Winter 12-2019

Functional dissociations revealed by representational similarity analysis of color-word Stroop

Michael Freund
Washington University in St. Louis

Washington University in St. Louis

Department of Psychological & Brain Sciences

Functional Dissociations Revealed by Representational Similarity Analysis of Color-Word Stroop

by

Michael Freund

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

December 2019
Saint Louis, Missouri

Contents

List of Tables

List of Figures

Acknowledgments

First, I would like to thank my advisor, Dr. Todd Braver, for his support during this process. I would also like to thank Drs. Julie Bugg and Jeff Zacks for serving on my thesis committee. Finally, as this work was financially supported by grant NIH R37MH066078 and the Cognitive, Computational, and Systems Neuroscience pathway award, I would like to thank American taxpayers and the McDonnell Center for Systems Neuroscience.

Michael Freund

Washington University in Saint Louis

December 2019

ABSTRACT OF THE THESIS

Functional Dissociations Revealed by Representational Similarity Analysis of Color-Word Stroop

by

Michael Freund

Master of Arts in Psychology

Washington University in St. Louis, December 2019

Research Advisor: Professor Todd Braver

The color-word Stroop task is often used in cognitive neuroscience as a common platform for both theoretical and experimental approaches to cognitive control. Yet traditionally, there has been tension between these two approaches. Theoretical models of Stroop have focused on representation: for example, how distributed and overlapping representations of the two stimulus dimensions (color, word) are prioritized, and how conflict between these dimensions is represented and used to regulate control. In contrast, neuroimaging experiments have primarily focused on 'univariately' (uniformly) mapping the effects of conflict to particular brain regions. This focus on univariate changes in brain activity limits the specificity with which neural representations can be measured — which limits the bearing of results on representational models. To address this limitation, the current study provides a novel, retrospective application of representational similarity analysis (RSA), a multivariate analytic approach that enables specification and comparison of representational models, to functional magnetic resonance imaging data acquired while participants (N=49) performed the classic

color-word Stroop task. Through RSA, we disentangled coding of the target (color naming), distractor (word reading), and congruency (conflict) dimensions across cortex, observing robust and predicted dissociations in the neuroanatomical profile, representational structure, and functional relevance of these distinct coding schemes. These results highlight the utility of RSA as tool for addressing key questions in cognitive control, and we provide guidance on how to apply, both retrospectively and prospectively, this technique in neuroimaging.

Chapter 1

Introduction

The color-word Stroop task is a hallmark paradigm of cognitive control (Stroop, 1935; see MacLeod, 1991 for a not-so-recent review). Within a single multidimensional stimulus, the task straightforwardly captures what is thought to be an essential cognitive control function: enabling the selection of a less automatic target process (i.e., color naming) in the face of concurrent activation from a more automatic distractor process (i.e., word reading). Because of its simplicity, the Stroop paradigm of conflicting task dimensions has afforded a platform useful for developing theories of cognitive control. But, although the Stroop task has been used in investigation for almost 100 years, there is much we still do not understand about how theorized target and distractor processes are embedded and regulated within neural systems.

A useful first step to understanding the Stroop task and the kind of cognitive control it demands is to decompose the task into different *dimensions* and investigate how these dimensions may be represented in mind and brain. In particular, influential cognitive models of color-word Stroop explicitly represent *target* and *distractor* dimensions, corresponding to hue and wordform identities of the compound stimulus, which respectively feed into parallel streams of color naming and word reading processes (e.g., Cohen, Dunbar, & McClelland,

1990; Logan, 1980). These models have provided an important formal backdrop for a general neuroscientific framework of control, in which dorsolateral prefrontal cortex (dlPFC) and associated neural systems (e.g., intraparietal sulcal cortex, or IPS) preferentially encode features related to the target dimension, from abstract goals and task rules, to concrete stimulus and response information (Duncan, 2001; Miller & Cohen, 2001). By way of long-range excitatory projections and local (lateral) inhibition, these prefrontal target representations are thought to guide the flow of activation along the target pathway while inhibiting propagation along the distractor (Miller & Cohen, 2001; Munakata et al., 2011).

In parallel, other theoretical accounts have suggested the importance of a third, more abstract, Stroop dimension of *conflict* or *congruency*, which corresponds to whether the target and distractor dimensions indicate identical or conflicting responses. This property is highly informative of whether controlled processing may be useful on a given trial, and is hypothesized to be encoded by dorsomedial prefrontal (dmPFC), with a focus in anterior cingulate cortex (ACC) — which may serve, in part, to dynamically recruit a broader network of control systems (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Shenhav, Botvinick, & Cohen, 2013). In conjunction with substantial evidence indicating dmPFC encodes errors and performance-related information (e.g., Ito, Stuphorn, Brown, & Schall, 2003; Bonini et al., 2014; Brown & Braver, 2005; Sarafyazd & Jazayeri, 2019), this perspective has also supported a "dual mechanisms" framework of control. According to this framework, there are two control "strategies" at participants' disposal for successful task performance, with dissociable neural substrates and relations to behavior (Braver, 2012). On one hand, participants may proactively engage with the task, relying on structures within lPFC and IPS to implement top-down control. On the other, participants may adopt a reactive strategy,

loading more heavily on conflict and performance-monitoring functions of dmPFC. The optimal strategy for performance depends on contextual factors. For example, when control is likely to be required, proactive control may be beneficial, whereas reactive may be costly.

Support for these Stroop frameworks has been bolstered by decades of functional magnetic resonance imaging (fMRI) research — albeit at a course-grained level. Incongruent Stroop trials consistently evoke increased levels of activity in dlPFC, IPS, and dmPFC (Cieslik, Mueller, Eickhoff, Langner, & Eickhoff, 2015; Nee, Wager, & Jonides, 2007). Clear dissociations have emerged in the dynamics and behavioral relevance of these dlPFC and dmPFC activations: a stronger proactive dlPFC response, but weaker reactive dmPFC response, are associated with reduced Stroop interference (e.g., MacDonald, Cohen, Andrew Stenger, & Carter, 2000; for review, see Braver, 2012). Further, trial-by-trial modulations in control have been linked to enhancement of posterior sensory regions associated with representation of the target dimension (Egner & Hirsch, 2005).

The extent to which these findings bear on theory is limited, however, because it is less clear what information these fronto-parietal activations may contain. For example, given that dlPFC has been shown to encode multiple task features, does increased incongruent-trial activity in this region reflect conflict coding, or strengthened target coding as a result of conflict? This ambiguity is mitigated by focusing on modulations in activation within posterior sensory cortices (Egner & Hirsch, 2005) — a putative consequence of dlPFC function — as these regions tend to respond in a more specific manner (e.g., to faces). But, such a downstream investigatory angle is impoverished, as well, as attentional modulation of particular ventral visual "hubs" is unlikely to sufficiently account for the regulation of interference in the wide variety of ways it can arise — particularly in the Stroop task, which is thought to have a more central locus of interference (e.g., Duncan-Johnson & Kopell, 1981; MacLeod,

1991). Thus, testing these hypotheses with traditional neuroimaging methods has remained difficult.

This difficulty arises because the traditional analytic technique used in these studies (and in cognitive control research more broadly), "univariate voxel-wise encoding" analysis (Friston et al., 1994; Worsley et al., 1996) was developed for a fundamentally different purpose than to estimate regional representations. Rather, the purpose of a univariate analysis is to estimate the overall level of activity (e.g., mean) within a given region of interest (ROI) evoked by particular task conditions. In most cases, this goal is in opposition to one of isolating and estimating representations (e.g., of particular Stroop dimensions), as task variables are generally not thought to be mapped to cortical areas in a one-to-one manner. For example, in general, cortical areas are not thought to encode target information in a uniform, scalar manner (i.e., in which the level of activity directly indicates the extent to which color-related information is being processed — let alone *which* color is being processed). Instead, encoding of these types of variables is thought to occur in a given cortical area in a spatially distributed and overlapping manner (i.e., through a neural population-level code; Hebb, 1949; Saxena & Cunningham, 2019). As a result, the level of activity in a given region will likely reflect processing of a mixture of task dimensions (e.g., target and distractor), and will likely not distinguish particular exemplars of these dimensions (e.g., the hues blue and red would be expected to evoke similar mean levels of activity within a cortical region). Despite this inadequacy, univariate methods have been the overwhelmingly used analytic framework in Stroop investigations. This has resulted in a substantial inferential gap between cognitive neuroscience findings and cognitive theory.

Thus, what is currently needed is a method that "unmixes" the multivariate response of a region, furnishing interpretable measures of representation of particular task dimensions.

Multivariate pattern analysis methods of fMRI data accomplish exactly this purpose. By shifting to a more spatially fine-grained level of analysis, these methods capitalize on the fact that response preferences for subtle task features (e.g., particular hues or wordforms) are much more likely to emerge in individual voxels rather than the regional mean. Further, a form of pattern analysis, termed representational similarity analysis (RSA), enables particular models to be fit to the representations of regions, thus enabling different hypotheses regarding neural representations to be evaluated and compared (Kriegeskorte, 2008; Kriegeskorte & Kievit, 2013; Nili et al., 2014). Although these methods are well-aligned to test frameworks of cognitive control function, they have been infrequently used to this end — and surprisingly, to the best of our knowledge, RSA has never been applied to investigate coding of task dimensions in the color-word Stroop task.

Here, we conduct a retrospective analysis of an fMRI dataset acquired while subjects performed a color-word Stroop task, to provide an initial "proof of concept" demonstration of the feasibility and potential theoretical advantages of using RSA for estimating the neural processing of component Stroop dimensions. In combination with a multi-modal parcellation atlas (Glasser et al., 2016), we used RSA to estimate distributions of target, distractor, and congruency representations across the cortical hierarchy. The use of the RSA framework enabled us (1) to compare representation of each dimension in terms of neuroanatomical profile, (2) to graphically depict these representations in an intuitive, data-driven manner, and (3) to assess whether dimension representations are differentially associated with behavior. Because of the retrospective nature of this project, however, it is important to keep in mind that the experimental design was not optimized for RSA; thus, our results are subject to several limitations (see *Discussion* section *Limitations*). But, despite these limitations, our results demonstrate clear dissociations between representation of Stroop dimensions, and largely in predicted ways, suggesting that we were successful in measuring dissociable

neural processing of each task dimension. We interpret these results as providing strong initial support regarding the utility of pattern analysis methods of non-invasively obtained measurements of brain activity — in particular, RSA of fMRI — to enable stronger tests of neuroscientific theory of cognitive control.

