVPA results

We finally aimed to explore the effect of reward directly on decoding accuracies, employing MVPA (Haxby, Connolly, & Guntupalli, 2014), as it has been previously reported during rule encoding in a classic, repetitive task-switching setting (Etzel et al., 2016). We discriminated between the two conditions of each instruction-related variable (i.e., one among faces and food-related trials, other for single versus sequential response requirements, and a last one for within versus across-dimension integration instructions) separately for rewarded and non-rewarded trials. We trained and tested our classifiers across the whole brain using searchlight and obtained, as a result, an accuracy map for each motivation

condition and variable. Nonetheless, while classification was above chance in different brain regions for the three variables, we did not detect any differences in accuracies between rewarded and non-rewarded trials, as no cluster survived at the group-level the *t*-test assessing above zero differences between the two motivation conditions.

5.4. Discussion

In the present study, we aimed to characterize the representational space for novel instructions during their proactive preparation. We assessed whether variables linked to proactive control organized encoding activity patterns and whether this structure was affected by reward expectations. Our results portrayed a complex landscape, where different organizational principles governed instruction encoding in FP cortices and lower-level perceptual and motor areas.

The left IFG/IFJ reflected the most complex and overarching representational structure, with activity patterns structured by dimension integration, response complexity and target category. Robust evidence supports the role of the IFJ in task-set reconfiguration (Brass, Derrfuss, Forstmann, & Cramon, 2005) in practiced (e.g. Woolgar, Hampshire, Thompson, & Duncan, 2011) and novel contexts (e.g. González-García et al., 2016; Muhle-Karbe et al., 2017), orchestrating neural dynamics during attentional selection (e.g. Baldauf & Desimone, 2014). This region seems to be involved in task-set maintenance (Sakai, 2008), selecting task-relevant information represented in perceptual regions (Cole, Reynolds, et al., 2013; Miller & Cohen, 2001). The current study advances our knowledge about the structure underlying *how* information is coded during novel instruction encoding, and stresses the diversity of task

parameters that orchestrate task encoding in the IFG/IFJ. Such a complex, multidimensional representational space (Rigotti et al., 2013) could be key to support the richness and flexibility of human behavior in novel environments. This perspective qualifies recent research, based on MVPA, that highlights the compositionality characterizing representations held in the IFG (Cole, Laurent, et al., 2013; Deraeve, Vassena, & Alexander, 2019; Reverberi, Görgen, & Haynes, 2012), by which complex tasks are coded by combining their simpler constituent elements.

The IPS also encoded novel rules proactively, but now according to response complexity. While this is quite consistent with previous studies linking the parietal cortex to action preparation, it is worth noticing the distinction found in our data between parietal and prefrontal regions, a finding further confirmed with a more sensitive ROI analysis. Dimension integration, the variable manipulated to appeal to a higher-level task goal representation, had an effect only on LPFC, while the IPS was linked to the more specific response-set complexity (De Baene & Brass, 2014; Rubinstein et al., 2001). The frequent coativation of IFG/IFJ and IPS in demanding paradigms (Duncan, 2010) had complicated the identification of their separate contributions. The differential pattern we observed is highly relevant to disentangle their proactive role. Interestingly, the emerging picture portraits the IFG/IFJ and the IPS collaborating during novel task representation, with the former maintaining overarching representations of all relevant variables, and the latter activating the relevant stimulus-response contingencies (see also Muhle-Karbe et al., 2014). The use of RSA in our paradigm provides a deeper understanding of this process, emphasizing that the proposed two-stage preparatory mechanism also guides

task-set encoding in FP cortices. In this sense, variables key for abstract goal or specific S-R settings become relevant differentially depending on the region.

Additional medial and lateral frontal cortices also participate in the FP network and are frequently recruited during task preparation (Duncan, 2010). Consequently, we also examined instruction coding in these MD regions. Our findings highlighted other LPFC areas reflecting target category (both the RLPFC and MFG) and dimension integration (MFG). The overall pattern of results obtained both with whole-brain and with ROI approaches reflects high heterogeneity within the FP network in general, and in the LPFC in particular, in terms of the attributes structuring task-set representation. In contrast, we did not obtain evidence supporting proactive task-set encoding in the ACC/preSMA and the aIfO regions. This finding fits with the subdivision of the FP network into two differentiated components: one anchored in the LPFC and IPS, and a second one composed by the ACC and the aIfO (Dosenbach et al., 2007; Palenciano et al., 2019). In line with our results, anticipatory task coding has been predominantly found in regions from the former rather than in the latter (Crittenden, Mitchell, & Duncan, 2016). Ultimately, the variability found within the FP control network during proactive novel task setting (Palenciano et al., 2019), with different processes and representational formats being combined, could be key to maximize flexibility.

Fronto-parietal cortices were not the sole brain regions encoding novel instruction parameters. Activity in fusiform gyri was organized according to target category, whereas patterns in somatomotor cortices reflected response complexity. While these regions are not associated *per se* with proactive control, their involvement reflects that their representational geometry is tuned in an

anticipatory fashion by relevant task parameters conveyed by instructions. It is important to stress that all the results discussed were locked to instruction encoding, where no target stimuli had been presented, neither any specific motor response could have been prepared. These findings suggest that FP areas exert a bias in posterior cortices, according to the content of instructions. Supporting this, increments of mean activity (Esterman & Yantis, 2010) and target-specific information encoding (e.g. Stokes, Thompson, Nobre, & Duncan, 2009) have been reported in perceptual and motor regions during preparation. Importantly, these changes have been linked to boosts in functional connectivity between the FP and posterior cortices (González-García et al., 2016; Sakai & Passingham, 2006). In direct relation to our findings, a recent study showed that the representational organization in regions along the visual pathway is dynamically adapted to task demands (Nastase et al., 2017). Our current results add to these findings by showing that representational space tuning could be a mechanism of preparatory bias, which could reflect predictive coding principles where iterative loops of feedback and feedforward communication shape cognition (Friston, 2005).

Crucially, the structure of information encoded by all these regions was sensitive to trial-wise motivational states. Surprisingly, reward expectation diminished the dissimilarities between the representations of the instructions although preserving the organizational scheme found in each area. Based on recent findings of increased task decodability (Etzel et al., 2016), we had hypothesized that reward would either polarize the representational structure or increase the representational distances overall. Results were, however, in the opposite direction, even when our reward manipulation was successful at boosting

performance and also increased activity in control and reward-related regions (Parro et al., 2017). Most importantly, decreases in dissimilarities were also robustly correlated with behavioral improvements. Taking into account that additional analysis employing MVPA and using data from the implementation stage corroborated these results, their implication must be thoughtfully considered. One possibility is that the decrease in dissimilarities is generated by a general boost of reward in signal-to-noise ratio. Although our results persisted after normalizing data across trials, a reward-related reduction of multivariate noise pattern could still be possible, and it could benefit task coding in the absence of the hypothesized RSA results. However, the MVPA did not reveal improved task classification accuracy in the rewarded condition, and thus this interpretation remains uncertain. Alternatively, motivation could have influenced task coding in ways that our searchlight procedure was not sensitive to. That would be the case if reward affected the spatial distribution of information: as ROIs were defined by size-fixed searchlight spheres, and were equal in rewarded and non-rewarded conditions, an effect like that would remain shadowed. Finally, the task complexity could also be key. In less demanding situations such as repetitive task switching (Etzel et al., 2016), reward could directly sharpen task encoding representations. In novel environments, however, motivation could exert a more general effect at the process level -instead of at the representational one. It could increase the efficiency of task reconfiguration (Braem & Egner, 2018), as indexed by the improvements in behavior, while the specific rule representations would remain equally structured. Nonetheless, more research is needed to properly characterize the intricate interactions among proactive control and motivation (Pessoa, 2017) in rich task environments, more

akin to daily life situations.

The current study entails some limitations that constrain the scope of our findings and call for further research. On the one hand, the nature of our paradigm demanded the selection of a few instruction-organizing variables. Some other dimensions, critical for anticipatory encoding, may have been left unaddressed. Furthermore, non-linear combinations of variables could add to the organization principles governing control regions (Rigotti et al., 2013). Considering an increasing number of plausible models in more complex and/or naturalistic scenarios, together with data-driven methods such as multidimensional scaling or component analysis, will complement our results. On the other hand, our main dependent variable (fMRI hemodynamic signal) provided spatially precise, but temporal impoverished data. Temporally resolved techniques, such as electroencephalography or magnetoencephalography, could be key to unveil the temporal dynamics of the representational patterns.

Overall, our findings provide novel insights on how verbal complex novel instructions organize proactive brain activations. The emerging picture departs from pure localizationist approaches where brain regions carry fixed information about concrete cognitive processes. Rather, the different dimensions relevant for efficient instructed action shape brain activity across an extended set of areas, flexibly structuring encoding activity according to the relevant task parameters.

Chapter 6:

Temporal dynamics underlying different structures for novel task anticipatory coding – Study 3

Abstract

Anticipatory task configuration entails the encoding of rule parameters to guide perceptual and motor systems according to current goals. This is a dynamic process, where task representations change along successive differentiated stages (Stokes et al., 2013). Recent evidence, obtained with functional Magnetic Resonance Imaging (fMRI), indicates that novel complex instructions structure anticipatory activations simultaneously according to different proactive control and motivation-relevant variables (Palenciano, González-García, Arco, Pessoa, & Ruz, 2019). Nonetheless, the temporal unfolding of such variable and multidimensional preparatory codification remains unknown. To investigate it, in the current study we collected high-density electroencephalography data while participants followed the same set of novel verbal instructions, and employed state of the art pattern analyses. Our results did not provide strong evidence supporting a clear effect of control-related variables at organizing task sets. Motivation, however, exerted a clear impact on anticipatory representations, making rules more similar among each other when a monetary reward was expected. This replicated and further characterized previous fMRI findings (Palenciano et al., 2019), showing that reward has an impact on two separated temporal windows. Our results open debate about how to conceptualize motivation-control interplays.

6.1. Introduction

In our daily life, we perform multiple and diverse tasks, quickly shifting among them when contextual demands and internal preferences require (Braver, 2012b; Monsell, 2003). This flexibility is crucial in novel contexts, where fixed, habit-like behavior does not lead to success. In such circumstances, humans can make an efficient use of instructions to guide our actions – an ability key for our adaptation to changing environments (Cole, Laurent, et al., 2013). Functional magnetic resonance imaging (fMRI) studies have highlighted the role of lateral prefrontal and parietal cortices during novel task processing (Bourguignon et al., 2018; González-García et al., 2017; Hartstra et al., 2012), and recent results indicate that complex instructions containing several relevant parameters at once (Palenciano et al., 2019) orchestrate anticipatory representations in different brain areas. In the temporal domain, understanding how flexible novel behavior is implemented in dynamic brain activity patterns is key for cognitive neuroscience.

To guide our behavior, instructions must be transformed into action-based representations. These are known as tasks-sets (Sakai, 2008) and contain the relevant rules, target stimuli, and responses (Sakai, 2008). Crucially, part of the task-set reconfiguration is proactive, starting and developing before actual stimuli presentation. The efficiency of this anticipatory task preparation influences the posterior behavioral performance (Rogers & Monsell, 1995). Success on new and complex demands relies even more on this proactive reconfiguration due to the most costly novel task-set assembly (Cole et al., 2018). Despite its relevance, it is yet unknown how preparatory activity for such complex

novel action plans structures the different relevant pieces of information in the temporal domain.

Some insights about the dynamic nature of this proactive reconfiguration can be extracted from electrophysiological recordings in non-human primates. In these studies, when a task is coded in advance, the resulting representation does not remain static until execution. Contrary, it displays a complex temporal evolution across different stages during the preparation interval (King & Dehaene, 2014; Sigala, Kusunoki, Nimmo-Smith, Gaffan, & Duncan, 2008; Stokes et al., 2013). These anticipatory variable task representations seem to instantiate an encoding context for upcoming stimuli, ensuring that they are processed in a goal-related fashion (Stokes et al., 2013). Similar complex dynamics have also been found in humans (Hebart, Bankson, Harel, Baker, & Cichy, 2018). Nevertheless, all the evidence available comes from simple experimental settings employing few and highly practiced rules. No attempts have been made for addressing this dynamic reconfiguration in more complex and novel task scenarios.

On this line, a recent study showed that preparation for novel instructions is linked to a flexible, task-oriented organization of anticipatory brain activity (Palenciano et al., 2019). Dimension integration requirements, response set complexity and target category (all of them related to proactive control) structure novel task representations across multiple brain regions. As these results were obtained with fMRI, the temporal profile characterizing these effects remained completely obscured. However, there are theoretical reasons to hypothesize the existence of different time courses underlying the influence of each variable on the encoding structure. In this sense, it has been proposed that preparation entails two differentiated stages: a first global task goal setting, followed by a

more concrete stimulus-response reconfiguration (De Baene & Brass, 2014; Rubinstein et al., 2001). This leads to the possibility that initial stages of novel task setting could be organized by more abstract, goal-related attributes (as dimension requirements and target category), followed by a response-oriented representational organization. Such a pattern could potentially reflect how instruction preparation entails the transit from a more abstract format to an increasingly proceduralized representation (Brass et al., 2017). To shed some light upon this issue, we adapted the study of Palenciano and colleagues (2019) to measure high-density electroencephalography (EEG) data. Extending advanced multivariate analysis (Haxby et al., 2014; Kriegeskorte et al., 2008) to the temporal plane (Grootswagers, Wardle, & Carlson, 2017) allowed us to track the temporal dynamics underlying the distinct proactive representational spaces.

Furthermore, continuing with our previous research, here we also assessed the impact of the motivational state upon task preparation, by including economic incentives in our paradigm. Reward expectations, traditionally associated with behavioral improvements, have been linked to a boost in proactive control mechanisms (Chiew & Braver, 2016; Shen & Chun, 2011). In line with this, Palenciano et al. (2019) found a robust motivation effect on preparatory task encoding. Intriguingly, reward expectations robustly decreased the dissimilarity among instructions representations, and this influence was tightly linked to benefits on performance. Previous results in practiced and simpler task scenarios pointed toward an opposite effect, with reward enhancing rule discriminability on neural patterns (Cole et al., 2011; Hall-McMaster et al., 2019). Consequently, we aimed to replicate our puzzling motivation-related finding, as well as to

characterize its temporal profile, which could help to better understand its implications.

6.2. Methods

Participants

We recruited 36 students (22 women; average age = 22.28 years, SD = 3.2 years), all of them right-handed, native Spanish speakers, and with normal or corrected-to-normal vision. They all signed a consent form approved by the Ethics Committee of the University of Granada and received monetary compensation in exchange for their participation (10-20€, according to their performance). The sample size was computed with a power analysis focused on the detection of a two-way behavioral interaction term (assuming a small-medium effect size of Cohen's d = .3 and 80% power). Five participants were excluded from the final sample, three due to poor performance (<70% accuracy rate) and the remaining two due to excessive artifacts in the EEG signal. Thus, data from 31 participants were submitted to the analysis.

Stimuli and procedure

The main experimental paradigm replicated the one employed in Palenciano et al. (2019). The instructions, stimuli and timings parameters remained the same, and all the details are available in Chapter 5. Crucially, as in the previous fMRI study, the independent variables manipulated here were integration across or within stimuli dimension, response set complexity, target category and reward expectation.

In addition, to aid in the detection and removal of ocular artifacts in the EEG signal, participants also completed a reading task. This was built employing random sentences from Wikipedia articles, sharing either the same number of words or letters than the instructions employed in the main task. We also ensured that only those with descriptive content were selected, avoiding sentences containing any kind of procedural information. A total of 60 sentences were presented in random order during 2500ms each, separated by a 3000ms ISI.

When participants arrived at the laboratory, and once the EEG equipment was set, they completed a practice session. This was identical to the experimental task, with the exception of the specific instructions employed, which were drawn from a different set to maintain rule novelty. Participants had to reach 80% of accuracy to continue with the main phase of the study. For timing reasons, we established a maximum of four repetitions of the practice block to allow continuance (with a maximum duration of 36 minutes). We did not record EEG data during this phase.

Afterward, participants performed the reading task, which was introduced with the purpose of maximizing the amount of data containing clear saccade artifacts, but in the absence of the neural signal of interest (i.e., activity linked to instruction encoding and preparation). During this task, participants were instructed to read as naturally as possible, but avoiding blinks and sudden saccades outside of the sentence location. During the ISI, they were told to stare at the central fixation point. The full reading task had a duration of five minutes and a half.

Finally, participants completed the main experimental task, composed of 192 trials distributed in six independent blocks with self-paced pauses between them. This phase lasted around an hour.

EEG recordings and preprocessing

High-density EEG was recorded from a 65-electrodes system (actiCap slim, Brain Products) with a 1000Hz sampling rate and an online high-pass cutoff of 0.016Hz. Two electrodes, TP9 and TP10, were used as horizontal electrooculogram (EOG), placed lateral to both eyes, to record saccades. Impedances were kept below 5kΩ, and data were referenced online to the FCz electrode. All data preprocessing was performed using EEGLAB software running on Matlab (r2016a version). First, we visually inspected the raw data to detect major artifacts and bad channels. However, no channel interpolation was required for any participant. Then, the recordings were downsampled to 256 Hz, average re-referenced, and filtered using a low-pass FIR of 127Hz. The main analyses were done on epochs anchored at the onset of the instruction events, with 3500ms duration [-500, 3000ms], and baseline-corrected using the [-500, 0ms] interval, corresponding to the previous reward cue period. A control test, carried out to assess data quality (see *MVPA* section), was performed on wider epochs. As this analysis focused on the trials' motivation condition, the epochs included the time window where reward information was available (anchored at reward cues, lasting for 5500ms [-500, 5000ms], baseline-corrected employing the [-500, 0ms] preceding ISI period).

An identical preprocessing was done with data from the reading task. We extracted epochs anchored at sentence onset, of 3500ms duration [-500,

3000ms], with a baseline correction on the [-500, 0ms] period of the previous ISI. Then, for each participant, we concatenated epochs from the reading and the main experimental task and used them to perform ICA (*runica* algorithm in EEGLAB), after excluding the ocular channels. We used visual inspection of components' topographies and power spectrum, together with SASICA software (correlation of components time course with EOG channels) to identify blinks and lateral movement ICAs (Chaumon, Bishop, & Busch, 2015). This step was critical, as the nature of the main instruction-following task necessarily involved saccadic movements as participants read the instructions. We additionally identified components containing clear muscular or channel-specific noise (Chaumon et al., 2015). Artifact-related ICAs were removed from the data. Finally, after discarding the reading task epochs, we conducted an automatic trial rejection procedure, in which epochs displaying abnormal spectra (>d50dB in 0-2 Hz frequency window; <100dB or >25dB in 20-40Hz), extreme voltage values (±200V) or improbable data (voltages departing ±6SDs from the baseline) were eliminated.

Behavioral data analysis

Accuracies and reaction times were entered, separately, in 4-ways repeated measures ANOVA, including Integration demands (within vs. across dimension), Response set complexity (simple vs. sequential responses), Target category (faces vs. food) and Reward (rewarded vs. non-rewarded trials), as factors. Post-hoc comparisons were Bonferroni-corrected for multiple comparisons. Behavioral analyses were performed with the software JASP (JASP Team, 2019; https://jasp-stats.org/).

EEG data analysis

Our main goal was to address the temporal unfolding of novel task representational structure, tracking the effect of proactive control-related variables. We also assessed their interaction with motivation. The analyses followed those in Palenciano et al. (2019) as much as possible, with the goal of maximizing the comparability of EEG and fMRI results. A detailed description of the analysis procedures (including figures) can be found in Palenciano et al (2019), included in Chapter 5. Here we include mainly the information that deviates from these previous analyses, mostly due to the different nature of the neuroimaging method. We employed two multivariate techniques: Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008) and Multivariate Pattern Analysis (MVPA; Haxby, Connolly, & Guntupalli, 2014). Importantly, both techniques were applied in a time-resolved fashion, iterating at each time point of the encoding interval, and yielding time courses of the measure of interest as a result.

Taking into account the particularities of these analyses, we applied a series of preparatory steps on the data to avoid inflated or biased results. First, we avoided imbalanced data sets, which could affect the interpretability of MVPA results (see below). We ensured that an equal number of trials of each condition were included in the MVPA (e.g., same number of faces and food-related trials in all Target category analyses). Furthermore, as motivation was manipulated concurrently with the proactive control-related variables, we also controlled that the same proportion of rewarded and non-rewarded trials were included for each condition. This way, effects linked to the motivational state could not be

misattributed to any of the other factors. In both cases, this was done by downsampling the data from the majority class.

Once we had a balanced set of trials to be included in the analyses, we extracted the corresponding epochs (which, unless otherwise stated, were anchored at the presentation of instructions) and smoothed the signal with a sliding window encompassing 10-samples (≈40 ms). We then downsampled each trial data to one third (selecting one of every three time points) to reduce computational costs. Finally, data were z-scored (ensuring that M = 0 and SD = 1) for each trial and channel separately. We then used these data to build the trial-wise multivariate activity patterns, generated time point by time point, which consisted of the normalized voltage values of all 63 electrodes (excluding EOG channels). Despite the proved benefits of trial averaging (i.e., the employment of supertrials; Grootswagers, Wardle, & Carlson, 2017) on EEG signal decoding to increase the signal-to-noise ratio, the number of observations in our design was not well suited for this approach, which forced the use of trial-by-trial estimators. All analyses were performed employing custom-developed MATLAB code (López-García, Sobrado, Peñalver, Górriz & Ruz, 2019).

Representational Similarity Analysis (RSA)

We first assessed when, during the instruction encoding, Task-set complexity, Response set complexity and Target category explained the instructions' encoding organization. RSA allowed us to compare, time point by time point, theoretical models based on these three variables with the estimated neural representational space.

To do so, we first generated the theoretical representational dissimilarity matrices (RDMs), in which each cell contained the expected distance (1 = maximal dissimilarity; 0 = minimal dissimilarity) between a pair of instructions, according to each of the three control-related variables manipulated. For example, in the Category RDM, pairs of face-related instructions had a dissimilarity of 0, and pairs of face and food-related instructions, of 1 (see Fig. 5.2.a). Then, at each time point, we built a neural RDM, in which each cell contained the actual distance among a pair of instructions' voltages vector. This was computed as 1 – the Pearson correlation coefficient between the multi-channel activity patterns of both trials. Then, we performed Spearman correlations among the theoretical and the neural RDMs, to assess the extent to which the former explained the latter. The resulting three correlation values were normalized (with the Fisher transformation) and stored, iterating this procedure through the whole epoch. This way, for each participant we obtained three correlation time courses, one per variable.

To assess for significant time windows where the variables effectively organized the patterns, we followed permutation-based testing (Stelzer, Chen, & Turner, 2013). This was suitable for the non-parametrical nature of the measure at hand, and additionally corrected for the multiple comparisons performed. For each participant and control-related variable, we permuted 100 times the trials' labels, used them to build empirical RDMs and repeated the time-resolved RSA. Then, we generated 10000 null group-level results, selecting one random correlation time course per participant and averaging across them. This allowed us to build an empirical null distribution of correlation values per each time point, from which we took the values corresponding to the 95 percentile as the threshold for

significant positive correlations ($P < .05$) with the theoretical RDM. We further controlled for inflated false positive rates associated with multiple testing with a clustering approach. For this purpose, we generated an empirical null distribution of cluster sizes (understood as contiguous, significant time points) from the permuted and thresholded ($p < .05$) correlation time courses, and assigned each cluster size an FDR-corrected P value. The actual results obtained were averaged across the sample, tested against the correlation thresholds, and then assigned a P value according to the clustering present in the data.

We then analyzed how reward expectation interacted with proactive control-related variables during task encoding. We tested the same two hypotheses explored in Palenciano et al. (2019), namely, whether reward polarized representational organization (Hypothesis 1), or alternatively, whether it increased general task dissimilarities (Hypothesis 2). To do so, we distinguished among same and different-condition trials (for example, two faces-related trials and a face and a food-related trial, respectively). For each time point, we built two neural RDMs, one per motivation condition. Within each RDM, we averaged the cells corresponding to same and to different-conditions trials. Hypothesis 1 was assessed by comparing if the subtraction of different minus same-condition trials was higher in the rewarded than the non-rewarded condition. To test Hypothesis 2, we just checked whether mean dissimilarity (collapsing between same and different-condition trials) was greater when reward was expected than when it was not. Again, we used permutation for statistical inference in this analysis. We followed the same procedure as above, but building empirical null distributions of the two differences among reward conditions assessing Hypotheses 1 and 2.

Multivariate Pattern Analysis (MVPA)

We further performed MVPA to investigate whether the information about control-related variables could be decoded from EEG signals. This was motivated by the different nature of this and the previous RSA. While RSA informs about the higher-level representational structure, with MVPA we access informational content. We conducted three MVPAs, aiming to classify between trials: (1) requiring within vs. across dimension integration; (2) with single vs. sequential responses; and (3) face vs. food-related.

To do so, we used the same data as in the previous RSA. Again, we repeated the MVPAs at each time point of the instruction-anchored epoch. We trained a binary linear Support Vector Machine classifier (C = 1; from libSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/) in 90% of the two conditions trials, and then tested it against the 10% (unlabeled) remaining ones. The percentage of trials correctly classified during the testing is informative regarding how readable the variable information is. To maximize data usage and generalization, we followed a 10-fold cross-validating scheme, which ensured that all trials were used once for testing, iterating across the full data set. At this point, we not only ensured balanced data among conditions, but also within each iteration of the cross-validation scheme (same number of each condition's trials within fold, and the same overall amount of data amount across all folds), which otherwise could compromise the results. The accuracies obtained across the cross-validation were averaged and assigned to the corresponding time point. Statistical inference followed the same approach as the RSA (Stelzer et al., 2013).

We also assessed reward effects with this technique, similarly as it has been done in the past (Etzel et al., 2016). We explored if motivation benefited task information coding (enhancing classification accuracy) at any moment of the

instruction encoding period. To do so, we repeated the three MPVA but separately for rewarded and non-rewarded trials. We then performed paired t-tests to identify statistical differences between the two motivation conditions.

We finally performed one extra analysis as a sanity check, as several aspects of our experimental settings are remarkable novel in the literature. The employment of paradigms involving naturalistic reading is still quite infrequent with EEG recordings, and absent in combination with decoding techniques. Moreover, ocular artifact cleaning techniques are recent and under development (e.g.: Dimigen, Sommer, Hohlfeld, Jacobs, & Kliegl, 2011). To ensure that the data were suitable for our analysis approach, we conducted a control test decoding motivation. At least in fMRI, reward expectation is linked to strong mean activity increases (Palenciano et al., 2019; Parro, Dixon, & Christoff, 2019) and its encoding is widespread in the brain (Wisniewski, Forstmann, & Brass, 2018). Thus, we decided to assess its effect on our data, as this manipulation we expected would be easier to detect than other abstract task attributes such as the ones explored in the main analysis. We classified rewarded from non-rewarded trials following the same procedure as specified before, with the exception that a wider epoch was employed, including the reward cue interval in addition to instruction presentation. We expected to find clear above-chance accuracies from reward cue onwards.

6.3. Results

Behavioral

The findings relating RTs and accuracy were highly congruent with our previous results (Palenciano et al., 2019). Dimension integration displayed a robust

significant effect on both RT (F1, 30 = 63.82, P <.001, η_p^2 = .68) and accuracy (F1, 30 = 73.56, P = .001, η_p^2 = .71), with faster (within dimensions: M = 0.98s, SD = 0.21s; across dimensions: M = 1.06s, SD = 0.23s) and more accurate responses (within dimensions: M = 0.85, SD = 0.12; across dimensions: M = 0.77, SD = 0.13) when the integration required was within stimulus dimension than across them. Target category also resulted significant in both behavioral measures (RT: F1, 30 = 47.28, P <.001, η_p^2 = .61; accuracy: F1, 30 = 17.90, P <.001, η_p^2 = .37), with face-related trials linked to more accurate (faces: M = 0.83, SD = 0.12; food: M = 0.79, SD = 0.14), but slower responses (faces: M = 1.06s, SD = 0.23s; food: M = 0.99s, SD = 0.21s). This category-related pattern has been previously found in studies using the same or similar paradigms (Palenciano et al., 2019; Palenciano, González-García, Arco, & Ruz, 2018).

Regarding the effect of reward, it was significant on RTs (F1, 30 = 4.97, P < .05, η_p^2 = .14), with faster responses when an incentive was expected (rewarded: M = 1.01s, SD = 0.23s; non-rewarded: M = 1.04s, SD = 0.22s). Rewarded trials were linked to a numerically higher accuracy rate (rewarded: M = 0.82, SD = 0.13; non-rewarded: M = 0.80, SD = 0.13), but the difference did not reach statistical significance (F1, 30 = 2.90, P = .09, η_p^2 = .09).