Chapter 2

Method

Unless noted, all analyses were conducted in R, version 3.4.4 (R Core Team, 2018).

2.1 Availability of data and code

This study was conducted on data from the Dual Mechanisms of Cognitive Control project. Additional procedural details, illustrations, task scripts can be accessed via our project website[1] and Open Science Framework page[2]. RSA-level data (e.g., similarity matrices) and R scripts for all analyses, figures, and this manuscript will be made available on the first author's GitHub.[3]

[1] `https://pages.wustl.edu/dualmechanisms/tasks`
[2] `https://osf.io/xfe32/`
[3] `https://github.com/mcfreund/stroop-rsa`

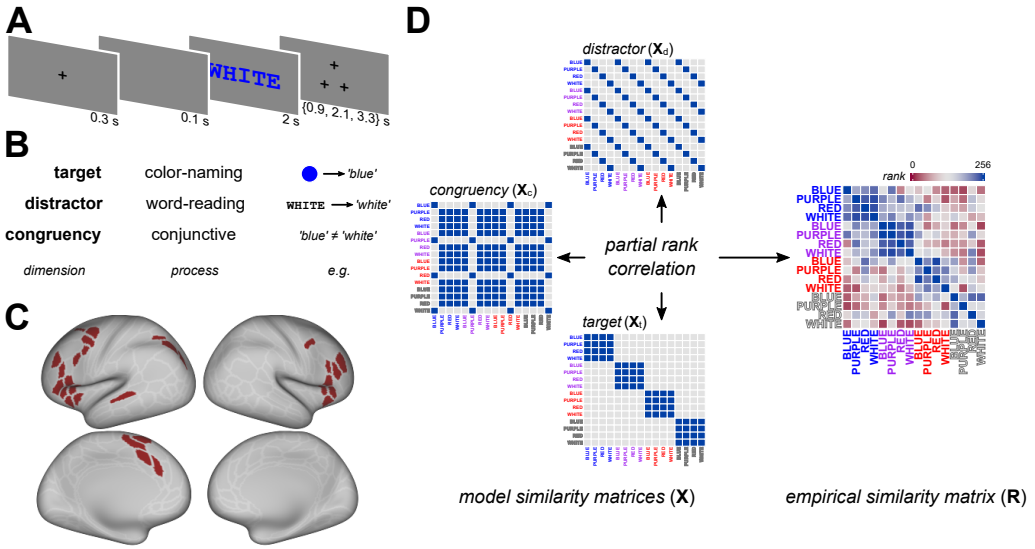

Figure 1: Schematic of task paradigm (**A**), conceptual (**B**) and analytic (**C**–**D**) framework. Participants performed a color-word Stroop task (**A**) while undergoing an fMRI scan. To decompose task-driven fMRI activity into three conceptual task dimensions of *target*, *distractor*, and *congruency* (**B**) — associated in turn with task-relevant *color-naming*, task-irrelevant *word-reading*, and higher-order *conjunctive* processes — a general linear model estimated the BOLD response evoked by sixteen unique Stroop stimuli (e.g., "WHITE" displayed in blue hue) independently for each voxel. The multi-modal atlas of Glasser et al. (2016) was then used to parcellate cortex (**C**, light silver borders), and within each parcel, linear correlations among response patterns from the sixteen stimuli were estimated to form an *empirical similarity matrix* (**D, right**). Through partial rank correlation, these matrices were fit to three representational models (**D, left**), which corresponded to the three hypothesized dimensions of the Stroop task (**B**). The resulting correlation statistics summarized the extent to which a parcel emphasized, within its distributed activity patterns, the representation of each unique task dimension. This framework may support inference regarding regional involvement in processing of a particular aspect of a task or the quality of an individual's particular task representations—both of which are obscured in univariate fMRI analysis of common cognitive control tasks such as Stroop.

2.2 Participants

At the time of analysis (February 2019), 78 individuals were recruited from the Washington University and surrounding St Louis metropolitan communities for participation in the Dual Mechanisms of Cognitive Control project. The present study began with a subset (N = 67) of these participants: those with a full set of imaging and behavioral data from the Stroop task during a particular scanning session (the "proactive" session), which we selected for methodological utility (for session selection reasoning, see *Method* section *Selection of data for analysis*). Of this subset, 1 individual was excluded for an atypically high rate

of response omission ($> 10\%$) and 17 for being genetically related as twins (by randomly selecting a set of unrelated co-twins). This left a final, unrelated sample of $N_{subj} = 49$, which we used in all analyses (i.e., *Results* sections *Representational similarity analysis* and *Brain–behavior correlations*), with the exception of the reduced-dimension plots in Figure 4. For this analysis, we included data from the held-out sample of co-twins, as this inclusion helped stabilize the observed configurations. (Note that the goal of this analysis was descriptive, and thus is not impacted by any assumptions associated with treating the participants as a random effect.)

2.3 Stimuli and procedures

We used a version of the standard color-word Stroop task (Stroop, 1935). Participants saw names of colors that were displayed in different hues and were instructed to "name the color, not read the word, as fast and accurately as possible".

2.3.1 Stimuli creation

The set of stimuli consisted of two subsets of color-word stimuli (randomly intermixed during the task): a *mostly incongruent* and an *unbiased* set. Each stimulus set was created by pairing four color words with four corresponding hues in a balanced factorial, forming 16 unique color-word stimuli within each set. The *mostly incongruent* group consisted of stimuli with hues (and corresponding words) "blue" (RGB = 0, 0, 255), "red" (255, 0, 0), "purple" (128, 0, 128), and "white" (255, 255, 255); the *unbiased* group, of "black" (0, 0, 0), "green" (0, 128, 0), "pink" (255, 105, 180), and "yellow" (255, 255, 0). These words were centrally

presented in uppercase 18-point, bold Courier New font on a grey background (RGB = 191, 191, 191). We focus our analysis solely on *mostly incongruent* stimuli (see *Method* section *Selection of data for analysis*), and thus do not describe the unbiased set further.

2.3.2 Task parameters.

Fixation and color-word stimuli were displayed in 18-point, bold Courier New. Each trial (e.g., Figure 1A) began with a central fixation cross, presented for 300 ms on a grey background (RGB = 191, 191, 191). The color-word stimulus, preceded by a blank screen following fixation offset (100 ms), was centrally presented for a duration of 2000 ms, fixed across trials. The duration of the inter-trial interval (triangle of fixation crosses) was either 900, 2100, or 3300 ms, selected randomly. These trials were organized into three blocks of 36, between which a fixation cross appeared for 30 s, forming a mixed block-event design (Chawla, Rees, & Friston, 1999; e.g., Dosenbach et al., 2006). Each of the 16 *mostly incongruent* stimuli were presented in both runs. However, of the 16 *unbiased* stimuli, 6 were presented in only in the first run and 6 in the second. Within each run for each participant, *mostly incongruent* stimuli were presented an equal number of times within each block. Within each block, stimulus order was randomized.

2.3.3 Hardware and software for task display and behavioral data collection.

The experiment was programmed in EPrime 2.0 ("E-Prime," 2016), ran through a Windows 7 Desktop, and displayed through a projector. Verbal responses were recorded for offline

transcription and response-time (RT) estimation. The first number of participants spoke into a standard MR-compatible electronic microphone; due to mechanical failure, however, we replaced this microphone with the noise-cancelling FOMRI III, which the subsequent participants used. A voice-onset processing script (from the MATLAB Audio Analysis Library) was used to derive response time estimates on each trial.

2.4 Selection of data for analyses

We focused our representational similarity analysis (RSA) solely on trials from the *mostly incongruent* stimulus group within the "proactive" scanning session of our Stroop task for methodological reasons: this was the only stimulus group and scanning session in our larger Dual Mechanisms project in which each unique Stroop stimulus (e.g., "BLUE" displayed in blue hue) was presented an equal number of times (9) to each participant.[4] These sixteen unique stimuli constituted the "conditions" for the RSA by forming the columns (and rows) of the similarity matrices (Figure 1D). By selecting this session and stimulus group for analysis, we ensured that any pattern differences observed between the stimulus conditions were not due to differences in the number of trials that contributed towards pattern estimations — a factor which strongly impacts the reliability of multi-voxel activation pattern estimates (Dimsdale-Zucker & Ranganath, 2018) — and without having to resort to under-sampling, which reduces the precision of estimates.

[4]In contrast, for example, we presented each *unbiased* stimulus three times more often if it was congruent (9) versus incongruent (3) within the "proactive" session. This trial frequency manipulation was performed to investigate questions outside the scope of the current analysis (see, e.g., Gonthier, Braver, & Bugg, 2016 for a similar manipulation).

2.5 Image acquisition, preprocessing, and mass-univariate general linear model

The fMRI data that used in these analyses were acquired with a 3T Siemens Prisma (32 channel head-coil; CMRR multi-band sequence, factor = 4; 2.4 mm isotropic voxel, with 1200 ms TR), and subjected to the minimally pre-processed functional pipeline of the Human Connectome Project, outlined in Glasser et al. (2013). After pre-processing, to estimate activation patterns, we fit a mass-univariate general linear model (GLM) to blood-oxygen-level dependent (BOLD) timecourses via a mixed block-event design in AFNI, version 17.0.00 (Cox, 1996). We convolved with a hemodynamic response function 16 boxcar regressors, each coding for the initial second of presentation of a *mostly incongruent* stimulus that prompted a correct response [via AFNI's `BLOCK(1,1)`]. We also included (1) two regressors [similarly created via `BLOCK(1,1)`] to capture trial-driven BOLD signal variance associated with congruent and incongruent stimuli of non-interest (*unbiased*) that prompted correct responses, (2) an "error regressor" coding for any trial in which a response was incorrect or omitted, (3) a sustained regressor coding for task versus rest (via `BLOCK`), (4) a transient regressor coding for task-block onsets [as a set of 7 finite impulse response functions [via `TENTzero(0,16.8,8)`], (5) 6 orthogonal motion regressors, (7) 5 polynomial detrending regressors (order automatically set) for each run, and (8) an intercept for each run. These models were created via `3dDeconvolve` and solved via `3dDeconvolve`. The data for each subject's model consisted of 2 runs $\times$ 3 blocks $\times$ 36 trials per subject (144 from the *mostly incongruent* stimulus group, 72 from *unbiased*). Frames with FD > 0.9 were censored.