Finally, the interaction of Reward with Response set complexity (F1, 30 = 6.02, P = .02, η_p^2 = .17) and Dimension integration with Target category (F1, 30 = 4.31, P =.05, η_p^2 = .13) were significant on RTs. We thus performed posthoc comparisons to access to the directionality of these effects. In the case of the former, the interaction was driven by a significant decrease in RTs in rewarded versus non-rewarded simple response trials, with less robust differences in the case of

sequential response ones (P > .05). Regarding the second interaction, it was generated by a non-significant difference between face-related trials requiring within-dimension integration and food-related ones integrating across dimensions, while the rest of pair-wise comparisons were statistically different. Nonetheless, both behavioral interactions fell beyond the scope of this work and are not further considered.

EEG results

Dynamics underlying novel task-set encoding structure

The main goal of this study was to track, with high temporal precision, the instauration of three representational structures known to be relevant for novel task representation and conveyed by verbal complex instructions. The three encoding structures were those governed by dimension integration, response set complexity and target category. We extracted this information by generating theoretical RDMs and correlating them, time point by time point, with the neural RDMs actually established during the instruction encoding interval. The time courses of the correlation coefficients for the three variables with the neural RDMs are available in Fig. 6.1.

Requirements for integrating within or across stimuli dimensions structured task representation halfway the encoding interval, with significant correlation clusters (P < .05 FDR-corrected for multiple comparisons) found between 1100 and 1500ms after instruction onset.

Response set complexity captured encoding organization later on, during the posterior jitter interval, reaching statistical significance almost 3000ms after rule presentation. However, and unexpectedly, an additional significant cluster was

found 100ms before instruction onset. The implications of this artifact will be later elaborated in the discussion.

Finally, target category organized task-set encoding during the first half of the encoding, with significant peaks at latencies between 900 and 1300ms. However, and again, another significant cluster appeared 50ms before instruction onset.

Temporal profile of reward-proactive control interactions

In a second step, we further evaluated whether the establishment of organized representational spaces was affected by the motivational state of the participants. Following Palenciano et al. (2019), we assessed two potential mechanisms by which this modulation could take place.

The first hypothesis, regarding a reward-induced polarization of encoding spaces, did not obtain evidence to be supported (see Fig. 6.2.A). The effect of our three control-related variables remained equivalent between the two motivation conditions through the whole encoding interval. We only found two exceptions, involving small clusters in which the test reached significance, but with an impoverishment of representational structure linked to reward expectancy. They appeared in response set complexity RSA (at around 250ms of latency) and in the target category RSA (after 2750ms, already during the ISI).

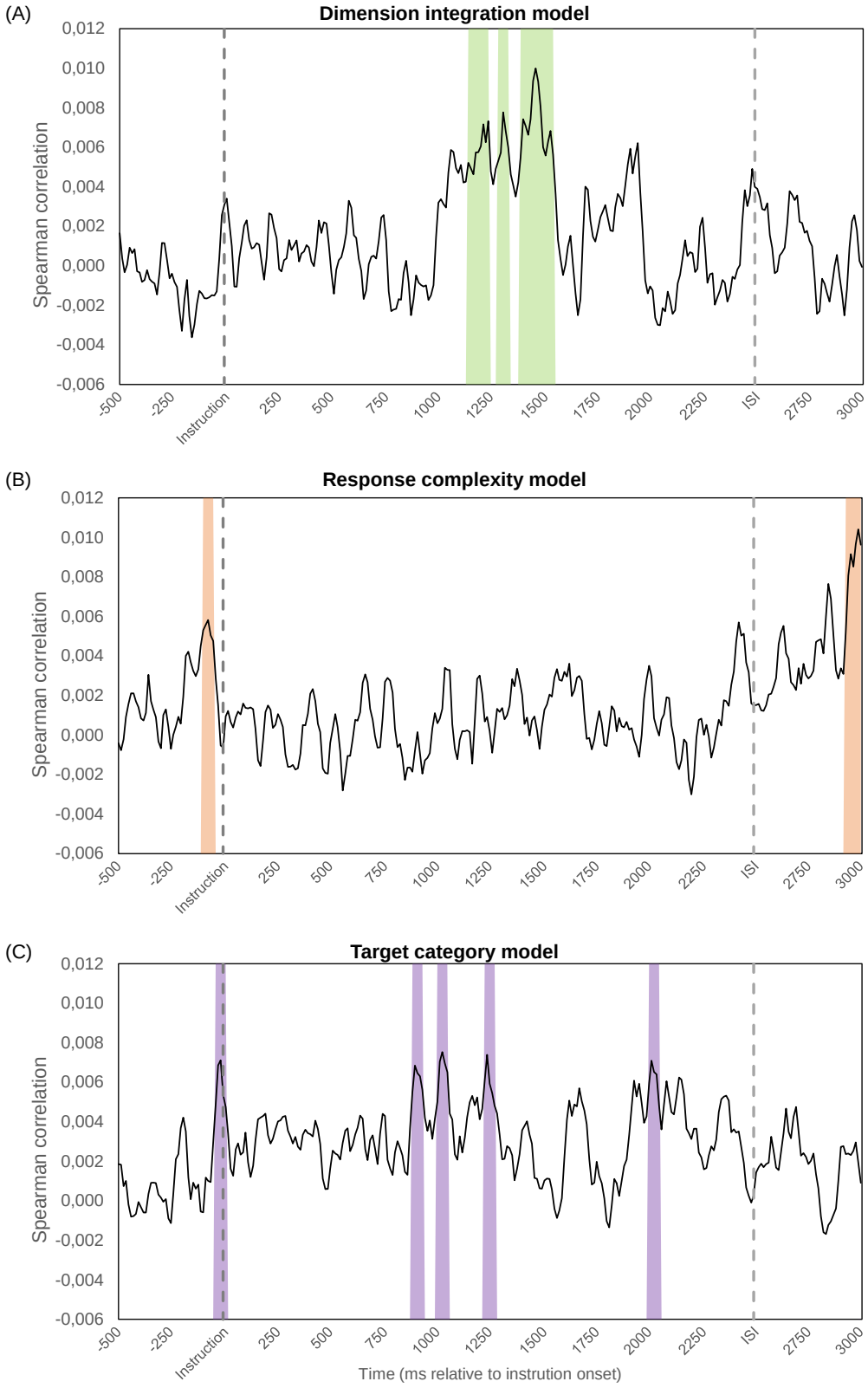

Figure 6.2: Time courses of the Spearman correlations for the dimension integration (A), response set complexity (B) and target category (C) models RSAs with the neural one across the encoding interval. Shaded areas indicate significant positive correlations ($P < .05$, cluster-wise corrected for multiple comparisons).

When then tested the second possibility, we found that motivation exerted a general decrease in dissimilarities among instruction representations (Fig 6.2.B), as in our previous fMRI results (Palenciano et al., 2019). Furthermore, this effect displayed a remarkable clear pattern across the control-related variables examined. In the three analyses, two differentiated significant time windows were found: the first appearing approximately 200ms after instruction onset and a second one during the last 500ms of the encoding interval. In the category RSA, two additional clusters appeared before instruction onset, however, motivation information was already available during this interval through the reward cue.

MVPA decoding of proactive control variables

To complement the RSA results, we introduced our data into three separate MVPAs, aiming to decode between the two levels of each one of our control-related variables. On one hand, the integration within or across dimensions could be significantly decoded in small, distributed clusters along the encoding epoch (see Fig. 6.3.A). Single and sequential response requirements, on the other hand, were predominantly decoded at the end of the encoding window, with significant accuracy clusters appearing around 1600ms after instruction onset, and again at 2800ms latency (Fig. 6.3.B). Finally, target category was decoded mainly from the first half of the epoch (Fig. 6.3.C), with an early significant classification peak at 200ms, and more compacted clusters around between 500 and 1000ms latencies. Additional peaks were found toward the end of the encoding.

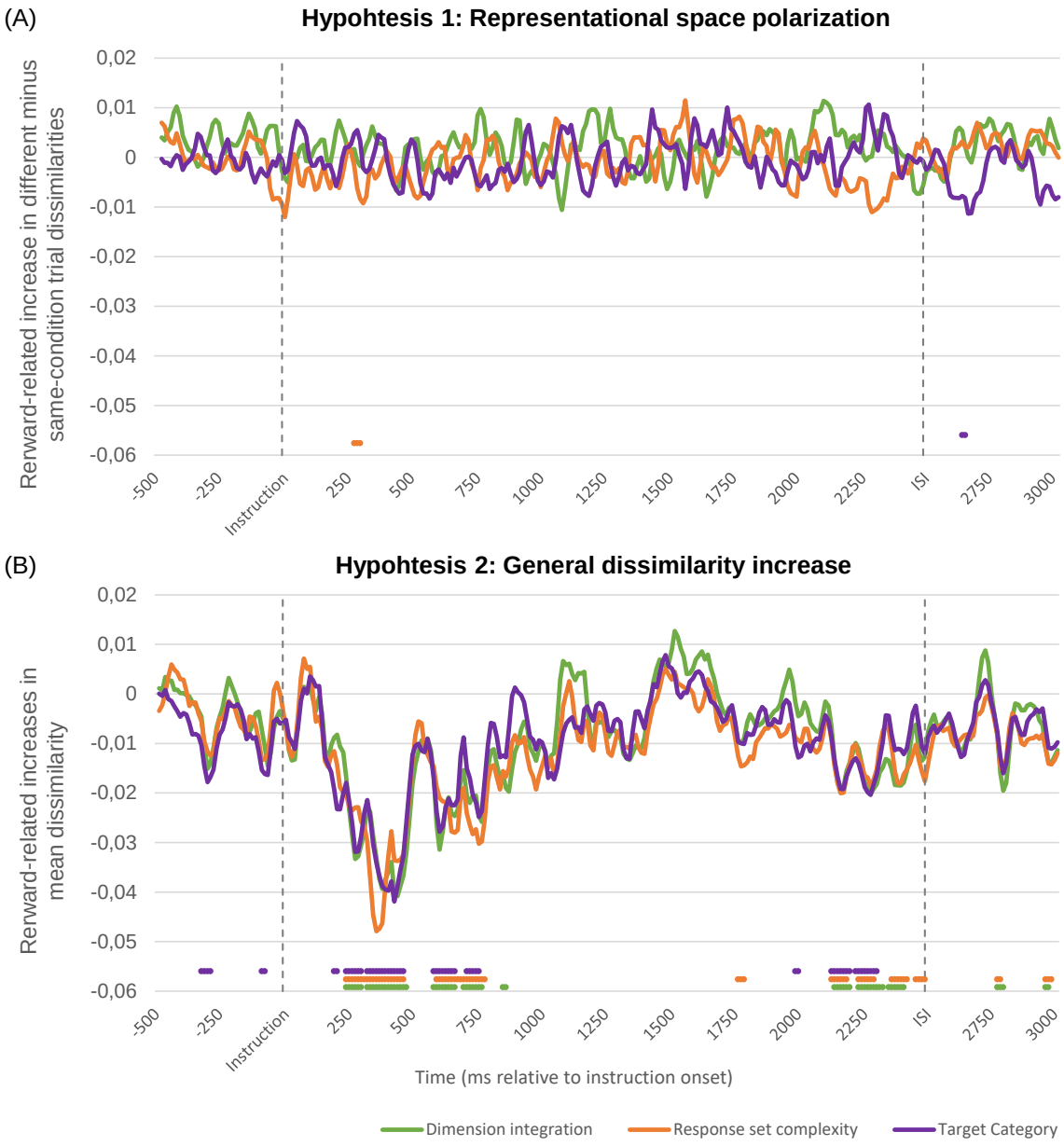

Figure 6.3: Time courses of the difference explored for the two interaction hypotheses. (A) Results from subtracting same-condition trials from different-condition ones, and then comparing among reward conditions. Above 0 values correspond to a greater increase in different (versus same) condition trials when reward is expected, which will indicate a polarization of the representational space. (B) Results from subtracting the mean dissimilarity of non-rewarded trials from rewarded ones. Above 0 values correspond to higher dissimilarity under the rewarded condition. In both (A) and (B), bars underneath the graphs depict significant deviations from 0 ($P < .05$, cluster-wise corrected for multiple comparisons).

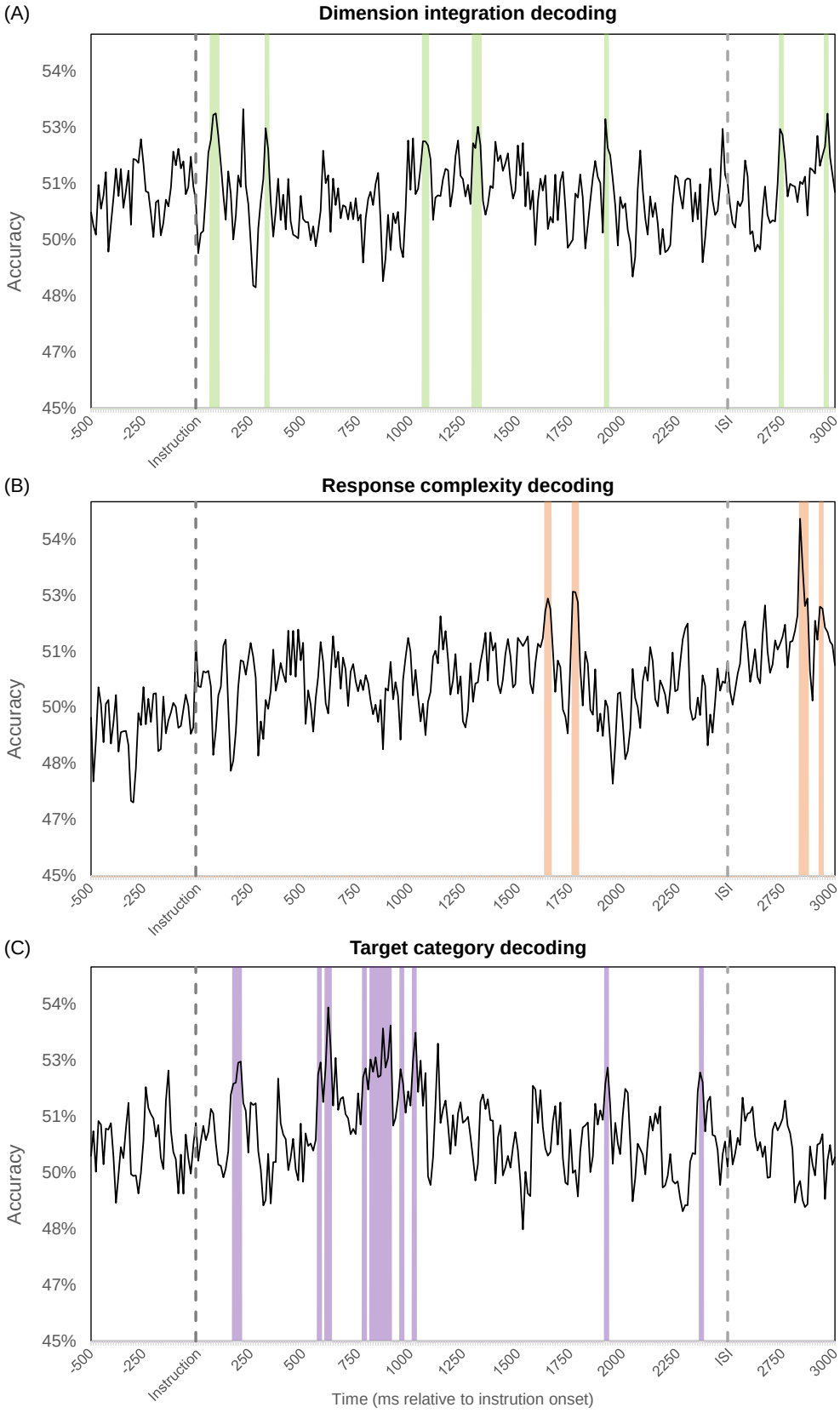

Figure 6.3. Time courses of the classification accuracies for the dimension integration (A), response set complexity (B) and target category (C) decoding analysis. Shaded areas indicate significant above-chance accuracies ($P < .05$, cluster-wise corrected for multiple comparisons).

Motivational state decoding.

We performed a final analysis as a quality control check, classifying among reward expectation conditions. Importantly, reward information was available during two differentiated intervals: during the reward cue, which by its physical properties (small 1.5º images centered on the screen) was a better match with traditional experimental settings, but also during the following instruction interval, in which reading-related saccades took place. Thus, we could disentangle – in case of absence of the expected results – whether the instruction period, even after eye artifact correction, was not well suited for RSA or MVPA.

The results of these MPVAs are displayed in Fig. 6.4. Motivation information was robustly decoded through both reward cue and instruction intervals. Classification accuracy rose after reward cue onset, reaching a peak around 500ms afterwards, and slowly decaying by the end of the period. Once the instruction was presented, motivation information was readable again during two time intervals, one within the first 500ms, and a second one with an approximate latency of 1000ms. Additional peaks were found dispersed by the end of the encoding.

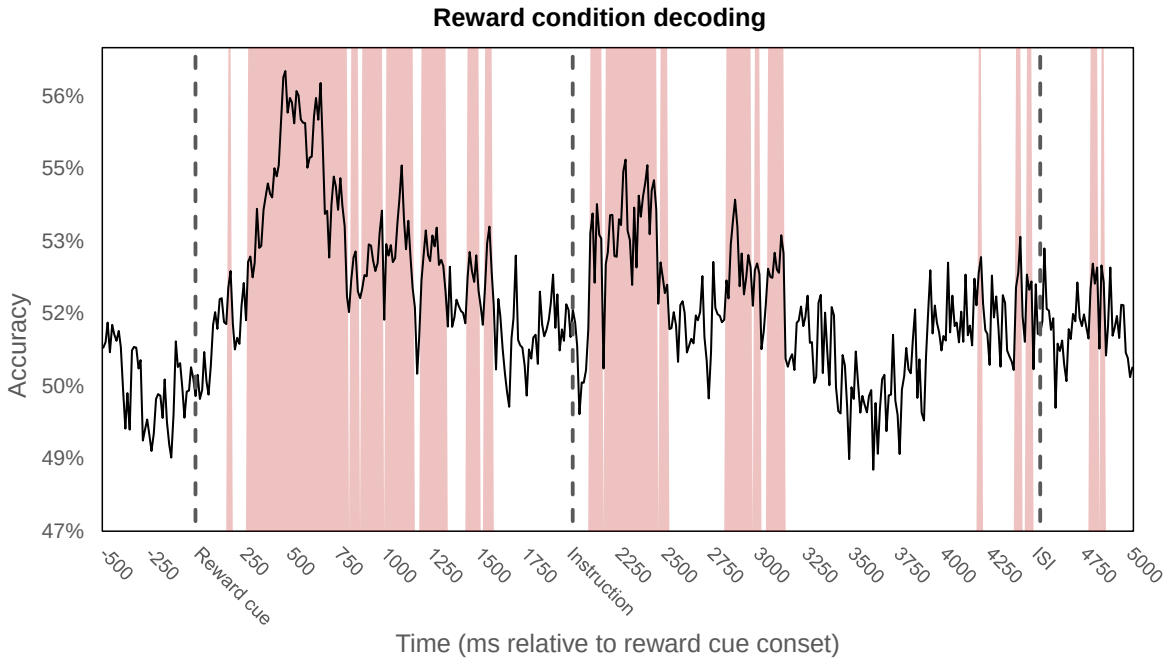

Figure 6.4: Time course of the classification accuracy for the motivation MVPA. Shaded areas indicate significant above-chance accuracies (P < .05, cluster-wise corrected for multiple comparisons).

6.4. Discussion

In this study, we aimed to characterize the preparatory representation for novel, complex tasks in a time-resolved manner. We recently found that the neural encoding space for new instructions was structured by proactive control-related variable, and was further modulated by reward expectations (Palenciano et al., 2019). Here, we aimed to uncover the temporality of these effects on anticipatory coding. We performed model-based RSAs to identify the temporal windows where the anticipatory activity was structured by these variables. Additional MVPA analyses revealed when the EEG signal contained information about those variables.

The main RSA results showed that the instructions' anticipatory activations were structured by the integration demands during a stable time window 1100ms after instruction presentation. When we explored response set complexity, we found

that its effect on the encoding tended to increase towards the end of the instruction and the beginning of the ISI. Finally, target category structured task sets in two brief time windows just before and after the effect of dimension integration. The overall pattern of results obtained displayed a tendency toward our hypotheses regarding the three variables. On the one hand, the earlier effect of integration requirements and target category could reflect an initial and more abstract task goal reconfiguration process. On the other, the later effect of response set complexity could imply a shift towards concrete, stimulus-response setting. These findings, when taken together, could reflect the existence of a two-staged preparation process, as it has been previously proposed (De Baene & Brass, 2014).

The MVPAs carried out showed a presence of integration requirements information on distributed peaks of above-chance classification across the whole preparation period. The response set complexity was decoded on brain data during the later portion of the encoding interval and afterward jittered interval, converging with the RSA results. Finally, category information showed a wide decoding window during the first half of the preparation period. These findings showed that, among the three attributes, target category was the information most robustly readable in activity patterns. This converges with previous results showing widespread anticipatory category decoding in novel task contexts (González-García, Mas-Herrero, de Diego-Balaguer, & Ruz, 2016), and extended them, by showing the presence of this information only during the first second of instruction presentation. Regarding dimension integration, the decoding contrasted with the RSA results, not showing a clear temporal window where this information was decodable. This highlights the different information provided by

RSA and MVPA: while higher level task-set organization could follow the kind of integration required, this information itself may not need to be explicitly encoded. Nevertheless, response set complexity temporal profile did show consistency between RSA and MVPA, which generally could reflect the late nature of more concrete sensorimotor rule preparation. Overall, RSA and MVPA findings highlight the complexity inherent to proactive instruction preparation, characterized by dynamic representational organizations in combination with equally flexible coding of task-attributes.

Nonetheless, several particularities of our RSA and MVPA findings enforce serious caution when interpreting these results. Crucially, for the response set and category RSA models, we found two significant peaks during the baseline period, where no task-information was available. These could be a byproduct of the signal preprocessing pipeline followed, especially the smoothing conducted to increase the signal-to-noise ratio (Grootswagers et al., 2017). However, we averaged within 40ms window for this step, while the baseline significant peak for the response model appeared 100ms before instruction onset. Alternatively, the RSA could be capturing previous trial response or category information, due to the high sensitivity of this and other pattern analysis techniques (Arco, González-García, Díaz-Gutiérrez, Ramírez, & Ruz, 2018). Even when trials were randomly ordered, spurious dependencies among subsequent trials' conditions could drive significant results (e.g. Todd, Nystrom, & Cohen, 2013). Nonetheless, we did not find significant sequential correlations among main task conditions, and therefore, this explanation seems implausible. Finally, they could just reflect spurious positive results. At this point, though, it is necessary to stress the conservative nature of our permutation-based statistical approach, which used

empirical null distributions built upon 100 individual and 10000 group permutations.

A second general concern about the proactive-related findings, in both RSA and MVPA, is the presence of really brief significant clusters dispersed along the encoding period. Our multiple comparison correction involved computing a null distribution of cluster sizes, so the significant time points reported pertained to clusters larger than the 5% of above threshold peaks in 10000 permuted maps. Nonetheless, we were expecting to find more stable findings, as happened in the dimension integration model RSA and the category decoding. The prevalent fractionated pattern of RSA and MVPA results (see Fig. 6.1 and Fig. 6.3) cast doubts about the suitability of the signal during the instruction encoding period. Even after the exhaustive cleaning of ocular artifacts, residual noise could have contaminated the data. Nonetheless, a control MVPA decoding between reward conditions rendered robust and stable above-chance classification windows during both the reward cue and instruction epochs. Therefore, we did not obtain evidence supporting that our results were caused by contaminated signal.

Overall, we find it risky and complex to extract strong conclusions regarding the temporality of dimension integration, response set complexity and target category effects on novel instruction encoding. Crucially, our previous fMRI study showed the impact of the three variables on theoretically-congruent brain regions. It is also important to stress that the three variables displayed here an equivalent behavioral effect as in Palenciano et al. (2019), with strong effects of dimension integration and target category, while response set complexity effect remained undetected in both data sets. Consequently, factors inherent to the current study could underlie our pattern of results: we employed complex and

long-lasting stimuli (novel instructions) and applied a time-resolved analysis on EEG signals. First, the efficiency at task reconfiguration is variable both within and across subjects (Miyake & Friedman, 2012), and thus it is reasonable to assume that the different preparation subprocesses had particular and variable timings across trials. This would have a lighter effect on fMRI studies, as the nature of the BOLD signal entails averaging across several seconds. Nonetheless, EEG (as well as other electrophysiological recordings) captures these subtle timing differences, and neither the RSA nor MVPA employed were optimized for coping with this variability (Vidaurre, Myers, Stokes, Nobre, & Woolrich, 2019). New and promising approaches in pattern analyses may be able to incorporate this temporal information (Vidaurre et al., 2019), and its future employment with this dataset could help to better characterize our results. Another important concern is that neural representations studied here were based on data from all recording channels simultaneously (which is the default procedure in this kind of study; e.g.: Hebart et al., 2018). Contrary, in fMRI either regions of interest or searchlight procedures (Kriegeskorte et al., 2006) are followed, so that the activity patterns analyzed usually come from constrained brain regions. Recent advances in RSA allows combining EEG and fMRI (as well as other techniques; Kriegeskorte et al., 2008) data, together with theoretical models. This approach, known as fusion models (Cichy, Pantazis, & Oliva, 2014; Hebart et al., 2018), generates time courses indicating representational coherence among a particular point on EEG data, a particular region registered with fMRI and theoretical expectations. Thus, incorporating information about the regions involved in Palenciano et al (2019) could provide interesting insights. Finally, multiple lines of research highlight the structuring role of different frequency bands for

segregating top-down (in alpha and beta bands) and bottom-up (in gamma) information (Fries, 2015). Hence, it would be interesting to conduct further analysis directly on proactive control-related frequencies' power, to assess if that allows the detection of our effects of interest.

The second main goal of this work was to explore the temporal profile of the interplay between motivation and proactive preparation in novel instruction encoding. As expected, we replicated the reward-related general decrease in task-set dissimilarities (Palenciano et al., 2019). This finding is highly relevant when we consider than practiced and simpler task scenarios generated an opposite effect in the past – with motivation increasing task distinguishability (Etzel et al., 2016; Hall-McMaster et al., 2019). Obtaining the same pattern twice, with different neuroimaging techniques, is compelling evidence in favor of our motivation effect. Nevertheless, how more similar task representation could lead to better performance is still uncertain. Some tentative ideas can be extracted from the comparison of our paradigm with others. For example, Hall-McMaster and colleges (Hall-McMaster et al., 2019), also employing RSA on EEG data, found that high motivation enhanced the organizing effect of different task attributes on neural representation (corresponding to the Hypothesis 1 tested here). Crucially, they used a classic task switching paradigm (Monsell, 2003) in which only two cues were alternated, and employed bivalent stimuli which could generate strong conflict during the execution (i.e., when the irrelevant stimuli dimension was associated with a response incompatible with the one required by the rule). Instead, we used 192 different rules which were applied on novel combinations of stimuli. As there was no fixed stimulus-response mapping, no conflicting information was present at the execution. Thus, rule variability and/or

interference during performance could determine the most suitable strategy for optimizing task coding. Motivation would ultimately boost the most appropriate mechanism instead of exerting a fixed effect regardless of task context. Further research would be highly useful to understand better the intricate motivation-control relationship.

We also provided the time course of the motivation-induced decrease in instructions dissimilarities, which was characterized by two pronounced temporal windows. The first one displayed the strongest effects, and appeared early in the encoding, after 200ms. This could reflect the instantiation of a generic rule pattern ("*IF [stimuli] THEN [responses]*) in which the specific novel task parameters could be mapped during preparation. In this sense, as the pattern would be shared by all the instructions, it would generate decreases in representational dissimilarity. Previous fMRI findings (Bourguignon et al., 2018) also pointed toward the presence of this mechanism during novel instruction coding. In our experiment, it could also be boosted by reward expectations. A second, less pronounced window was found later on, towards the end of the epoch (approximately 2000ms after instruction onset). This result could still be generated by the updating of the generic rule pattern once the whole instruction is fully processed. Alternatively, it could also be explained from the compositional coding account, which proposes that individual task components are reused for generating novel rule representations. As these components would be shared across some instructions, its activation would also diminish dissimilarity. In any case, future studies explicitly assessing this possible interpretation are necessary to better understand the dynamic effect of motivation.

6.5. Conclusion

In conclusion, this study provided important insights into the temporal profile underlying proactive coding of new instructions. It was characterized by the dynamic establishment of representational encoding structures. We found a tendency towards early task-set organizations governed by abstract goal attributes (dimension integration and category), which later on followed response requirements. However, the novelty of the experimental setting and analysis followed entailed considerable variability in our results, enforcing caution regarding potential interpretations. Crucially, motivation displayed a strong decrease in task-set dissimilarities, replicating previous fMRI results (Palenciano et al., 2019). This finding and its temporal characterization are of high relevance to the current debate about control-motivation interactions. Specifically, we stress that the mechanisms by which reward boosts proactive control may depend on the particularities of the task context. Overall, the present work contributed to shed some light on an unexplored field on the instructed-behavior literature: its fine-grained temporal flexibility. Further application of recently developed analytical approaches to this data set could be key to shed additional light on this topic.