2.6 Representational similarity analysis

RSA consists of three steps — similarity estimation, model fitting, and model evaluation — and our procedures generally followed the originally recommended methods for each step (Nili et al., 2014). To parcellate cortex, however, instead of using a data-driven searchlight analysis, we used a combination of an atlas-based parcellation scheme and an independent univariate region-of-interest (ROI) analysis. This atlas-based ROI approach enabled us to a conduct whole-cortex analysis that was not subject to known limitations associated with searchlights (Etzel, Zacks, & Braver, 2013), while maintaining a suitable level of power, particularly within regions sensitive to control demand.

2.6.1 Similarity estimation and atlas selection.

To estimate our empirical similarity matrices, beta coefficient images of the $N_{stimuli} = 16$ mostly incongruent stimuli were first extracted from the GLMs. We next used a volumetric version (in MNI) of the Human Connectome Project's Glasser Parcellation (Glasser et al., 2016) to divide each image into $N_{parcel} = 360$ parcels tiling the brain (Figure 1C, light silver borders). The Glasser parcellation is useful as it is whole-cortex and constructed from multimodal sources (resting-state functional connectivity, myelin density, cortical thickness estimates, and task fMRI activations). Further, explicit links have been drawn between several Glasser parcels and areas defined within the broader neuroanatomical literature (see supplementary material in Glasser et al., 2016), in addition to the hypothetical "multiple demand" network implicated in functional neuroimaging (Assem, Glasser, Essen, & Duncan, 2019). Finally, using each parcel's stimulus activation patterns, we estimated the across-voxel linear correlation between each of the pairwise combinations of stimulus conditions, and

collated these correlations into an $N_{stimuli} \times N_{stimuli}$ empirical similarity matrix, $\mathbf{R}$ (Figure 1D, right).

2.6.2 Representational model fitting.

To decode task information from these correlation matrices, we first built three models, each corresponding to one of the three Stroop dimensions. We formulated each of these models as an $N_{stimuli} \times N_{stimuli}$ correlation matrix $\mathbf{X}$, indexed by $\mathbf{X}(i, j)$, that took only binary values (Figure 1D). These models make different predictions regarding the similarity structure of a region's measured activity patterns. The *target* model ($\mathbf{X}_t$) predicts that the region will show a unique pattern of activity for each stimulus hue (or equivalently, correct response), such that the correlation $\mathbf{X}_t(i, j)$ is equal to one when the two stimuli have the same hue (e.g., "BLUE" and "GREEN" in red hue). Similarly, the *distractor* model ($\mathbf{X}_d$) predicts a region's activity patterns will cluster purely by the status of the stimulus word (i.e., $r = 1$ if the two stimuli have the same word, e.g., "BLUE" in red and green hues, 0 otherwise). Finally, in the *congruency* model ($\mathbf{X}_c$), activity patterns cluster purely by the congruency status of stimuli (1 if the two stimuli are both incongruent or congruent, 0 elsewhere). Thus, each of these models is categorical, essentially reflecting the similarity matrix $\mathbf{R}$ that would be obtained from a hypothetical area that responds to each level of a given dimension (e.g., to each hue of the target stimulus) with a unique and noiseless pattern.

To fit these models, we extracted the unique off-diagonal elements of each $\mathbf{X}$ and of $\mathbf{R}$, which we denote as vectors $\mathbf{x}$ and $\mathbf{r}$, and estimated the partial rank correlations between them. The partial correlation captures the unique association between $\mathbf{r}$ and a model (e.g., $\mathbf{x}_t$), that remains after removing the variance component that each vector shares with the

other two models ($\mathbf{x}_c$ and $\mathbf{x}_d$). Partial correlation was advisable here, as our model vectors were not orthogonal (i.e., the correlation between $\mathbf{x}_t$ and $\mathbf{x}_d$ is $r_{t,d} = -0.25$, $r_{t,c} = -0.10$, and $r_{d,c} = -0.10$). We opted for rank correlation to provide robustness against univariate outliers and to keep with RSA convention (e.g., Diedrichsen & Kriegeskorte, 2017; Nili et al., 2014).

In an additional step — prior to the model fitting described above — we removed a specific nuisance component from the empirical similarity vector of each parcel through a rank regression procedure. This component stemmed from the task design: though each *mostly incongruent* stimulus occurred an equal number of times throughout the course of a session, these stimuli were not fully balanced across the two scanning runs. Specifically, half of the stimuli were presented three times in run 1 versus six in 2, and vice versa for the other half of stimuli. As each scanning run contains a large amount of run-specific noise (Alink, Walther, Krugliak, Bosch, & Kriegeskorte, 2015; Henriksson, Khaligh-Razavi, Kay, & Kriegeskorte, 2015), this imbalance across runs could lead to a bias in the resulting correlation coefficients between stimulus activation patterns, in which similarity among patterns from stimuli that mostly occurred within the same run would be inflated. We formalized this component of bias as a model matrix $\mathbf{X}_{bias}$ (equal to 1 where the run in which stimulus i most frequently occurred = the run in which stimulus j most frequently occurred, 0 elsewhere). As $\mathbf{x}_{bias}$ was correlated to our models of interest (albeit weakly, at $r_{bias,t} = 0.03$, $r_{bias,d} = -0.13$, and $r_{bias,c} = -0.05$), we removed this component from each parcel's similarity structure via ordinary least-squares regression (i.e., by regressing each parcel's rank-transformed $\mathbf{r}$ onto $\mathbf{x}_{bias}$ and subtracting this component from $\mathbf{r}$). The resulting bias-corrected vector of residuals thus formed the dependent variable in our RSA models (the $\mathbf{r}$ of the previous paragraph).

As a result of these procedures, we obtained three partial Spearman's correlation coefficients (ρ_t, ρ_d, ρ_c), each indexing the magnitude of coding of a unique feature of the task, per participant and parcel.

2.6.3 Region of interest definition

A subset of cortical ROIs were of a priori theoretical interest due to their putative involvement in cognitive control computations. These ROIs were identified based on a separate analysis of independent data from the Dual Mechanisms project (i.e., a univariate 'conjunction' analysis of the baseline session, involving all four tasks scanned in the project). The specifics of this conjunction analysis are beyond the scope of the current study, but it yielded a set of 29 cortical parcels. Because of the a priori identification and interest in these parcels, they were evaluated within the RSA through a separate p-value correction procedure (see Representational model evaluation). Notably, many of these ROIs are also highly consistent with prior cognitive control neuroimaging studies, including a recent conjunction analysis published using the Human Connectome Project data and the same cortical parcellation scheme: lateral (IFJp, p9-46v, i6-8) and medial (SCEF, a32pr) prefrontal cortex, anterior insula (AVI, FOP5), and intraparietal cortex (LIPd, IP1) regions (Assem et al., 2019).

2.6.4 Representational model evaluation.

We evaluated the fits of our RSA models in two ways. First, to assess whether a parcel's activity patterns carry any information about a given task dimension, we performed one-sided Wilcoxon sign-rank tests over participants (the default recommendation for inferential testing within an RSA framework; Nili et al., 2014), predicting that the distribution of

participants' model fits should be greater than zero. The resulting p-value was then adjusted, independently for each model, to maintain a consistent false-discovery rate (FDR; Benjamini & Hochberg, 1995) either over all 360 parcels, for non-ROI parcels, or over $N_{ROI} = 29$ parcels, for our ROIs (see *Region of interest definition*).[5] Parcels for which the distribution of model fits was greater than zero with an FDR-adjusted $p < 0.05$ we took for evidence of "task-dimension coding".

Then, to assess the relative strength of task representations, we used model comparison. among the parcels we found to code for a given task dimension, we test whether the given representation was stronger than the other two task-dimension representations by performing paired sign-rank tests (two-tailed) on the model fits. These three pairwise comparisons were FDR-corrected within each parcel.

To interpret and organize the results, we combined these two evaluation methods (testing against zero and pairwise comparisons), sorting parcels into different sets. Membership to each set was constrained by the representational preferences that parcels displayed across the three task dimensions — that is, by the *coding profile* of each parcel. The most inclusive of these sets were those that merely required significant coding of a given task dimension (i.e., greater than zero across participants at $p < 0.05$). We refer to parcels within these sets as *target*, *distractor*, or *congruency coding* parcels, and indicate the respective membership constraints with $\mathbf{t} > 0$, $\mathbf{d} > 0$, and $\mathbf{c} > 0$ (where $\mathbf{t}$, $\mathbf{d}$, and $\mathbf{c}$ denote the model-fit distribution across participants, and > 0 indicates the sign-ranked test hypothesis was supported at $\alpha = 0.05$). More stringently, we used the pairwise comparisons to create subsets of these *coding sets* with parcels that displayed preferential, or *selective coding*, of task dimensions. *Target-selective* parcels, for example, included all *target coding* parcels for which the target

[5]The results of the RSA are not strongly impacted by this correction: a highly consistent set of parcels would be obtained if this ROI approach had not been included.

representation was not only greater than zero, but greater than distractor and congruency representations, in addition to the latter two model fits not being significantly greater than zero (i.e., $\mathbf{t} > \{0, \mathbf{d}, \mathbf{c}\}$, and $\{\mathbf{c}, \mathbf{d}\} \not> 0$, where $\not>$ indicates the sign-rank test hypothesis was not supported at $\alpha = 0.05$].