Chapter 7:

GENERAL DISCUSSION

7. GENERAL DISCUSSION

The goal of this thesis was to study the neural control mechanisms that scaffold performance guided by novel instructions. We addressed this aim through three neuroimaging studies, which explored in progressively increased detail the preparatory state engaged by instructions. The main results obtained will be briefly reviewed in the following section. Wrapping up our findings, we will describe three core notions about proactive control in novel task settings that were reached along this thesis. First, the deployment of proactive processes acts at overlapping, yet distinct, timescales. Second, these control mechanisms operate in interaction with motivation. Third, and finally, proactive control processes show a striking overlap on the IFS during novel and complex verbally instructed-behavior. We will conclude by highlighting the open questions that our research led us to. Addressing these is key both for advancing our understanding of novel instruction following and furthermore, for general models of action control.

7.1. Brief results summary.

In Study 1, we tried to characterize the overall transient and sustained control processes (Dosenbach et al., 2008) triggered by task novelty. Using fMRI and a mixed design (Petersen & Dubis, 2012), we found that mainly FPN regions were recruited at both timescales. This was especially the case for the IFS, transiently involved for practiced instruction encoding, and during the implementation of novel ones. This region was also associated with tonic activations through blocks of new tasks. We complementary assessed the maintenance of distributed rule representations within the trial window. Our results showed higher consistency

in the CON - an effect that was potentiated by the experience with the instructions. Thus, converging with previous literature (reviewed in Chapter 1), our findings supported the distinction between the FPN and CON (Crittenden, Mitchell, & Duncan, 2016; Dosenbach et al., 2007, 2006), although with a more complex subdivision of roles between the two systems. On the one hand, the FPN was key for this complex behavior, adopting flexible temporal dynamics depending on novelty. Crucially, at the phasic timescale, new instructions engaged the IFS later than practiced ones, which could reflect the most costly and prolonged preparation process for novel tasks (Cole et al., 2018). The sustained participation of this LPFC region could implicate more general, and compensatory, proactive control adjustments. On the other hand, CON regions were key in sustained task maintenance, but in a narrower timescale than previously reported, and especially for practiced tasks. This result contrasts with the scarce evidence supporting task coding in these areas (Crittenden et al., 2016; Woolgar, Jackson, & Duncan, 2016), and highlights the necessity of better understanding the computations carried out by CON regions.

In Study 2 we focused on the proactive preparation triggered by novel instructions. We assessed whether anticipatory activity was structured according to relevant task parameters. Employing fMRI and pattern analysis, we found flexible representational organizations across the brain. Dimension integration requirements organized the encoding in the IFS, while response complexity did so in the IPS and motor areas. Target category also guided task coding in the fusiform gyrus and the precuneus. Crucially, the three representational structures converged on the right IFS. Across all the regions explored, reward expectations modulated the encoding structure, making individual task-set representations

more similar to each other. These findings further characterize the mean increases of instruction encoding activations reported in these areas (Cole et al., 2010; Demanet et al., 2016; Hartstra et al., 2011; Ruge & Wolfensteller, 2010). We showed that the IFS and IPS were involved in the active representation of task-sets at different levels of abstraction. Moreover, the comprehensive task information reflected on the activation patters of the IFS stresses the importance of this region in orchestrating the preparation process. Crucially, perceptual and motor areas also adapted their tuning to upcoming targets according to response requirements. Future studies assessing effective connectivity will help to uncover whether or not FPN regions are the source of the representational calibration found in lower-level areas. Finally, the motivation effect was opposite to previous findings (Etzel et al., 2016), but also highly robust and relevant for performance. Addressing how the tuning to task parameters is modulated by control-motivation interactions is an important goal for future research.

Finally, in Study 3 we investigated the temporal unfolding of the preparatory processes triggered by novel instructions. We shifted to EEG recording and explored whether the encoding structures found in the previous study also influenced the dynamics of preparatory representations. We concurrently investigated the fine-grained modulation of motivation of these processes. Our results show a tendency toward earlier effects of dimension integration requirements and category, followed by response set complexity. This whole pattern fits with the presence of two sequential stages during preparation: the encoding of abstract task goals, followed by more specific stimulus-response rule reconfiguration (De Baene & Brass, 2014; Muhle-Karbe et al., 2014). However, it is important to stress that stronger evidence is required to add robustness to our

results. Importantly, we replicated the intriguing reward-related increases on task similarity, which followed a biphasic temporal profile. A first, early peak of task similarity increase appeared around 200ms after instruction onset. Interestingly, this could be reflecting the establishment of a general rule template, which should be updated with instructed content during task-set building (Bourguignon et al., 2018). The effect reappeared at a later window, toward instruction offset, but the significance of this effect is more uncertain and calls for more detailed investigations.

Overall, the present thesis provided several pieces of evidence that could be key for our understanding of how instructions are transformed into action-oriented representations to guide performance. In the following sections, we will describe in detail the main insights reached with this work.

7.2. Proactive control is deployed at different timescales in contexts of task novelty.

The link between proactive-control and instructed behavior is a cornerstone of this line of research since its beginning (Cole, Laurent, et al., 2013), and several studies have addressed novel task-set reconfiguration at the neural level (e.g. Cole, Etzel, Zacks, Schneider, & Braver, 2011; Dumontheil, Thompson, & Duncan, 2011; González-García, Arco, Palenciano, Ramírez, & Ruz, 2017; Ruge & Wolfensteller, 2010; Stocco, Lebiere, O'Reilly, & Anderson, 2012). However, a finer conceptualization of the specific processes underlying this phenomenon is lacking. In this thesis, we studied the neural basis of instructed-behavior employing multiple designs (event-related, mixed), analysis approaches (univariate, RSA, MVPA) and recording techniques (behavior, fMRI and EEG).

Thanks to this rich approach and the comprehensive data accumulated, we can shed some light on these processes.

The first relevant result pointed toward the participation of the IFS for both practiced and novel instructions, during their encoding and implementation, respectively (Study 1), a pattern resonating with previous fMRI and MEG data (Cole et al., 2010). It is tempting to interpret this neural signature as the culmination of task-set preparation, delayed for novel rules (Cole et al., 2018). However, alternative accounts also fit with these results, ranging from more general semantic instruction processing to more specific computations as the binding of novel stimulus-response associations (Demanet et al., 2016; Hartstra et al., 2012; Huang, Hazy, Herd, & O'Reilly, 2013). Crucially, compelling evidence for the IFS role in task-set representation was provided by Study 2. Results from this study showed that the anticipatory activity observed in this area was highly structured, reflecting the encoding of instructions in a space with axes defined by relevant task attributes. It is important to stress that the new instruction encoding found in the IFS was assessed in the context of fMRI and classic event-related designs – where several seconds of neural events are collapsed. Different approaches are needed to characterize complementary mechanisms with slower or faster temporal profiles, which could be also key for performance on novel tasks.

Influential proposals emphasize that proactive control also acts at longer timescales, displaying sustained activation patterns that transcend individual trials (Braver, Paxton, Locke, & Barch, 2009; Dosenbach et al., 2008). In cognitive terms, these would implement task-set maintenance through long periods of time. This sustained component is thought to be key for the stability of

our behavior –until a sudden disturbance appears (Braver, 2012a). Coinciding with this view, we found sustained activations across blocks of novel instructions in FPN regions (Study 1). As the rules were considerable variable within blocks, the classical interpretation in terms of task-set maintenance (Braver et al., 2009; Braver, Reynolds, & Donaldson, 2003; Jimura et al., 2010) does not fit with our results. Nevertheless, at this point, it is important to highlight the hierarchical nature of human control, stressed across multiple theoretical models (Badre, 2008; Badre & Nee, 2018; Duncan, 2010; Koechlin & Summerfield, 2007). Our complex organized behavior is thought to rely on multilevel control task representations (Duncan, 2010), which may involve high levels of abstraction, both in the temporal plane (Fuster, 2001) and regarding the task goal-structure (Badre, 2008). From this view, the individual instructions in our paradigm may have constituted subgoals to be accomplished at a short term. Concurrently, a higher level, general task model could be built upon the general indications given to the participants or their initial experience with the experimental settings (Bhandari & Duncan, 2014; Duncan, Emslie, Williams, Johnson, & Freer, 1996; Duncan et al., 2008; Niv, 2019). The maintenance of such model though blocks could be implemented by the sustained involvement of the FPN mentioned above. Nonetheless, little is known yet about information contained in such tonic activity patterns.

On the other side of the temporal spectrum, there are proactive adjustments taking place in quick succession at faster timescales. These would happen in the order of hundreds of milliseconds or below, and their dynamics could not be accessed by fMRI studies. In this line, successive transitions in task-set representations have been described in non-human primates with invasive

electrophysiological recordings (Sigala et al., 2008; Stokes et al., 2013). With the goal of addressing fast transitions in representations in a non-invasive fashion, we assessed the dynamic instantiation of variable representational structures during the encoding and preparation for novel instructed tasks (Study 3). Our data suggest that the initial stages of encoding organization were compatible with global task-goal reconfiguration (affected by task-set complexity and category), later shifting to a structure based on response requirements. These fast sequences could be the building blocks generating the rule encoding across FPN, perceptual and motor cortices shown in Study 2. Importantly, the reward-related decrease in task-set dissimilarities pointed toward an additional process taking place shortly after instruction presentation: the instantiation of a global rule pattern. This would follow the overall task structure used in our paradigm while leaving the trial relevant attributes non-specified (e.g.: "If there are [stimuli A] and [stimuli B], press [response A], otherwise [press B]"). Such a template could sustain task-set building by updating the instructed targets and responses. While the identification of this mechanism is based on its potentiation by reward, previous findings point toward the same process (Bourguignon et al., 2018; Stocco et al., 2012) underlying novel instructed performance.

Overall, our results showed a complex landscape with time-varying, overlapping proactive processes complementing each other during novel instruction preparation.

7.3. Proactive control – motivation interactions.

So far, we have addressed proactive mechanisms as isolated from other interrelated processes. Nonetheless, integration is ubiquitous in the human

brain, and accounting for the intricate relationships across domains is a necessity for a comprehensive understanding of neural cognitive function (Pessoa, 2017). In an attempt to move in this direction, we included a manipulation of motivation with economic incentives in our paradigm (Studies 2 and 3) and assessed how reward expectations affected novel task organization.

In this thesis, we uncovered three main aspects of the proactive control-motivation synergy in novel task contexts. First, in contexts characterized by variability in relevant rules where accomplishing the instructed task requires the combination of information, reward expectation increases similarity among task codes. This effect is behaviorally relevant: it displays robust correlations with reward-related performance improvements. Secondly, this finding is replicated across different brain regions, both at high and lower-level in the processing stream. In this sense, it resembles a general principle modulating task coding wherever it is relevant for preparation. Finally, it displays a dual temporal profile, biasing the encoding at the early beginning of the preparation and again, when the instruction information is no longer available.

Our interpretation considers that high motivational states increase the efficiency of novel task-set generation (for example, via rule template instauration, as previously mentioned). However, taking into account the scarce literature on this topic, making strong assumptions is risky and problematic. Furthermore, two unanswered questions come to mind regarding this explanation. In the spatial plane, does reward exert an equivalent effect on task coding across all the regions explored here? In the temporal one, are all the sequential time windows sensitive to reward reflecting a similar or different underlying process? While further research is needed, a key approach would be to combine the obtained fMRI and

EEG data (Cichy et al., 2014; Hebart et al., 2018). RSA-based fusion models allow the integration of representational spaces obtained from different sources and participants (Hebart, Donner, & Haynes, 2012; Kriegeskorte et al., 2008). Specifically, we think it would be key to track where in the brain the two reward-related peaks of increased task similarity are taking place.

To conclude, one final and critical issue must be addressed regarding motivation-control interplays. Twice we found an effect going in the opposite direction than previously reported (Etzel et al., 2016; Hall-McMaster et al., 2019). It could be argued that better task coding leads to more distinguishable rule representations, which should be reflected in more accurate classifications on MVPA (e.g.: Cole, Ito, & Braver, 2016) or increased dissimilarity distances in RSA (Bourguignon et al., 2018). In this line, reward expectations have been linked to increased task decoding accuracies in the only two studies that to our knowledge have explored this phenomenon to date (Etzel et al., 2016; Hall-McMaster et al., 2019). On the other hand, it could also be claimed reward-related boosts on task-set integration would improve the preparation on contexts benefiting from conjoint rule processing. The discrepancy between the two sets of results we argue could be based on the considerably distinct experimental settings employed. One crucial difference affected the task execution stage, which was not analyzed here, but whose demands could affect how rules are encoded in advance. Our instructions were applied over novel combinations of stimuli that did not generate conflict but rather required the eventual integration of codes to reach an accurate response. In contrast, previous studies used much simpler bivalent targets, with conflicting attributes in half of the trials. The presence of competition between stimulus dimensions may entail a context where maximal task separability improves

processing efficiency. Supporting this view, Hall-McMaster et al. (2019) found that reward-related benefits on rule coding were limited to switch trials –where interference from the previous task must be overcame-. Also, while our experiments exploited flexibility by using a total of 192 complex and abstract rules, the two studies from other laboratories alternated among only a couple of simpler ones. Maybe simpler task contexts would benefit to a higher extent from establishing clear rule distinctions, while a more variable one could exploit task commonalities to boost efficiency. Overall, motivation could potentiate the most suitable strategy upon specific task demands (Locke & Braver, 2008) instead of having fixed effects. In conclusion, these a priori conflicting results should open a window of opportunity to keep investigating how reward is integrated with control processes to optimize our performance.

7.4. Left IFS as a core region for flexible novel behavior.

We conclude by further highlighting one brain region that appeared in all the different findings conforming this thesis: the left IFS. Its involvement was a constant, regardless of the approach. Fig. 7.1.A displays the overlap between the results displaying different temporal profiles of activation for novel and practiced tasks (Study 1) and those showing structured anticipatory rule encoding (Study 2). While significant clusters were spread across the LPFC, when taking into account only the statistically significant voxels across the six results maps, all of them coincided in the posterior section of the IFS (see Fig. 7.1.B). Importantly, as Fig. 7.1.C (taken from Bourguignon, Braem, Hartstra, De Houwer, & Brass, 2018) shows, a remarkable similar anatomical pattern has been found in previous studies about novel (Demanet et al., 2016; Hartstra et al., 2012; Muhle-Karbe et al., 2017; Ruge & Wolfensteller, 2010) or complex (Reverberi, Gorgen, et al.,

2012) task processing. More generally, almost all current theoretical models about control assume a core role for the LPFC (e.g.: Badre, 2008; Botvinick & Cohen, 2014; Duncan, 2001; Koechlin, Ody, & Kouneiher, 2003; Miller & Cohen, 2001). It is reasonable to ask, thus, how is such functional diversity accomplished by this brain region?