2.7 Data-driven dimensionality reduction

To interpret the representations of an area, some form of dimensionality reduction is required. RSA model-fitting (above) can be considered a form of hypothesis-driven dimensionality reduction, in which the similarity structure of high-dimensional activity patterns is summarized along three particular axes that correspond to the hypothesized dimensions of Stroop. Complementary to this approach is data-driven dimensionality reduction: instead of projecting to a low-dimensional space defined *a priori*, data-driven methods typically seek a space that optimizes some overall criterion of fit. This enables a "hypothesis-free" examination of an area's representations, which may reveal aspects of an area's representational structure that are lost within the hypothesis-based space.

To accomplish this, we used non-metric multidimensional scaling (NMDS; Kruskal, 1964). NMDS is a flexible non-parametric technique that operates on dissimilarities, rather than similarities (e.g., $1 - r$). In NDMS, a lower-dimensional embedding, termed *configuration*, is found through iterative procedures that seek to minimize *stress*, a goodness-of-fit measure of the configuration. Stress refers to the error obtained from a monotonic regression of all interpoint Euclidean distances within the estimated configuration (d_{ij} between points i and j) onto the observed (high-dimensional) dissimilarities: specifically, the proportion of unexplained root sum-of-squared deviations (i.e., $\textit{stress} = \frac{\sqrt{\sum_{i<j}(d_{ij} - \hat{d}_{ij})^2}}{\sqrt{\sum_{i<j} d_{ij}^2}}$, where $\hat{d}_{ij}$ represents the

fitted values[6]). In other words, in the low-dimensional configuration produced by NMDS, as one moves from the closest to the furthest pairs of points, one is also moving, with some degree of error, from the most similar to the most dissimilar pairs of high-dimensional patterns. This degree of error is given by the stress of the solution. Thus, NMDS is ideal for our purposes: it assumes only that the rank order of pattern similarities is meaningful (i.e., that they are not embedded within a metric space), while producing an optimal low-dimensional representation within the intuitive Euclidean space.

We performed NMDS on five select parcels (Figure 4). These parcels were chosen, from each of the five groups of "representational profiles" highlighted in Figure 4, as the parcels demonstrating the largest effect sizes on the relevant Stroop dimension(s) within each group. [That is, if W_D is the Wilcoxon sign-rank statistic for coding of Stroop dimension D, the parcel chosen within the *task-relevant selective* group gave $\max(W_t)$; for the *task-relevant & congruency selective*, $\max(W_t+W_c)$; *distractor & target*, $\max(W_d+W_t)$; *distractor*, $\max(W_d)$; and *congruency*, $\max(W_c)$.] For each selected parcel, we estimated the mean similarity matrix across participants and subtracted these values from 1 to obtain dissimilarities. Before averaging, we applied Fisher's z-transform (inverse hyperbolic tangent), and after averaging, transformed back to r: $\bar{\mathbf{D}} = \mathbf{J} - \tanh(\sum_{s=1}^{N_{subj}} \operatorname{arctanh}(\mathbf{R}_s)/N_{subj})$, where $\mathbf{J}$ is an $N_{stimuli} \times N_{stimuli}$ matrix of all ones, and $\bar{\mathbf{D}}$ is the resulting mean dissimilarity matrix.[7] Each $\bar{\mathbf{D}}$ was submitted to an implementation of Kruskal's NMDS in R [`MASS::isoMDS()`] to generate a 2-dimensional configuration.

[6]This equation assumes that the dissimilarity matrix is symmetric with an all-zero diagonal, so that $\sum_{i<j} d_{ij}^2$, e.g., captures all unique entries.

[7]Similar to our RSA, we regressed from each subject's $\mathbf{R}$ the $\mathbf{X}_{bias}$ model prior to conducting this procedure.

2.8 Brain–behavior models

In a final analysis, we assessed whether our estimates of Stroop-dimension representation may have indexed behavioral processes that diverge across individuals. First, we estimated each participant's behavioral Stroop effect. Then, through bivariate correlation and multiple regression, we related these estimates to RSA model fits from several task-modulated brain regions.

2.8.1 Behavioral Stroop effect estimation.

We estimated each participant's Stroop effect in RTs and errors as a random slope parameter within mixed-effect models using the R packages `lme4` (Bates, Maechler, Bolker, & Walker, 2014) and `nlme` (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2018). This framework gave a straightforward way to benchmark the level of individual variability in the Stroop effect (see *Behavioral model specification and evaluation*), that is, relative to the level of unexplained variance in the response. Note that these mixed-effect estimates, though precision-weighted, were similar to those obtained from simple, independent linear contrasts ($r_{rt} = 0.94$, $r_{error} = 0.82$).

Error coding and exclusion criteria. We defined "errors" as any non-target color word spoken by a participant prior to the correct response (e.g., including distractor responses, but not disfluencies) or as a response omission. Error trials (137 commissions and 52 omissions of 10,548 trials) and all trials with RTs greater than 3000 ms or less than 250 ms (2) were excluded from the RT model. Additionally, the responses on some trials were unable to be transcribed due to poor recording quality (from, e.g., high scanner noise or poor enunciation);

these trials were coded as "unintelligible" and were excluded from both RT and error models (54). Further, to help stabilize RT estimates, we adopted an additional, relatively liberal (i.e., inclusive) criterion of excluding all trials for a given participant with RTs that deviated greater than 3.5 SDs from their correct-trial mean RT (94, range of exclusions per participant = [0, 4]). We validated this latter exclusion by fitting a separate model on data from each run, and estimating the change in cross-run reliability of the resulting participant-level Stroop effects: this trimming procedure increased the estimated split-half correlation from $r = 0.60$ to 0.69 ($\Delta r = 0.14$, bootstrapped 95% CI = [0.01, 0.14]; note that these contrasts were calculated after z-transformation). Thus, our RT and error models were fit, respectively, to a total of 10,176 and 10,530 datapoints, with ranges from [179, 216] and [178, 215] per participant.

Behavioral model specification and evaluation. The RT model assumed a Gaussian distribution (with identity link function) and the error model assumed binomial (with logit link). Both models were fit with a fixed effect for the congruency of a trial (congruent, incongruent) and a random effect of participant, with a random intercept, slope of congruency, and a covariance parameter. Additionally, our RT model estimated a participant-specific parameter by which their residual standard deviation was scaled. This additional estimation relaxed the assumption that each participant (level-II factor) has equal variance. To accomplish this, we fit the model in `nlme`, estimating a diagonal residual matrix [i.e., with `weights = varIdent(form = ~ 1 | participant)`]. Though this addition made the RT model significantly more complex (estimating $N_{subj} - 1$ more parameters), it was warranted: homogeneity of variance was clearly violated, as indicated by the vastly improved fit of the heterogeneous-variance model ($\Delta\text{BIC}_{full-red.} = -4309$, $\chi^2_{48} = 4752, p < 10^{-22}$). Importantly, this also increased the robustness of our model: using participant-specific rather than

uniform variance "re-claimed" some of the between-participant variance (level-II) as within-participant (level-I), increasing the shrinkage of the Stroop estimates (Supplementary Figure 7B; Pinheiro & Bates, 2000, p. 4.3.2, pp. 188–190). If anything, this additional shrinkage made our brain–behavior analysis more conservative, and had the added benefit of bringing all participants' Stroop effects positive (Supplementary Figure 7B), a property largely thought to be "universal" (Haaf & Rouder, 2017).

Once we obtained these estimates, we sought to establish that there was enough variability within to be plausibly explained: searching for moderators of individual differences in Stroop would be of limited validity when there are limited differences to moderate. To this end, we compared the "full" models, specified above, to reduced models that omitted the congruency variance parameter (and corresponding covariance). For RT data, the full model was preferred ($\Delta\text{BIC}_{full-red.} = 75.90$, $\chi^2_2 = 94.36$, $p < 10^{-20}$). For the error data, however, the reduced model was preferred ($\Delta\text{BIC}_{full-red.} = 17.78$, $\chi^2_2 = 0.74$, $p < 0.69$), indicating that the Stroop effect in errors was not measurably variable across individuals. Thus, we focus our brain–behavior models on RTs.

2.8.2 Selection and definition of regions for brain–behavior models.

To assess the functional significance of our RSA-derived task representations, we selected a set of seven cortical regions per hemisphere, plus one bilateral region (ventral somatomotor strip), that we expect are linked to variability in color-word Stroop task performance: (a) V1–V3, (b) ventral occipito-temporal, (c) intra-parietal sulcal, (d) ventral somato-motor

cortex, (e) mid-dorsolateral prefrontal, (f) inferior frontal, (g) frontal insular, and (h) dorsomedial prefrontal cortices. Corresponding to each of these regions, we defined a spatially contiguous cluster of Glasser parcels (a "super-parcel") in which at least one task-dimension representation was successfully decoded (i.e., with FDR-adjusted $p < 0.05$). For ventral primary motor cortex, however, we used the bilateral "'somato-motor–mouth"' community from the Gordon atlas (Gordon et al., 2016), as the Glasser atlas does not contain a parcel with exclusive coverage of this area. Table 6 contains the full list of parcels included within each super-parcel, and Supplementary Figure 9 depicts their surface locations. Next, for each super-parcel, we created a single mask and used the activity patterns across the entire super-parcel to re-estimate our RSA model fits, which were subsequently correlated with behavioral estimates (see *Brain–behavior model fitting and evaluation*).

Conducting our brain–behavior correlations at this "wider angled" level of analysis, versus at the level of individual parcels, enabled us to reduce both the number of statistical tests and the between-individual heterogeneity in coverage of a given functional area, with a corresponding loss, of course, in the spatial precision of our inferences.

2.8.3 Brain–behavior model fitting and evaluation.

We assessed the relationship between an individual's strength of task-dimension coding and their behavioral Stroop effect in two ways: first, via robust bivariate correlations, then, through multiple linear regression.