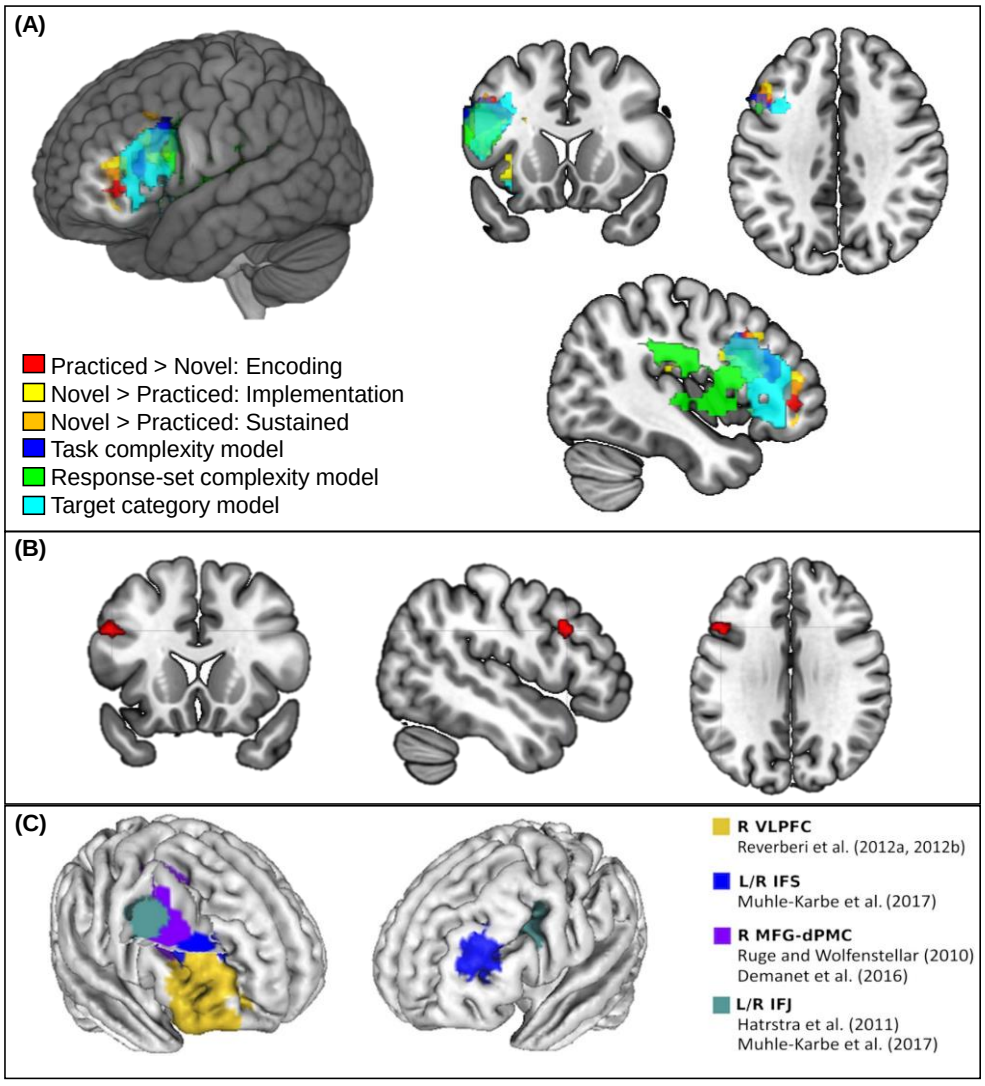

Figure 7.1: (A) Overlap across the findings of Studies 1 and 2. Results from univariate ANOVAs and t-test from Study 1 are shown in warm colors. The direction (greater transient or sustained activation for novel than practiced instructions or vice versa) is indicated in the legend. Results from model-based RSAs from Study 2 are displayed in cold colors. All the statistical maps were thresholded at $P < .05$, and FWE-corrected for multiple comparisons. Significant results displayed were restricted to the left lateral prefrontal cortex for illustrative purposes. **(B)** Statistically significant voxels coinciding across the six results from (A). **(C)** A similar overlap has been found in previous studies (adapted from Bourguignon et al., 2018).

We argue that what is special about the IFS, and the LPFC in general, is not the ability to implement a vast catalog of different cognitive processes. Rather, more general organizational principles governing prefrontal function could ultimately generate this heterogeneous pattern of results. Among them, neurons in LPFC display both abstract and adaptive (Duncan, 2001) receptive fields, responding to complex task goals in a context-relevant fashion (Woolgar, Hampshire, Thompson, & Duncan, 2011). This entails the LPFC with outstanding representational flexibility (Cole, Laurent, et al., 2013). Our findings, directly relating the IFS with success upon novel demands, stress that this portion of LPFC may be crucial to the first assembly of such complex representations. In addition, the nature of our experimental material adds important evidence regarding the nature of this adaptive coding in the human brain. Whereas previous studies used simpler materials where all the relevant variables were manipulated in different, separate trials, each of our complex novel instructions incorporates several variables at once. Our fMRI findings indicate that during an instruction episode, the IFS contains multiplexed neural patterns that refer to the different relevant dimensions coded by the instructions. Our EEG findings, on the other hand, suggest that these codes do not remain active throughout the whole interval, but activate at different time intervals. Although a full comprehension of this complex adaptive pattern is yet to be achieved, our strategy opens exciting new avenues of future research.

7.5. Open questions and future directions.

The interest in instructed-guided behavior is only recent in Cognitive Neuroscience. As a consequence, this is still a largely unexplored but exciting field. Among the multiple topics requiring further investigation, here we would

like to stress two of them. First, what kind of neural code could sustain the limitless variety of novel tasks a human can achieve? And second, how are novel task representations put together for the first time?

Regarding the first topic, two main proposals have been made. On the one hand, the compositional perspective emphasizes the reuse of previously acquired task component representations, in a combinatory fashion (Cole, Laurent, et al., 2013; Reverberi, Görgen, et al., 2012). The recursive application of such strategy could potentially explain highly complex goal encoding. On the other hand, the mixed-selectivity approach is based on explicit task representations via the non-linear combinations of relevant parameters (Rigotti et al., 2013). This proposal fits well with the adaptive coding principle found in LPFC (Duncan, 2001). Importantly, both perspectives are not incompatible a priori, and a combination of both could be key for flexible human cognition. Additionally, we advocate for addressing the dimensions of such encoding space. So far, we have pointed toward three potential axes (task and response set complexity, plus stimulus category) structuring LPFC representations. While further research is needed to ascertain the specificity of these variables for action-oriented task coding (Sobrado et al., *in prep*), it is equally important to explore additional dimensions organizing novel rule representation. In this sense, naturalistic experimental settings would be critical to capture the complexity and flexibility of our adaptive behavior. Data-driven approaches could also be helpful in identifying emerging representational structures (Huth, De Heer, Griffiths, Theunissen, & Gallant, 2016).

In second place, one of the most intriguing open questions refers to the assembly process itself. What are the mechanisms that allow a novel task representation to be built for the very first time? Addressing this topic is crucial for not falling into

implicit homuncular explanations. Investigations framed in computational modeling have stressed the role of hierarchical reinforcement learning in acquiring these control representations (Botvinick & Cohen, 2014; Niv, 2019). These models highlight the extraction process of complex, hierarchical task structures from experience. This approach, however, ignores the facilitating role of language (and symbolic transmission in general), which allows direct access to task procedures via instruction. Further research is needed to connect both areas of research, studying whether action-oriented representations built upon experience and instructions are equivalent or how they differ. Nevertheless, one key insight provided by the reinforcement learning perspective is the link between controlled and flexible behavior with more rigid learning principles (Braem & Egner, 2018). These have been traditionally seen as the two irreconcilable ends of a spectrum, with control being defined in opposition to automatic, habit-like behavior. Dissolving this dichotomy and addressing the learned nature of flexibility is thus an exciting new window into the understanding of cognitive control mechanisms.

Chapter 8:

CONCLUSIONS

8. CONCLUSIONS

- Guiding our behavior by novel instructions requires transient and sustained proactive processes implemented in the FPN – especially, on the left IFS. Costly novel task-set reconfiguration could rely on the long term maintenance of a more abstract and general task context models. These two proactive mechanisms could be multiplexed on the variable temporal dynamics followed by the IFS.

- The CON supports flexible preparation especially when some experience with instruction has been acquired. Rule representations are held in these areas from the encoding until stimuli are available – with practice improving the quality of this sustained coding. Consequently, CON could be key for experience-related increases in task processing efficiency.

- FPN and CON behave like segregated systems during novel instructed behavior. However, the distinction among them seems to be beyond their temporal profiles, as was previously proposed (Dosenbach et al., 2007). A more sophisticated distinction should be pursued, in which the information content, as well as its representational format, could play a key role.

- Novel instructions are anticipatorily coded in flexible representational spaces, whose axes are defined by relevant task parameters. This proactive tuning of encoding spaces could be a general principle for optimal preparation in variable scenarios. Crucially, this property is widely distributed across prefrontal and parietal control regions.

- The IFS encodes new instructions following a complex, overarching representational architecture, simultaneously organized by several task attributes. This points out the potential role of this region orchestrating the proactive adjustments – at least in contexts where higher flexibility is required.
- Motor and perceptual cortices also show structured anticipatory activations according to upcoming target and motor requirements. Importantly, this preparation state is engaged by abstract task information conveyed in verbal instructions. The potential biasing influence of the FPN on these lower-level regions during anticipatory task coding is still an open question on the field.
- Complex temporal dynamics underlie the codification of novel instructions. The encoding structure flexibly weights different task parameters throughout the preparation. Importantly, the representational geometry seems to transit from more abstract organizations to more specific, response-based ones. However, further investigation addressing this interpretation is needed.
- Motivational state modulates proactive task coding in a behavioral-relevant fashion, leading to performance improvements. On variable and novel scenarios, reward leads to more similar rule representations. Potential mechanisms boosted by motivation could be the establishment of a general rule template or the compositional reutilization of individual task components representations.

REFERENCES

Anderson, M. L., Kinnison, J., & Pessoa, L. (2013). Describing functional diversity of brain regions and brain networks. *Neuroimage*, 73, 50–58. https://doi.org/10.1016/j.neuroimage.2013.01.071

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

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, 12(5), 193-200. https://doi.org/10.1016/j.tics.2008.02.004

Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences*, 22(2), 170-188. https://doi.org/10.1016/j.tics.2017.11.005

Baldauf, D., & Desimone, R. (2014). Neural Mechanisms of Object-Based Attention. *Science*, 344(6182), 424–427. https://doi.org/10.1126/science.1247003

Barch, D. M., Braver, T. S., Sabb, F. W., & Noll, D. C. (2000). Anterior cingulate and the monitoring of response conflict: Evidence from an fMRI study of overt verb generation. *Journal of Cognitive Neuroscience*, 12, 298–309. https://doi.org/10.1162/089892900562110

Bhandari, A., & Duncan, J. (2014). Goal neglect and knowledge chunking in the construction of novel behaviour. *Cognition*, 130(1), 11–30. https://doi.org/10.1016/j.cognition.2013.08.013

Botvinick MM, Braver TS, Barch DM, Carter CS, & Cohen JD (2001). Conflict monitoring and cognitive control. *Psychol Rev*. 108:624–652.

https://doi.org/10.1037/0033-295X.108.3.624

Botvinick, M. M., & Braver, T. S. (2015). Motivation and Cognitive Control: From Behavior to Neural Mechanism, *Annual Review of Psychology*, 66(1), 83–113. https://doi.org/10.1146/annurev-psych-010814-015044

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, 38(6), 1249–1285. https://doi.org/10.1111/cogs.12126

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108, 624–652. https://doi:10.1037/0033-295X.108.3.624

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

Braem, S., & Egner, T. (2018). Getting a Grip on Cognitive Flexibility. *Current Directions in Psychological Science*, 27(6), 470–476. https://doi.org/10.1177/0963721418787475

Brass, M., Wenke, D., Spengler, S. & Waszak, F. (2009). Neural Correlates of Overcoming Interference from Instructed and Implemented Stimulus-Response Associations. *Journal of Neuroscience*, 29, 1766–1772. https://doi.org/10.1523/JNEUROSCI.5259-08.2009

Brass, M., & von Cramon, D. Y. (2004). Decomposing components of task preparation with functional magnetic resonance imaging. *Journal of*

Cognitive Neuroscience, 16, 609–620. https://doi.org/10.1162/089892904323057335

Brass, M., Derrfuss, J., Forstmann, B., & Cramon, D. Y. von. (2005). The role of the inferior frontal junction area in cognitive control. *Trends in Cognitive Sciences*, 9(7), 314–316. https://doi.org/10.1016/J.TICS.2005.05.001

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

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, 16, 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Braver, T. S., Paxton, J. L., Locke, H. S., & Barch, D. M. (2009). Flexible neural mechanisms of cognitive control within human prefrontal cortex. *Proceedings of the National Academy of Sciences*, 106(18), 7351–7356. https://doi.org/10.1073/pnas.0808187106

Braver, T. S., Reynolds, J. R., & Donaldson, D. I. (2003). Neural Mechanisms of Transient and Sustained Cognitive Control during Task Switching, Neuron, 39, 713–726. https://doi.org/10.1016/S0896-6273(03)00466-5

Bressler, S. L., & Richter, C. G. (2015). Interareal oscillatory synchronization in top-down neocortical processing. *Current Opinion in Neurobiology*, 31, 62–66. https://doi.org/10.1016/j.conb.2014.08.010

Brett, M., Anton, J., Valabregue, R., & Poline, J. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage*, 16, 99. Retrieved from

http://www.mrc-cbu.cam.ac.uk/Imaging/marsbar.html

Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307, 1118–1121. https://doi.org/10.1126/science.1105783

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, 315, 1860–1862. https://doi.org/10.1126/science.1138071

Buschman, T. J., & Miller, E. K. (2014). Goal-direction and top-down control. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369, 20130471. https://doi.org/10.1098/rstb.2013.0471

Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron*, 76, 838–846. https://doi.org/10.1016/j.neuron.2012.09.029

Capilla, A., Schoffelen, J.-M., Paterson, G., Thut, G., & Gross, J. (2014). Dissociated α- band modulations in the dorsal and ventral visual pathways in visuospatial attention and perception. *Cerebral Cortex*, 24, 550–561. https://doi.org/10.1093/cercor/bhs343

Chaumon, M., Bishop, D. V. M., & Busch, N. A. (2015). A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *Journal of Neuroscience Methods*, 250, 47–63. https://doi.org/10.1016/j.jneumeth.2015.02.025