First, for each of the fifteen super-parcels, (7 unilateral areas $\times$ 2 hemispheres + 1 bilateral area) we assessed the correlation between each of the three RSA model fits (ρ_t, ρ_d, ρ_c) and the Stroop effect estimate across participants. For each of these 45 relationships, we

estimated both Pearson's r and Spearman's ρ with a trimmed correlation procedure, in which multivariate outliers were identified and excluded through a projection technique similar to the Stahel-Donoho measure of "outlyingness" (e.g., Maronna & Yohai, 1995), with code adapted from the development version of the `WRS2` package (Mair & Wilcox, 2018). No bivariate associations reported in the text, however, contained outlying values. For inference, we used 95% confidence intervals, estimated through bootstrap resampling (10,000 replicates). This method provided robustness to heteroskedasticity (Wilcox, Rousselet, & Pernet, 2018). Corrections for multiple comparisons were not conducted.

Next, we further characterized seven of these task-dimension correlations (see *Results*, *Brain–behavior correlations* for variable inclusion criteria) with a model selection procedure using ordinary least-squares regression. This model selection procedure enabled us to find a set of explanatory variables (regional task-dimension representations) that parsimoniously accounted for unique variance in the behavioral Stroop effect. Specifically, we fit all 127 unique combinations of these regressors and calculated three fit statistics for each model: Akike Information Criterion (AIC), Bayseian Information Criterion (BIC), and error obtained from a leave-one-out cross-validation scheme (LOO error). To obtain LOO-error, we first fit a given model on all possible "training" sets of $N_{participant} - 1$, then reconstructed a Stroop effect with the regressors of each held-out participant. These reconstructed Stroop effects formed a vector, $\hat{\mathbf{y}}$. LOO error was then given by $1 - r_{\hat{\mathbf{y}},\mathbf{y}}$, where $r_{\hat{\mathbf{y}},\mathbf{y}}$ is the Pearson correlation between the reconstructed and observed Stroop effects. Models that minimized these criteria are considered in the results.

Chapter 3

Results

3.1 Representational similarity analysis

In a hypothesis-driven representational similarity analysis (RSA), we first fit three models (Figure 1C) to the correlation structure of each Glasser parcel's (Figure 1B) activity patterns. These models enabled us to isolate and estimate regional coding, then compare the cortical distributions of three conceptual dimensions of information within the Stroop task: *target*, *distractor*, and *congruency*. (Figure 2 displays the resulting statistical maps, thresholded at an FDR-corrected $p < 0.05$; Tables 3–4 contain statistical results relevant to these analyses.) Examining the representational profiles of select areas (Figure 3) suggested the target, distractor, and congruency-coding sets of parcels may be amenable to further decomposition through within-parcel model comparison. Finally, these within-parcel model comparisons (Figure 3, Supplementary Figure 8, Figure 4 center, Tables 1–2), demonstrate neuroanatomical dissociations in task-dimension preferences.

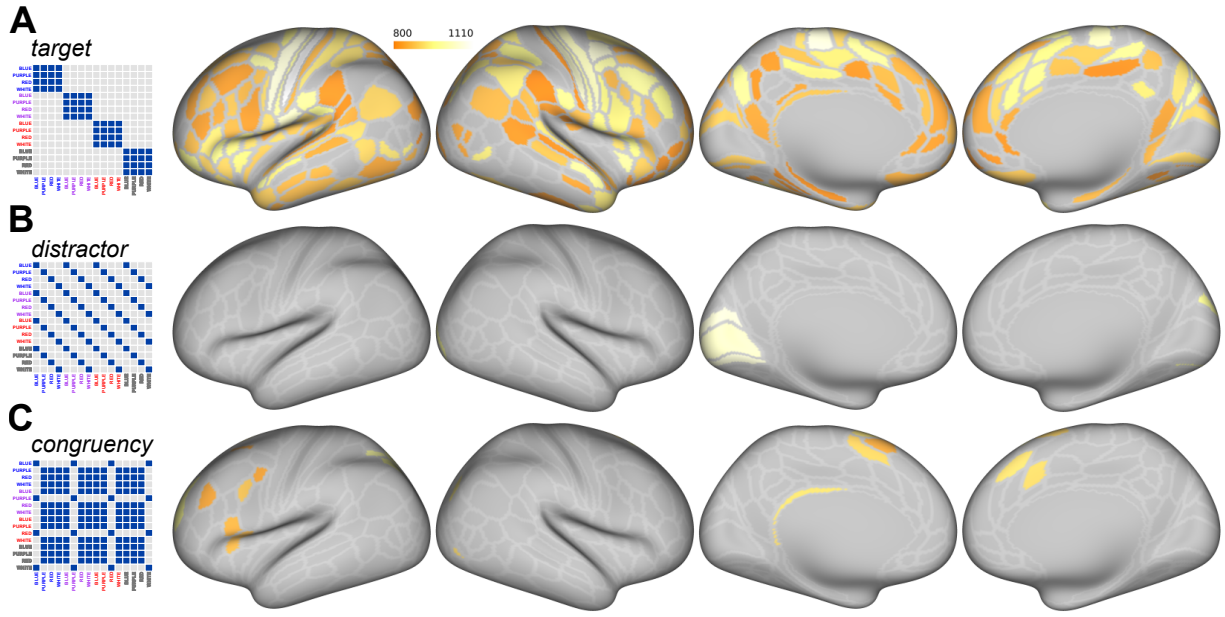

Figure 2: Representational similarity analysis of color-word Stroop. **A**–**C**, Parcels for which the *target* (**A**), *distractor* (**B**), or *congruency* (**C**) model statistics were significantly greater than zero over $N_{subj} = 49$ participants after FDR correction. The hue indicates the value of the test statistic (the sum of signed ranks). Each row (**A**–**C**) is plotted with the same color scale. **A**, The *target* model (*left*) is correlated with parcel representations across cortex. **B**, The distractor model (*left*) exclusively captures representations in early-to-mid visual cortex (V1–V4). **C**, The congruency model (*left*) is correlated with lateral and medial prefrontal, intraparietal, and insular parcels, in addition to left retrosplenial and right lateral occipital cortex.

Target representations were found widely across cortex.

A majority of parcels (236/360) had representations that were correlated with our *target* model (Figure 2A). This set of parcels tiled most of the frontal, insular, superior parietal, lateral and ventral temporal cortices, without a strong overall preference for hemisphere ($N_{left} = 122$, $N_{right} = 114$).

For further examination, we subdivide this collection of "target coding" parcels into two overlapping sets: *target-selective* parcels, and *task-relevant–selective* parcels.

'Target selective' parcels. We defined "target selective" parcels as those for which (1) the target model fits were greater than zero (i.e., $\mathbf{t} > 0$ via one-tailed sign-rank; Figure

1A), (2) whereas congruency and distractor fits were not ($\{\mathbf{c}, \mathbf{d}\} \not> 0$), and (3) target model fits were greater than *both* the latter two models ($\mathbf{t} > \{\mathbf{c}, \mathbf{d}\}$ via two-tailed paired sign-rank). This revealed a reduced set of 19 parcels (Supplementary Figure 8), which included parcels in bilateral somatomotor strip (parcel–*hemisphere*: 3a–*l*, 3b–*l*, 4–*l*); bilateral superior temporal gyrus (STG), near the lateral sulcus (PBelt–*r*, A4–*r*) and more anterior (STGa–*l*); and bilateral inferior temporal cortex (IT; TE2a–*l*, TF–*r*, TGv–*r*). Additionally within this set were bilateral orbitofrontal parcels towards the frontal pole (OFC; 11l–*l*, 11l–*r*, a10p–*r*), a left rostral inferior parietal lobular parcel (IPL; PFop–*l*), and a right precuneal parcel (POS2–*r*).

'Task-relevant–selective' parcels. Influential theories of prefrontal cortex function hold that these regions orchestrate goal-directed behavior by emphasizing representation of task-relevant information (e.g., Miller & Cohen, 2001). Within our Stroop-dimension framework, it follows that these regions should encode the target dimension stronger than distractor. The prediction for the congruency dimension, however, is weaker: current-trial congruency is a higher-order property of Stroop, not an explicitly relevant or irrelevant feature. Our *target-selective* constraints may have therefore been a poor match for the representational profiles that control-related fronto-parietal parcels generally exhibited. Thus, we defined a *task-relevant* contrast ($\mathbf{t} > \{0, \mathbf{d}\}, \mathbf{d} \not> 0$) to identify which regions are preferentially selective for the relevant, versus irrelevant, dimension. Note that this *task-relevant–selective* set was constructed according to the same criteria as the *target-selective* (above), except that all constraints regarding coding of the congruency dimension were relaxed. This collection of parcels therefore contained all *target-selective* parcels, in addition to a wider set of 99 parcels that spanned bilateral lateral and medial PFC, frontal opercular, IPS, ventral visual, and posterior cingulate cortex (Figure 4 center, dark and light blue).

3.1.1 Distractor representations were found exclusively in V1–V3.

In striking contrast to the widespread distribution of target representations, the parcels that were measurably correlated with our distractor model were confined to early-to-mid visual cortex, from V1 to V3 (Figure 2B), most prominently in V1–*l*.

3.1.2 Congruency representations were found primarily in pre-frontal and intra-parietal parcels.

In a third neuroanatomical distribution, congruency information was successfully decoded from a set of 20 parcels that were mostly situated within the left frontal lobe and along left IPS (Figure 2C). Within left hemisphere, these frontal lobe parcels were located in inferior frontal junction and premotor cortex (IFJ-p, PEF), inferior frontal gyrus (44), insula (FOP4), mid-dlPFC (p9-46v), superior frontal gyrus (s6-8), lateral fronto-polar cortex (9-46d), and, relatively strongly, (pre-)supplementary motor regions (SCEF, $\rho = 0.09$; 6ma, $\rho = 0.08$). In right frontal lobe, only three parcels significantly represented congruency information: a dmPFC cluster (8BM, p32pr), and a superior frontal gyral parcel (s6-8). Numerically, the largest effect sizes, however were found in a cluster of left IPS parcels (IP1, $\rho = 0.13$; MIP, $\rho = 0.12$; LIPd, $\rho = 0.11$), in addition to a more inferior, right IPS parcel (IP0, $\rho = 0.09$). Two other parcels—the left retrosplenial complex (RSC–*l*), and a right lateral occipital parcel (LO2–*r*)—also measurably represented congruency information.