Chiew, K. S., & Braver, T. S. (2016). Reward favors the prepared: Incentive and task-informative cues interact to enhance attentional control. *Journal of Experimental Psychology: Human Perception and Performance*, 42(1), 52–

66. https://doi.org/10.1037/xhp0000129

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3), 455–462. https://doi.org/10.1038/nn.3635

Clayton, M. S., Yeung, N., & Kadosh, R. C. (2015). The roles of cortical oscillations in sustained attention. *Trends in Cognitive Sciences*, 19, 188–195. https://doi.org/10.1016/j. tics.2015.02.004

Cole, M.W., & Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *Neuroimage*. 37:343–360. https://doi.org/10.1016/j.neuroimage.2007.03.071

Cole, M. W., Bagic, A., Kass, R., & Schneider, W. (2010). Prefrontal Dynamics Underlying Rapid Instructed Task Learning Reverse with Practice. *Journal of Neuroscience*, 30(42), 14245–14254. https://doi.org/10.1523/JNEUROSCI.1662-10.2010

Cole, M. W., Braver, T. S., & Meiran, N. (2017). The task novelty paradox: Flexible control of inflexible neural pathways during rapid instructed task learning. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2017.02.009

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, 1–13. https://doi.org/10.3389/fnhum.2011.00142

Cole, M. W., Ito, T., & Braver, T. S. (2016). The Behavioral Relevance of Task Information in Human Prefrontal Cortex. *Cerebral Cortex*, 26(6), 2497–

2505. https://doi.org/10.1093/cercor/bhv072

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., & 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

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*, 16(9), 1348–1355. https://doi.org/10.1038/nn.3470

Coste, C.P., & Kleinschmidt, A. 2016. Cingulo-opercular network activity maintains alertness. *Neuroimage*. 128:264–272. https://doi.org/10.1016/j.neuroimage.2016.01.026

Crittenden, B. M., Mitchell, D. J., & Duncan, J. (2015). Recruitment of the default mode network during a demanding act of executive control. *ELife*, 4, e06481. https://doi.org/10.7554/eLife.06481

Crittenden, B. M., Mitchell, D. J., & Duncan, J. (2016). Task Encoding across the Multiple Demand Cortex Is Consistent with a Frontoparietal and Cingulo-Opercular Dual Networks Distinction. *Journal of Neuroscience*, 36(23), 6147–6155. https://doi.org/10.1523/JNEUROSCI.4590-15.2016

Crone, E. A., Wendelken, C., Donohue, S. E., & Bunge, S. A. (2006). Neural evidence for dissociable components of task-switching. *Cerebral Cortex*, 16, 475–486.

https://doi.org/10.1093/cercor/bhi127

De Baene, W., & Brass, M. (2014). Dissociating strategy-dependent and independent components in task preparation. *Neuropsychologia*, 62, 331–340. https://doi.org/10.1016/j.neuropsychologia.2014.04.015

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. 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

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses and review. *Journal of Experimental Psychology: General*, 140(4), 552–572. https://doi.org/10.1037/a0023885

Dosenbach, N. U. F., Fair, D. A., Cohen, A. L., Schlaggar, B. L., & Petersen, S. E. (2008). A dual-networks architecture of top-down control. *Trends in Cognitive Sciences*, 12(3), 99–105. https://doi.org/10.1016/j.tics.2008.01.001

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K.,

Dosenbach, R. A. T., … Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences*, 104(26), 11073–11078. https://doi.org/10.1073/pnas.0704320104

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

Dubis, J.W., Siegel, J.S., Neta, M., Visscher, K.M., & Petersen, S.E. (2016). Tasks Driven by Perceptual Information Do Not Recruit Sustained BOLD Activity in Cingulo-Opercular Regions. *Cereb Cortex*. 26:192–201. https://doi.org/10.1093/cercor/bhu187

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. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, 2(11), 820–829. https://doi.org/10.1038/35097575

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

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology*, 30(3), 257–303. https://doi.org/10.1006/cogp.1996.0008

Duncan, J., Parr, A., Woolgar, A., Thompson, R., Bright, P., Cox, S., … Nimmo-Smith, I. (2008). Goal Neglect and Spearman's g: Competing Parts of a Complex Task. *Journal of Experimental Psychology: General*, 137(1), 131–148. https://doi.org/10.1037/0096-3445.137.1.131

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, 8, 1784– 1790. https://doi.org/10.1038/nn1594

Ekman, M., Derrfuss, J., Tittgemeyer, M., & Fiebach, C.J. (2012). Predicting errors from reconfiguration patterns in human brain networks. *Proc Natl Acad Sci.* 109:16714–16719. https://doi.org/10.1073/pnas.1207523109

Engelmann, J. B., Damaraju, E., Padmala, S., & Pessoa, L. (2009). Combined effects of attention and motivation on visual task performance: Transient and sustained motivational effects. *Frontiers in Human Neuroscience*, 3. https://doi.org/10.3389/neuro.09.004.2009

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16, 143–149. https://doi.org/10.3758/BF03203267

Esterman, M., & Yantis, S. (2010). Perceptual Expectation Evokes Category-Selective Cortical Activity. *Cerebral Cortex*, 20(5), 1245–1253. https://doi.org/10.1093/cercor/bhp188

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

Etzel, J. A., Cole, M. W., Zacks, J. M., Kay, K. N., & Braver, T. S. (2016). Reward Motivation Enhances Task Coding in Frontoparietal Cortex. Cerebral Cortex, 26(4), 1647–1659. https://doi.org/10.1093/cercor/bhu327

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110(41), 16616–16621. https://doi.org/10.1073/pnas.1315235110

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102, 9673– 9678. https://doi.org/10.1073/pnas.0504136102

Freedman, D.J., Riesenhuber, M., Poggio, T., & Miller, E.K. (2001). Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. *Science*, 291:312–316. https://doi.org/10.1126/science.272.5270.1905.

Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. *Neuron*, 88(1), 220–235. https://doi.org/10.1016/j.neuron.2015.09.034

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Fuster, J. M. (2001). The prefrontal cortex - An update: Time is of the essence. *Neuron*, 30(2), 319-333. https://doi.org/10.1016/S0896-6273(01)00285-9

Fuster, J. M. (2004). Upper processing stages of the perception-action cycle. *Trends in Cognitive Sciences,* 8(4), 143-145. https://doi.org/10.1016/j.tics.2004.02.004

Gläscher, J. & Gitelman, D. (2008). Contrast weights in flexible factorial design with multiple groups of subjects. Available in http://www.sbirc.ed.ac.uk/cyril/download/Contrast_Weighting_Glascher_Gitelman_2008.pdf

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., … & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536, 171–178. https://doi.org/10.1038/nature18933

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., Flounders, M.W., Chang, R., Baria, A.T., & He, B.J. (2018) Content-specific activity in frontoparietal and default-mode networks during prior-guided visual perception. *Elife*. 7. https://doi.org/10.7554/eLife.36068.

González-García, C., Mas-Herrero, E., de Diego-Balaguer, R., & Ruz, M. (2016). Task-specific preparatory neural activations in low-interference contexts. *Brain Structure and Function*, 221(8), 3997–4006. https://doi.org/10.1007/s00429-015-1141-5

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, 29(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Haber, S. N., & Knutson, B. (2009). The Reward Circuit: Linking Primate Anatomy

and Human Imaging. *Neuropsychopharmacology*, 35(10), 4–26. https://doi.org/10.1038/npp.2009.129

Hall-McMaster, S., Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2019). Reward boosts neural coding of task rules to optimise cognitive flexibility. *Journal of Neuroscience*, 0631–19. https://doi.org/10.1523/JNEUROSCI.0631-19.2019

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 Neuroscience*, 37(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci*. 7:523–534. https://doi.org/10.1038/nrn1931

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17, 323–328. https://doi.org/10.1016/j.cub.2006.11.072

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *ELife*, 7.

https://doi.org/10.7554/eLife.32816

Hebart, M. N., Donner, T. H., & Haynes, J. D. (2012). Human visual and parietal cortex encode visual choices independent of motor plans. *NeuroImage*, 63(3), 1393–1403. https://doi.org/10.1016/j.neuroimage.2012.08.027

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2014). 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

Heilbronner, S. R., & Hayden, B. Y. (2016). Dorsal anterior cingulate cortex: A bottom-up view. *Annual Review of Neuroscience*, 39, 149–170. https://doi.org/10.1146/annurev-neuro- 070815-013952

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109, 679–709. https://doi.org/10.1037/0033-295X.109.4.679

Huang, T. R., Hazy, T. E., Herd, S. A., & O'Reilly, R. C. (2013). Assembling old tricks for new tasks: A neural model of instructional learning and control. Journal of Cognitive Neuroscience, 25(6), 843–851. https://doi.org/10.1162/jocn_a_00365

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), 453–458. https://doi.org/10.1038/nature17637

JASP Team (2019). JASP (Version 0.11.1) [Computer software]

Jimura, K., Locke, H. S., & Braver, T. S. (2010). Prefrontal cortex mediation of cognitive enhancement in rewarding motivational contexts. *Proceedings of the National Academy of Sciences*, 107(19), 8871–8876. https://doi.org/10.1073/pnas.1002007107

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

Kleinsorge, T., & Rinkenauer, G. (2012). Effects of monetary incentives on task switching. *Experimental Psychology*, 59(4), 216–226. https://doi.org/10.1027/1618-3169/a000146

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6), 229–235. https://doi.org/10.1016/j.tics.2007.04.005

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, 302, 1181–1185. https://doi.org/10.1126/science.1088545

Krebs, R. M., Boehler, C. N., Roberts, K. C., Song, A. W., & Woldorff, M. G. (2012). The Involvement of the Dopaminergic Midbrain and Cortico-Striatal-Thalamic Circuits in the Integration of Reward Prospect and Attentional Task Demands. *Cerebral Cortex*, 22, 607–615. https://doi.org/10.1093/cercor/bhr134

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience,* 2, 4. https://doi.org/10.3389/neuro.06.004.2008

Landsiedel, J., & Gilbert, S. J. (2015). Creating external reminders for delayed intentions: Dissociable influence on "task-positive" and "task-negative" brain networks. *NeuroImage*, 104, 231–240. https://doi.org/10.1016/j.neuroimage.2014.10.021

Levy, B. J., & Wagner, A. D. (2011). Cognitive control and right ventrolateral prefrontal cortex: Reflexive reorienting, motor inhibition, and action updating. *Annals of the New York Academy of Sciences*, 1224, 40–62. https://doi.org/10.1111/j.1749-6632.2011.05958.x

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

Locke, H. S., & Braver, T. S. (2008). Motivational influences on cognitive control: Behavior, brain activation, and individual differences. *Cognitive, Affective, & Behavioral Neuroscience*, 8(1), 99–112. https://doi.org/10.3758/CABN.8.1.99

Lundqvist D, Flykt A, Öhman A. 1998. The Karolinska Directed Emotional Faces – KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet, ISBN 91-630-7164-9.

Luria, A. R. (1966). Higher Cortical Functions in Man. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4684-7741-2

Marini, F., Demeter, E., Roberts, K., Chelazzi, L., & Woldorff, M. (2016). Orchestrating proactive and reactive mechanisms for filtering distracting information: Brain-beha- vior relationships revealed by a mixed-design fMRI study. *The Journal of Neuroscience*, 36, 988–1000. https://doi.org/10.1523/JNEUROSCI.2966-15.2016

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *J Exp Psychol Learn Mem Cogn*, 22, 1423–1442. https://doi.org/10.1037/0278-7393.22.6.1423

Meiran, N. (2010). Task Switching: Mechanisms Underlying Rigid vs. Flexible Self-control. In: Hassin, R; Ochsner, K.; Trope Y, editor. Self-control in society, mind and brain. New York (NY): Oxford University Press, p. 202–220.

Meiran, N., Pereg, M., Kessler, Y., Cole, M. W., & Braver, T. S. (2015). The power of instructions: Proactive configuration of stimulus–response translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3), 768–786. https://doi.org/10.1037/xlm0000063

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

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, 21(1), 8–14. https://doi.org/10.1177/0963721411429458

Momennejad, I., & Haynes, J.-D. (2013). Encoding of prospective tasks in the human prefrontal cortex under varying task loads. *The Journal of*

Neuroscience, 33, 17342–17349. https://doi.org/10.1523/JNEUROSCI.0492-13.2013

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7, 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Muhle-Karbe, P. S., Andres, M., & Brass, M. (2014). Transcranial magnetic stimulation dissociates prefrontal and parietal contributions to task preparation. Journal of Neuroscience, 34(37), 12481–12489. https://doi.org/10.1523/JNEUROSCI.4931-13.2014

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., Poline, J.-B., & Poldrack, R. A. (2015). Orthogonalization of Regressors in fMRI Models. *PLOS ONE*, 10(4), e0126255. https://doi.org/10.1371/journal.pone.0126255

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti di Oleggio Castello, M., … Haxby, J. V. (2017). Attention Selectively Reshapes the Geometry of Distributed Semantic Representation. *Cerebral Cortex*, 27(8), 4277–4291. https://doi.org/10.1093/cercor/bhx138

Nichols, T., Brett, M., Andersson, J., Wager, T., & Poline, J. B. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, 25(3), 653–660. https://doi.org/10.1016/j.neuroimage.2004.12.005

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS*

Computational Biology, 10(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, 22(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

Norman, D. A., & Shallice, T. (1986). Attention to Action. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), Consciousness and Self-Regulation: Advances in Research and Theory Volume 4 (pp. 1–18). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4757-0629-1_1

Padmala, S., & Pessoa, L. (2011). Reward Reduces Conflict by Enhancing Attentional Control and Biasing Visual Cortical Processing. *Journal of Cognitive Neuroscience*, 23(11), 3419–3432. https://doi.org/10.1162/jocn_a_00011

Palenciano, A.F., Díaz-Gutiérrez, P., González-García, C., & Ruz, M. (2017). Neural mechanisms of cognitive control/Mecanismos neurales de control cognitivo. *Estudios de Psicología*, 38, 311–337. https://doi.org/10.1080/02109395.2017.1305060

Palenciano, A. F., González-García, C., Arco, J. E., & Ruz, M. (2018). Transient and Sustained Control Mechanisms Supporting Novel Instructed Behavior. *Cerebral Cortex*, 29(9), 3948–3960. https://doi.org/10.1093/cercor/bhy273

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

Parro, C., Dixon, M. L., & Christoff, K. (2017). The Neural Basis of Motivational Influences on Cognitive Control. Human Brain Mapping, 39(12), 5097–5111 https://doi.org/10.1101/113126

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl), S199-209. https://doi.org/10.1016/j.neuroimage.2008.11.007

Pessoa, L. (2009). How do emotion and motivation direct executive control? *Trends in Cognitive Sciences*, 13(4), 160–166. https://doi.org/10.1016/j.tics.2009.01.006

Pessoa, L. (2017). Cognitive-motivational interactions: Beyond boxes-and-arrows models of the mind-brain. *Motivation Science*, 3(3), 287–303. https://doi.org/10.1037/mot0000074

Petersen, S. E., & Dubis, J. W. (2012). The mixed block/event-related design. *NeuroImage*, 62(2), 1177–1184. https://doi.org/10.1016/j.neuroimage.2011.09.084

Pischedda, D., Görgen, K., Haynes, J.-D., & Reverberi, C. (2017). Neural Representations of Hierarchical Rule Sets: The Human Control System Represents Rules Irrespective of the Hierarchical Level to Which They Belong. *Journal of Neurosciene,* 37(50), 12281-12296 https://doi.org/10.1523/JNEUROSCI.3088-16.2017

Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, 174, 340–351. https://doi.org/10.1016/J.NEUROIMAGE.2018.03.041

Posner, M., & Petersen, S. E. (1990). The attention system of the human brain.

Annual Review of Neuroscience, 13, 25–42. https://doi.org/10.1146/annurev.ne.13.030190.000325

Qiao, L., Zhang, L., Chen, A., & Egner, T. (2017). Dynamic Trial-by-Trial Re-Coding of Task-Set Representations in Frontoparietal Cortex Mediates Behavioral Flexibility. *Journal of Neuroscience,* 37, 11037-11050. https://doi.org/10.1523/JNEUROSCI.0935-17.2017

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98, 676–682. https://doi.org/10.1073/pnas.98.2.676

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

Reverberi, C., Gorgen, K., & Haynes, J.-D. (2012). Distributed Representations of Rule Identity and Rule Order in Human Frontal Cortex and Striatum. *Journal of Neuroscience*, 32(48), 17420–17430. https://doi.org/10.1523/JNEUROSCI.2344-12.2012

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), 585–590. https://doi.org/10.1038/nature12160

Ritchie, J. B., Bracci, S., & Op de Beeck, H. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. *NeuroImage*, 148, 197–200. https://doi.org/10.1016/j.neuroimage.2016.12.079

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124(2), 207–231. https://doi.org/10.1037/0096-3445.124.2.207

Rowe, J. B., Eckstein, D., Braver, T. S., & Owen, A. M. (2008). How does reward expectation influence cognition in the human brain? *Journal of Cognitive Neuroscience*, 20(11), 1980–1992. https://doi.org/10.1162/jocn.2008.20140

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27, 763–797. https://doi.org/10.1037/0096-1523.27.4.763

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

Ruge, H., & Wolfensteller, U. (2010). Rapid formation of pragmatic rule representations in the human brain during instruction-based learning. *Cerebral Cortex*, 20(7), 1656–1667. https://doi.org/10.1093/cercor/bhp228

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., & Kastner, S. (2012). The pulvinar regulates information transmission between cortical areas based on attention demands. *Science*, 337, 753–756. https://doi.org/10.1126/science.1223082

Sadaghiani, S., & D'Esposito, M. (2015). Functional characterization of the cingulo-opercular network in the maintenance of tonic alertness. *Cerebral Cortex*,

25(9), 2763–2773. https://doi.org/10.1093/cercor/bhu072

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

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

Sakai, K., & Passingham, R. E. (2006). Prefrontal set activity predicts rule-specific neural processing during subsequent cognitive performance. *Journal of Neuroscience*, 26(4), 1211–1218. https://doi.org/10.1523/JNEUROSCI.3887-05.2006

Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H. … Greicius, M. D. (2007). Dissociable Intrinsic Connectivity Networks for Salience Processing and Executive Control. *Journal of Neuroscience*, 27(9), 2349–2356. https://doi.org/10.1523/JNEUROSCI.5587-06.2007

Shen, Y. J., & Chun, M. M. (2011). Increases in rewards promote flexible behavior. *Attention, Perception, & Psychophysics*, 73(3), 938–952. https://doi.org/10.3758/s13414-010-0065-7

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79, 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Sigala, N., Kusunoki, M., Nimmo-Smith, I., Gaffan, D., & Duncan, J. (2008). Hierarchical coding for sequential task events in the monkey prefrontal cortex. *Proceedings of the National Academy of Sciences*, 105(33), 11969–11974. https://doi.org/10.1073/pnas.0802569105

Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, 65, 69–82. https://doi.org/10.1016/j.neuroimage.2012.09.063

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. Co*gnitive, Affective and Behavioral Neuroscience*, 12(4), 611–628. https://doi.org/10.3758/s13415-012-0117-7

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., Thompson, R., Nobre, A. C., & Duncan, J. (2009). Shape-specific preparatory activity mediates attention to targets in human visual cortex. *Proceedings of the National Academy of Sciences*, 106(46), 19569–19574. https://doi.org/10.1073/pnas.0905306106

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology, 18, 643–662 https://doi.org/.10.1037/h0054651

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, 77, 157–165. https://doi.org/10.1016/j.neuroimage.2013.03.039

Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A. … Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, 168(3), 242–249.

https://doi.org/10.1016/j.psychres.2008.05.006

Townsend, J., & Ashby, G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), Cognitive theory (Vol. 3, pp. 200–239). Hillsdale, N.J: Erlbaum. Retrieved from https://labs.psych.ucsb.edu/ashby/gregory/publications/281

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Vidaurre, D., Myers, N. E., Stokes, M., Nobre, A. C., & Woolrich, M. W. (2019). Temporally Unconstrained Decoding Reveals Consistent but Time-Varying Stages of Stimulus Processing. *Cerebral Cortex*, 29(2), 863–874. https://doi.org/10.1093/cercor/bhy290

Visscher, K. M., Miezin, F. M., Kelly, J. E., Buckner, R. L., Donaldson, D. I., McAvoy, M. P., … Petersen, S. E. (2003). Mixed blocked/event-related designs separate transient and sustained activity in fMRI. *NeuroImage*, 19(4), 1694–1708. https://doi.org/10.1016/S1053-8119(03)00178-2

Voytek, B., Kayser, A., Badre, D., Fegen, D., Chang, E. F., Crone, N. E. … & D'Esposito, M. (2015). Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nature Neuroscience*, 18, 1318–1324. https://doi.org/10.1038/nn.4071

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

Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J., & Wagner, A. D. (2014). Frontoparietal representations of task context support the flexible control of goal-directed cognition. *Journal of Neuroscience*, 34(32), 10743–10755. https://doi.org/10.1523/JNEUROSCI.5282-13.2014

Wisniewski, D., Forstmann, B., & Brass, M. (2018). How exerting control over outcomes affects the neural coding of tasks and outcomes. *BioRxiv*. https://doi.org/10.1101/375642

Wisniewski, D., Reverberi, C., Momennejad, I., Kahnt, T., & Haynes, J. D. (2015). The role of the parietal cortex in the representation of task–reward associations. *Journal of Cognitive Neuroscience*, 35, 12355–12365. https://doi.org/10.1523/JNEUROSCI.4882-14.2015

Wisniewski, D., Reverberi, C., Tusche, A., & Haynes, J.-D. (2015). The neural representation of voluntary task-set selection in dynamic environments. *Cerebral Cortex*, 25(12), 4715–4726. https://doi.org/10.1093/cercor/bhu155

Woolgar, A., Hampshire, A., Thompson, R., & Duncan, J. (2011). Adaptive Coding of Task-Relevant Information in Human Frontoparietal Cortex. J*ournal of Neuroscience*, 31(41), 14592–14599. https://doi.org/10.1523/JNEUROSCI.2616-11.2011

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, 1433–1454. https://doi.org/10.1162/jocn_a_00981

RESUMEN

Los humanos destacamos por nuestra rápida adaptación a medios cambiantes. A su base, se encuentra la capacidad de implementar instrucciones en nuestros actos. Este medio de adquirir conductas supone una alternativa más rápida y eficaz que el aprendizaje por ensayo y error, del cual dependen otras especies animales ante la novedad. Por ello, es de enorme importancia para nuestra especie. En esta tesis, exploramos los mecanismos neurales a la base de esta compleja conducta.

El comportamiento en base a instrucciones se sustenta en el control cognitivo, un conjunto de procesos de alto nivel encaminados a guiar nuestra conducta a objetivos que no son alcanzables mediante patrones de conducta automáticos (Norman & Shallice, 1986). Estos mecanismos han sido ampliamente estudiados en Neurociencia Cognitiva, en especial con Resonancia Magnética Funcional (RMf). La investigación ha convergido en el rol de dos redes, una fronto-parietal y otra cíngulo-opercualar, que implementan el control actuando a distintas escalas temporales (Dosenbach, Fair, Cohen, Schlaggar, & Petersen, 2008). Sin embargo, los paradigmas empleados para ello están basados en tareas repetitivas y simples, dejando sin explorar gran parte de nuestra conducta flexible.

Investigaciones recientes asocian el seguimiento de instrucciones con el control proactivo (Cole, Braver, & Meiran, 2017; Cole, Patrick, & Braver, 2018), referido a ajustes anticipatorios que nos preparan para futuras demandas (Braver, 2012). Estos procesos transforman las instrucciones en representaciones de control (Cole, Laurent, & Stocco, 2013), conocidas como sets de tarea (Sakai, 2008), que contienen los parámetros relevantes para la ejeución (estímulos, respuestas, reglas). Los sets de tarea sesgan el procesamiento en regiones motoras y perceptivas, y en última instancia, guian nuestras acciones (Miller & Cohen,

2001). No obstante, se desconocen los mecanismos neurales que median esta preparación en contexto de novedad.

En esta tesis, hacemos uso de RMf y electroencefalografía (EEG) para investigar los procesos neurales de control que permiten guiar nuestro comportamiento en base a instrucciones. A lo largo de tres estudios, intentamos abordar cuatro objetivos específicos, detallados a continuación.

El primer estudio de RMf de la tesis (Palenciano, González-García, Arco, & Ruz, 2018) buscó conocer la participación transitoria y sostenida de las redes de control fronto-parietal y cíngulo-opercular durante el seguimiento de instrucciones. De esta forma, además de indagar en los procesos de control durante el comportamiento instruido, también buscamos estender el modelo dual de Dosenbach a contextos de novedad. Para ello, usamos un paradigma basado en la codificación e implementación de instrucciones verbales (González-García, Arco, Palenciano, Ramírez, & Ruz, 2017). Este fue adaptado a un diseño mixto, de bloques y eventos (Petersen & Dubis, 2012), que nos permitió la estimación simultánea de activaciones tónicas y fásicas (Visscher et al., 2003).

A continuación, en un segundo estudio de RMf (Palenciano, González-García, Arco, Pessoa, & Ruz, 2019), exploramos la preparación proactiva de instrucciones, focalizándonos en cómo los nuevos sets de tarea son representados en patrones de activación multivoxel (Haxby, Connolly, & Guntupalli, 2014). Específicamente, buscamos estudiar si parámetros relevantes para la ejecución estructuran la forma en que los sets de tarea se codifican anticipatoriamente. Para ello, generamos instrucciones manipulando la necesidad de integración a través de dimensiones estimulares, la complejidad del set de respuesta y la categoría del

target. El empleo de Análisis de Similitud Representacional (RSA; Kriegeskorte, Mur, & Bandettini, 2008) nos permitió estimar la estructura subyacente a la codificación de instrucciones en distintas localizaciones del cerebro (Kriegeskorte, Goebel, & Bandettini, 2006). Esta se comparó con modelos derivados de las tres variables manipuladas. De esta forma, evaluamos si la actividad anticipatoria en distintas regiones estaba organizada de acuerdo a uno o varios de los parámetros relevantes para la tarea.

Basándonos en el estudio anterior, un tercer objetivo consisitó en estudiar las dinámicas temporales que subyacen a las distintas estructuras representacionales antes descritas, durante la preparación de nuevas tareas. De esta forma, buscamos extraer el curso temporal caracterizando el efecto de los parámetros de tarea manipulados en el segundo estudio. Para ello, llevamos a cabo un tercer experimento, registrando datos de EEG de alta densidad, y replicando el mismo paradigma experiemntal que en el estudio anterior. Además, adaptamos el RSA para poder llevarlo a cabo de forma resuelta en el tiempo.

Por último, un cuarto objetivo perseguido en la tesis fue evaluar la interacción entre control proactivo y motivación, en contextos de novedad. Las expectativas de recompensa parecen incrementar la eficacia de mecanismos prapratorios (Chiew & Braver, 2016), habiéndose asociado este efecto a una mejora en la fidelidad de las representaciones de reglas relevantes (Etzel, Cole, Zacks, Kay, & Braver, 2016; Hall-McMaster, Muhle-Karbe, Myers, & Stokes, 2019). Sin embargo, este efecto se ha estudiado únicamente en tareas repetitivas y simples. En el segundo y tercer estudio, decidimos incluir incentivos económicos para evaluar si este efecto en la codificación se extrapolaba a contextos nuevos y variables. De forma importante, pudimos explorar el efecto de motivación a

través de distintas regiones cerebrales, y con alta precisión temporal, gracias a la recogida de datos de RMf y EEG, respectivamente.

Gracias a nuestra aproximación multidisciplinar, basada en distintas técnicas de registro y el empleo de métodos de análisis uni y multivariados, esta tesis ha aportado un conjunto de resultados que ayudan a comprender la generación de conducta nueva en base a instrucciones. Tres conclusiones principales emergieron de nuestros datos.

En primer lugar, el seguimiento de instrucciones descansa sobre procesos proactivos que se desarrollan a distintas escalas temporales. La corteza prefrontal lateral se encarga de la generación del set de tarea novedoso dentro de un mismo ensayo. Adicionalmente, esta región presenta actividad sostenida a lo largo de bloques, potencialmente implicando la instauración de un modelo general de tarea que complementa a la preparación de reglas individuales. Por último, dentro del intervalo de preparación, se dan dinámicas rápidas en cuanto a los parámetros estructurando la codificación de sets de tarea.

En segundo lugar, la motivación afecta la codificación proactiva de instrucciones robustamente, haciendo más similares entre sí las representaciones de distintas tareas. Este efecto está relacionado con la posterior ejecución en la tarea, enfatizando su relevancia y la necesidad de explorarlo en más detalle en el futuro.

Por último, nuestros resultados apuntan a la importancia de una región específica del surco frontal inferior (IFS) para el seguimiento de instrucciones nuevas. El IFS se vio involucrado con distintos perfiles temporales, y además, representó instrucciones novedosas de forma multidimensional, simultáneamente incorporando distintos parámetros en la codificación. Por ello, indagar en los

principios que subyacen al funcionamiento de esta versátil área es crucial para profundizar en el conocimiento del seguimiento de instrucciones, pero también para modelos generales de control cognitivo.