Notably, several of these congruency representations were coincident with task-relevant representations, and were also stronger than distractor representations (Figure 4 center, lime green). We refer to these as "task-relevant–congruency" selective parcels. In fact, only three

areas were "purely" selective for congruency, according to our criteria: left 9-46d (frontopolar), right 6ma (supplementary motor), and LO2 (lateral occipital; displayed in Figure 4 center, in pink).

3.1.3 Dissociations in select representational profiles.

While Figures 2 (A–C) and 4 (center) clearly demonstrate the neuroanatomical dissociations in encoding of these Stroop dimensions, to illustrate more clearly the variety of *representational preferences* across the cortical hierarchy, we selected four parcels from task-relevant regions — V1, primary motor (4), dorsal premotor (FEF), and ACC (p32pr). Figure 3 displays these preferences. In V1, distractor coding was numerically stronger than target ($\Delta\rho = 0.03$, $p = 0.07$), whereas in area 4, target coding predominated ($\Delta\rho = -0.07$, $p = 0$). In dorsolateral versus medial frontal cortex (FEF, p32pr), dissociations between congruency and target coding emerged: a preference for the target versus congruency dimension was stronger in FEF versus ACC ($\beta_{region \times dimension} = 0.07$, $p = 0.09$; pairwise comparisons not significant).

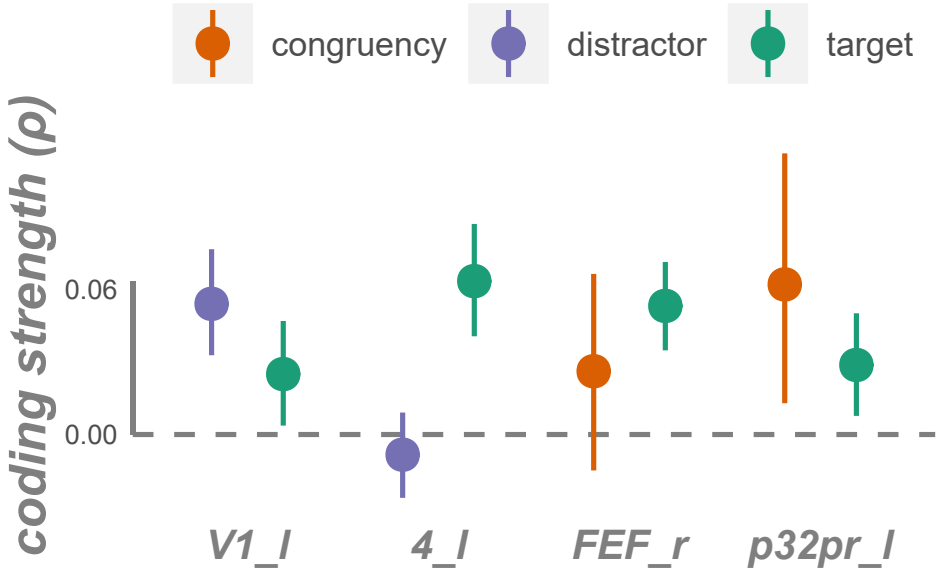

Figure 3: Representational preferences of select sensory, motor, and control-related areas. Circles are centered on the mean of participants' RSA model fits (for a given model and parcel); error bars span 95% confidence intervals bootstrapped from these samples. Left V1 (V1_l); left primary motor cortex (4_l); right dorsal premotor cortex (FEF_r); dorsomedial prefrontal cortex (p32pr_l).

3.2 Data-driven dimensionality reduction

Next, we conducted a data-driven analysis using non-metric multidimensional scaling (MDS), a non-parametric dimensionality reduction technique, on activity patterns from an exemplary group of task-dimension coding parcels (Figure 4, surround). Whereas the hypothesis-driven RSA sought to summarize high-dimensional fMRI activity patterns along particular dimensions in activity space that corresponded to dimensions of the Stroop task (e.g., Figure 3), this MDS analysis aims to summarize these patterns in low-dimensional configurations, or "representational geometries", that best capture the "full" representational structure of an area, without assuming what that structure may be. This approach is also complementary

to RSA, as it provides a compact visual representation of the task-dimension coding results and a means for hypothesis generation.

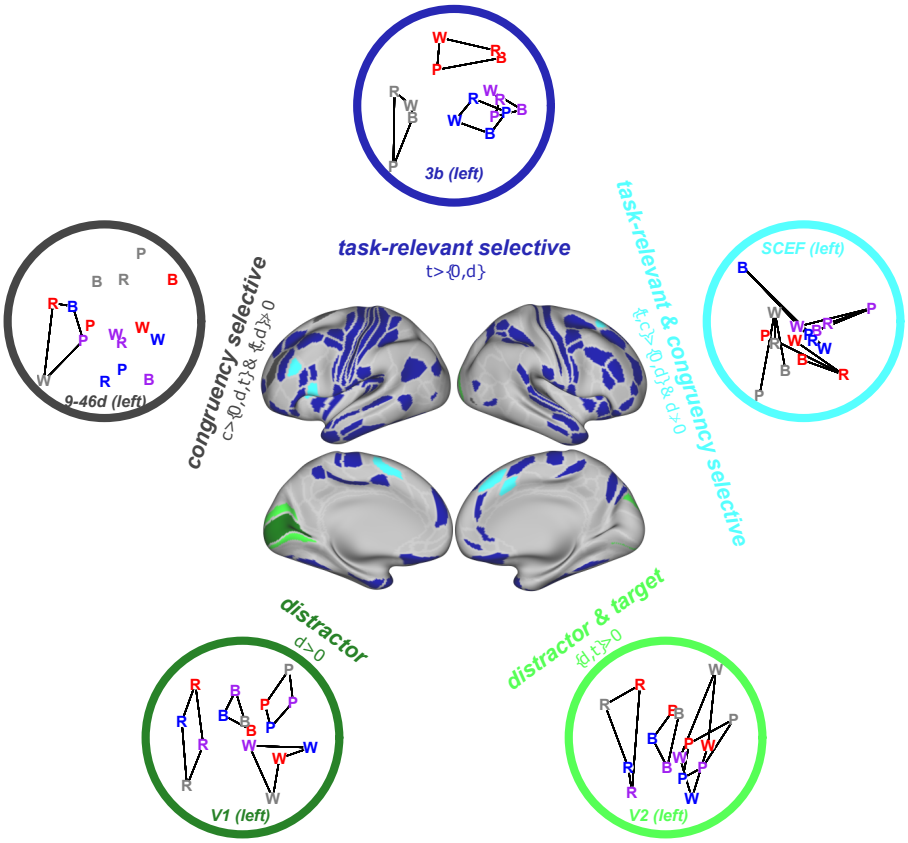

Figure 4: The spatial distribution and exemplary geometry of five "types" of areal representational profiles. **Center**, A "conjunction" map in which multiple representational profiles are overlaid. Dark blue indicates the parcels that were defined as being *task-relevant–selective* — that is, with target representations both stronger than zero and stronger than distractor coding (cf., the more expansive set of "target coding" parcels in Figure 2A, and the more restrictive set of "target selective" parcels in Supplementary Figure 8). These task-relevant–selective parcels encompass a broad set of somatomotor, inferior and superior temporal, intraparietal, and lateral, insular and anterior PFC areas. Notably, several of these parcels within left lPFC and bilateral mPFC were simultaneously selective for congruency (light blue). The distribution of these two representational profiles contrasted to that of distractor coding parcels, which were exclusively located in early-to-mid visual cortex (V1–V3). This distractor information, however, was not selectively represented by these visual areas: V1 distractor representations (dark green) were not stronger than congruency nor target, and in extrastriate cortex, activity patterns reflected a combination of distractor and target representations (light green). Finally, three parcels — areas within left fronto-polar, left supplementary motor, and right lateral occipital cortices — contained neither target nor distractor representations, but were instead selective for congruency (dark grey). **Surround**, The representational geometry of parcels that most strongly represented a task dimension in isolation (3b–*l*, top; V1–*l*, bottom left; 9-46d–*l*, top left) or in conjunction with another task dimension (IP0–*l*, top right; V2–*l*, bottom right), displayed via non-metric multidimensional scaling(Kruskal, 1964). Within each plot, distances between colored letters (i.e., color-word Stroop stimuli; white stimuli indicated by grey letters) represent the relative rank-ordering of similarities between higher-dimensional activity patterns within each region. Though connecting lines are arbitrarily imposed, they highlight the task-dimension structure within each parcel. Respectively, the MDS solutions for 3b–*l* (left somatosensory cortex), V1–*l*, and 9-46d–*l* (left fronto-polar cortex), clearly show pattern clustering by target, distractor, and congruency task dimensions. The solution for V2-*l* similarly reflects the hypothesis-driven analysis, roughly depicting simultaneous discrimination of distractor and target status along the horizontal and vertical axes, respectively. While the solution for SCEF–*l* [left (pre-) supplementary motor area] also demonstrates some degree of conformance to the target and congruency models, a feature of the representation not captured by the congruency model is the heightened dissimilarity among congruent stimulus patterns relative to among incongruent. This can be seen in the central contraction of incongruents and the peripheral expansion of congruents, suggesting this region responded in a stereotyped manner, specifically during incongruent trials.

3.2.1 MDS reveals a range of representational structures

The peripheral plots in Figure 4 display MDS configurations from five exemplary parcels. These parcels represented a given task dimension (or conjunction of task dimensions) most strongly within each of the five "types" of representational profiles displayed. Thus, we are by definition focusing on the representations that best conform to our hypothesized task-dimension representations. While we are likely missing much unpredicted and potentially interesting results, by restricting our scope in this way, we focus on interpreting the MDS configurations that are the most interpretable: those that are the least likely to be driven by task-independent noise, and those with features that map relatively well to our specified task dimensions. Further, as these analyses are not for inference (but rather for hypothesis generation), all participants, including twins and non-twins ($N = 67$) are retained, to help stabilize the geometries.

The differences between the geometries of parcels coding for each task dimension are clearly illustrated by the MDS solutions (Figure 4, surround). The solutions from regions displaying coding of only one dimension — target coding in left somatomotor cortex (Figure 4, top, dark blue), distractor coding in left V1 (bottom left, dark green), and congruency coding in left fronto-polar cortex (top left, dark grey) — display notable separation of points according to the respective task dimension (connecting lines are arbitrary, but highlight this separation). Similarly, in left V2, a region that displayed both target and distractor representations, the horizontal and vertical dimensions learned by MDS approximately map to the dimensions along which stimulus-evoked patterns are best discriminated (within the low-dimensional solution), respectively, on distractor and target status. And, in left dmPFC (SMA–pre-SMA; top right, light blue), incongruent patterns are located towards the origin (center),

while congruent patterns diverge in the periphery, and these patterns are superimposed over an approximate clustering by target.

3.2.2 'Incongruency', rather than congruency, coding marks the geometry of several parcels.

The MDS configuration of SCEF patterns illustrates a potential deficiency of the RSA congruency model. Effectively, this model proposes a representational geometry of two clusters of patterns, corresponding to incongruent and congruent stimuli. What is instead observed in SCEF is one central cluster of incongruent-evoked patterns, and four divergent congruent-evoked patterns. This suggests that SCEF may have responded in a stereotypical way to incongruent, but in divergent ways to congruent stimuli. To test this hypothesis, we built an "incongruency" coding model, which would be observed by a region that responds with a common pattern only on incongruent trials (i.e., $r = 1$ if stimulus i and j are both incongruent, 0 otherwise). Indeed, this "incongruency" model fit the SCEF geometry better than our original (congruency) model ($\Delta\rho = 0.02$, $p = 0.04$), in addition to the geometry of several other congruency-coding parcels within dmPFC and IPS (Table 5).

3.3 Brain–behavior correlations

In a final analysis, we attempted to link explicitly these RSA-derived indices of regional task-dimension representations to behavioral performance. By relating individual-level variability in the magnitude of a given task-dimension index to variability in the size of the Stroop effect, we test several hypotheses regarding the nature of the information carried by these indices.

Briefly, we defined a collection of 15 anatomical regions — 7 sets of Glasser parcels per hemisphere, and one bilateral Gordon community (Somato-Motor–mouth) — based on *a priori* evidence, refined each set to include only a single contiguous cluster of task-modulated parcels (a "super-parcel"), then re-estimated RSA indices using the entirety of each super-cluster. We then correlated each index with the behavioral Stroop effect across participants. Table 7 displays the strongest 25 of these bivariate correlations.

3.3.1 The strength of regional task representations explains individual variability in the Stroop effect in predicted ways.

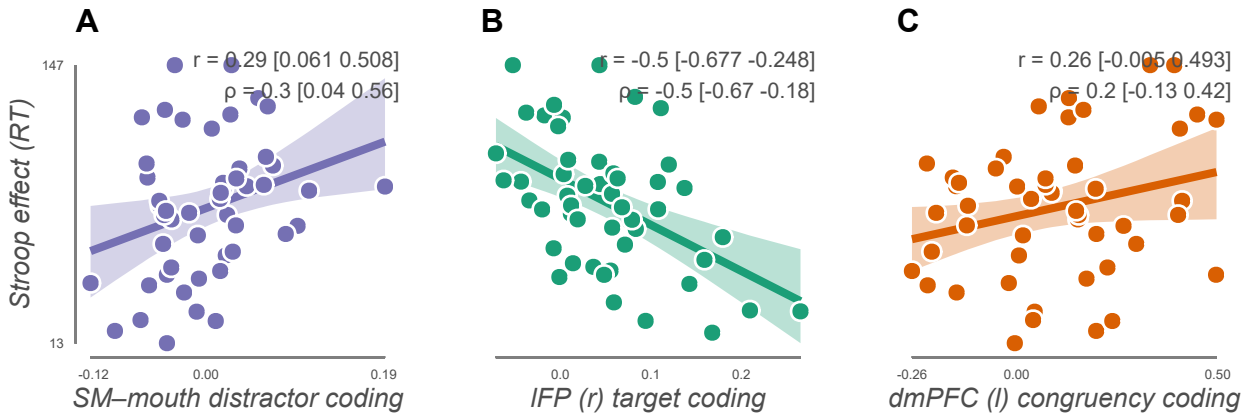

Figure 5: Dissociations in functional relevance of Stroop dimension coding. For clarity, we plot only the relevant dimension for each region (i.e., distractor coding for SM–mouth, target coding for lFP, congruency coding for dmPFC). Bivariate correlations from the omitted relationships were non-significant: $r_{SM,targ.} = -0.11$, $r_{dmPFC,targ.} = 0.12$, $r_{lFP,dist.} = -0.12$, $r_{lFP,cong.} = -0.02$. To test for double dissociations, however, we fit two 3-way interaction models, each with terms for coding strength (ρ), region, and dimension: one model for target and congruency coding in lFP versus dmPFC, a second for target versus distractor coding in lFP versus SM–mouth. Results from these models are reported in the text.

Four of these correlations were from regions and in directions that we predicted. Most prominently, stronger target representations in right IPS and left dlPFC were associated with an attenuation of Stroop interference (Supplementary Figure 10), supporting the notion that task-relevant information in these regions is a key locus of control function (Kane & Engle,

2002; Miller & Cohen, 2001). In contrast to target coding in lPF, in bilateral ventral somatomotor cortex, we found that stronger distractor representation was associated with *larger* Stroop effects (Figure 5A) This positive correlation suggests that the relationship between fronto-parietal target coding and interference resolution cannot be explained by a general factor of "decodability". Finally, participants with stronger congruency coding in left mPFC (supplementary-motor to anterior cingulate) tended to have larger Stroop effects (Figure 5C). Although this bivariate correlation is weak, it was predicted based on the assumption that the strength of an individual's dmPFC congruency coding reflects their reliance on reactive control processes — which, given the mostly incongruent list, is expected to be suboptimal for performance.

To test formally whether dlPFC and IPS target coding reflect dissociable functions from dmPFC congruency coding, we fit a single model on Stroop RTs with explanatory variables of coding strength (ρ), indicators for region (*dmPFC*, *IPS/dlPFC*) and dimension (*target*, *congruency*), and their interactions. As IPS and dlPFC coding readouts were relatively similar ($r_t = 0.57$, $r_c = 0.34$, $r_d = 0.33$), we averaged them for simplicity, forming a single lateral fronto-parietal estimate for each subject and dimension. (Note that this decision did not change the direction or significance of the effects.) As reflected in Figure 5 (B–C), the nature of the relationship between task coding and the Stroop effect depended both on region and dimension ($\beta_{\rho \times dimension \times region} = 238, t = 2.42, p = 0.02$). In other words, this positive and predicted correlation in dmPFC congruency coding forms a double dissociation with lateral fronto-parietal target coding.

Similarly, we tested for a double dissociation between lFP and SM–mouth in target versus distractor coding. While we did not find evidence for a double dissociation ($\beta_{\rho \times dimension \times region} =$

$-38, t = -0.28, p = -0.28$), a single dissociation was present within SM–mouth between target and distractor coding ($\beta_{\rho \times dimension} = -201$, $t = -2.18$, $p = 0.03$). This interaction was not detected within lFP ($\beta_{\rho \times dimension} = -163$, $t = -1.64$, $p = 0.11$) But, because this overall particular pattern of results (Figure 5) was predicted based on (a) established functional dissociations between medial and lateral PFC of reactive and proactive control, and (b) the intuitive hypothesis that distractor coding at the output level should be detrimental, these dissociations suggest that our representational models were sufficiently specific in indexing these functions.

3.3.2 Unpredicted relationships to behavior in task-involved areas.

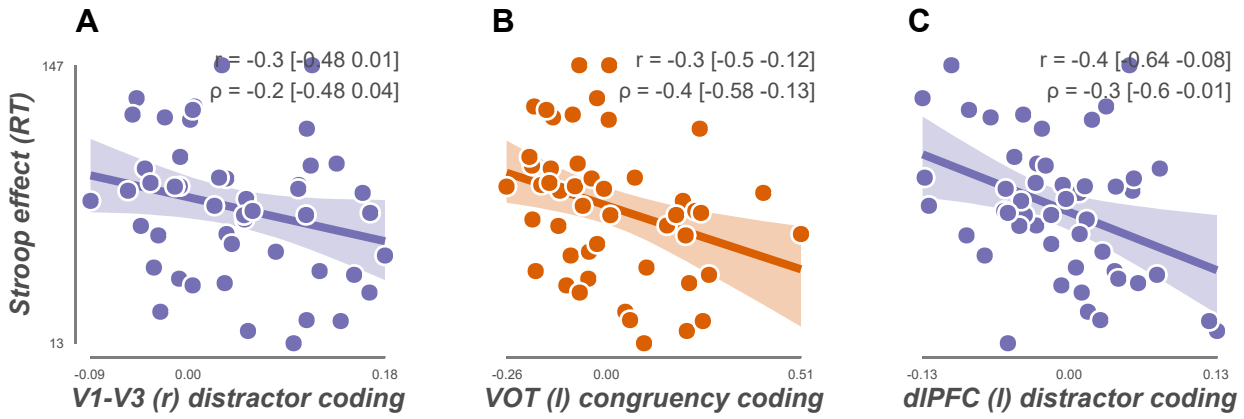

Figure 6: Unpredicted relationships between Stroop dimension coding and behavior. Ventral occipito-temporal cortex (VOT).

Three of the observed correlations, however, were unexpected, and one was directly contrapredicted. In early and extrastriate visual cortex, we found that the strength of distractor coding was weakly and negatively associated with the size of the Stroop effect (Figure 6A). In left ventral visual, by contrast, we found a moderately strong negative correlation between

congruency coding and Stroop (Figure 6B). Finally, against our predictions regarding the relative importance of task-relevant representations within fronto-parietal network, we found that larger distractor coding in left dlPFC was moderately associated with smaller Stroop effects (Figure 6C).

3.3.3 The variance explained by these representations is independent.

To explore whether these representational readouts predict independent variance in performance, we performed a simple model selection procedure. In brief, we fit linear regression models using all (127) combinations of these seven readouts as explanatory variables, and calculated three fit statistics for each model (BIC, AIC, and LOO error). These statistics agreed in accepting the best model (Supplementary Figure 11), which contained terms for right IPS target coding, left dlPFC distractor coding, left ventral visual congruency coding, and left dmPFC congruency coding (Table 8).

Notably, while the signs of each of these estimates matched their bivariate counterparts (cf., Table 7), the multiple regression revealed an interesting case of suppression: including ventral visual congruency coding in the same model as dmPFC congruency coding substantially increased the dmPFC estimate, to the extent that dmPFC congruency coding became (numerically) the strongest explanatory variable in the final selected model (Table 8, coefficients of partial determination). This suppression can also be clearly seen in the two-fold increase in R^2 that adding the mPFC term brings to a model with ventral visual congruency coding relative to the R^2 of a model with mPFC alone (0.15 versus 0.07).

Chapter 4

Discussion

We conducted a retrospective, and, to the best of our knowledge, novel, application of RSA (Kriegeskorte, 2008; Kriegeskorte & Kievit, 2013) to fMRI data obtained during the classic color-word Stroop task (Stroop, 1935). In combination with a multimodal atlas-based approach to parcellating cortex (Glasser et al., 2016), this framework enabled us to orthogonalize and estimate representation, within areas tiling the cortex, of three sources of task information in Stroop: the potentially conflicting target and distractor task dimensions, and their conjunction, or congruency. We found that these hypothesized dimensions — in predicted and specific ways — were dissociated in their mapping to cortex (Figures 2–4; Supplementary Figure 8), their representational structure (Figure 4, surround), and their association with individual differences in behavioral performance (Figure 5).

To avoid overinterpreting these results, it is important to acknowledge the many limitations that were associated with the particular experimental design used for analysis. These limitations largely stemmed from the retrospective nature of the analysis and thus could mostly be successfully addressed through a prospectively designed study. But, despite these limitations, the representational dissociations we observed meshes well with several decades-worth of neuroimaging research of Stroop and theoretic development in cognitive control —

suggesting that our simple RSA procedures were successful in specifically indexing component neural processes underlying performance in this task. This study therefore contributes an important "proof of concept" demonstration of the utility of the RSA framework for addressing key questions in cognitive control research.

In the sections that follow, we interpret the current results and their implications for understanding of cognitive control mechanisms in the Stroop task. Next, we discuss fruitful directions for extending the RSA framework to address a broader range of cognitive control questions and associated experimental paradigms. Lastly, we address the limitations of the study and analysis approach, and how these could be potentially remedied in a future prospective study.

4.1 Implications regarding Stroop mechanisms.

The motivation for the present work was based upon the assumption that dissociable processes of color naming, word reading, and conflict are represented in regional fMRI activity patterns and that these representations can be indexed in isolation through simple RSA. Importantly, we had strong a priori predictions for the how these component representations should be differentially distributed across the cortical hierarchy and how they should differentially associate with behavioral performance. To validate our motivating assumption, we tested whether these dissociations were borne out in our data. Our expectations were largely confirmed by these tests.

4.1.1 Neuroanatomical profiles.

Target coding. Our target coding model was built to capture representations associated with the task relevant process of color naming. These representations could relate to any aspect of the color-naming process, from early visual encoding of hue-related information, to the final neocortical output of articulatory representations in motor cortex. We sorted the regions coding for target into those that *exclusively* carried task-relevant representations (*target selective*), and those that preferred the task-relevant to irrelevant dimension (*task-relevant selective*).

Target-selective parcels were those that represented the target dimension more strongly than either distractor or congruency dimensions, and showed no evidence of representing either congruency or distractor dimensions (Supplementary Figure 8). Notably, several of these target-selective parcels were in regions associated with sensorimotor representations of target response information: somatomotor cortex (left 4, 3a, 3b, OP2-3, and, on the posterior bank of the postcentral gyrus, PFop) and auditory cortex (A4–r, PBelt–r). That is, because we focused our RSA solely on correct-response trials, we effectively constrained our scope to trials in which the motor output was target articulation, and therefore also the trials in which participants (could have only) heard themselves articulate the correct response.

We also notably observed target-selective coding in hetero-modal areas that are more deeply situated within processing pathways, such as bilateral inferotemporal cortex and superior temporal lobe areas. While posterior ventral temporal cortex was expected to carry target and distractor representations, finding that anterior IT regions, the culmination of the ventral visual stream, exhibited a target-selective profile is consistent with an interpretation that the target dimension was processed to a deeper extent than the distractor. Additionally,

several target selective parcels were situated within linguistically sensitive areas, such as left superior temporal gyrus (anterior) and sulcus. It is also worth noting that among the parcels just over the criteria for target selectivity was left inferior frontal gyrus (47–*l*, with $\rho_t = 0.03, p(|\mathbf{t}-\mathbf{c}| > 0) = 0.076$), a region traditionally associated with language production.

While unpredicted, there were also two other parcels within the target-selective group: a region within left orbitofrontal cortex and in right precuneus. We avoid speculating on the functions these regions may have played in this task, but note that recent studies using pattern analysis have similarly decoded task-relevant information from these regions (e.g., Schuck, Cai, Wilson, & Niv, 2016; Crittenden, Mitchell, & Duncan, 2015; Jackson & Woolgar, 2018), and that the precuneus has been identified as being activated during color-word Stroop by previous univariate analyses (Banich et al., 2001).

Task-relevant–selective parcels were those that represented the target dimension stronger than the distractor dimension (no constraints were placed on congruency coding; Figure 4, center). From the framework of top-down attentional modulation, we hypothesized that control-related lateral fronto-parietal regions would exhibit this representational profile (i.e., emphasizing target over distractor information). Although the resulting task-relevant selective group was much more expansive than the target-selective parcels, this hypothesis was supported: this set covered bilateral dorsal premotor, mid-dlPFC to IFG, insular, dMPFC, and IPS — all regions associated with fronto-parietal, cingulo-opercular and dorsal or ventral attentional control networks (e.g., Corbetta & Shulman, 2002; Dosenbach et al., 2006; Duncan, 2010; Tanji & Hoshi, 2008). Additionally, this group gave more extensive coverage of task-relevant areas such as ventral temporal and extrastriate cortex (V8, FFC, VVC) and STG.

Distractor coding. The distractor coding results stand in stark contrast to the target coding results: we found distractor representations exclusively within V1 to V3. While we expected a general cortical emphasis of target coding, finding no evidence of elevated distractor coding in left lateral occipito-temporal sulcus ("visual word form area") in particular was unexpected, as this region is an important sensory hub for reading (Dehaene, Cohen, Sigman, & Vinckier, 2005). In hindsight, however, our choice of atlas was inappropriate to localize this area (or any ventral visual area), as the parcel boundaries encompass heterogeneous areas within this region (Glasser et al., 2016). Another possibility is that participants were less reliant upon word reading processes as a result of the mostly incongruent nature of the list. Within mostly incongruent lists, the Stroop effect is greatly reduced. Behavioral evidence suggests that this reduction is, in part, mediated by a general attenuation of reading processes (Gonthier et al., 2016; Lindsay & Jacoby, 1994). It remains to be seen whether ventral visual distractor representations play a role in this list-level adaptation (e.g., as a locus of feature-based attentional suppression). One avenue future work could pursue is combining probabilistic atlases or functional localizers for left occipito-temporal cortex (e.g., Weiner et al., 2017) and proportion congruence manipulations to examine more precisely how intermediate sensory distractor representations are modulated by adaptive control.

Congruency coding. In a third and unique neuroanatomical profile, the higher-order congruency dimension was represented primarily within medial, lateral, and polar frontal cortex, in addition to intra-parietal cortex. The regions identified from this multivariate analysis were congruent with a substantial body of research demonstrating increased (univariate) activation in these regions: in particular, dmPFC (including SMA and pre-SMA), left mid-dlPFC, and left IFG (e.g., Cieslik et al., 2015; Nee et al., 2007). This profile also meshes with a recent report that used face–word Stroop paradigm, and decoded congruency information from lateral PFC and fronto-polar cortex (Jiang & Egner, 2014).

But, given the robustness of these prior univariate results, one might interpret the current congruency decoding effects as relatively weak, as we identified only a handful of dmPFC and dlPFC parcels that encoded the congruency status of trials. The paucity of congruency coding, however, is likely due to the mostly incongruent nature of the list, as the list-wide reductions in the behavioral Stroop effect are mirrored by brain activations (e.g., Carter et al., 2000; Wilk, Ezekiel, & Morton, 2012). Notably, a univariate analysis of these same data failed to identify any regions that responded with greater mean activation to conflict. This suggests that pattern analysis may be more sensitive than traditional univariate analyses, even for variables that are well-known to elicit robust increases in regional activity, such as "conflict". In extensions of the present study, it will be useful to explicitly compare univariate and RSA results to test this hypothesis directly.

We also found coexisting target and congruency representations in parcels within lateral and medial frontal cortex (Figure 4, center, light blue). In other words, these regions carried multidimensional representations, congruent with recent work demonstrating high-dimensionality of prefrontal representations (Fusi, Miller, & Rigotti, 2016; Rigotti et al., 2013). Notably, this highlights another advantage of pattern analysis methods, as demonstrating multidimensionality is challenging to establish with univariate methods.

4.1.2 Brain–behavior models.

One of the more robust bivariate correlations we found was negative, and between the strength of target coding in right IPS and individual differences in the magnitude of the Stroop effect (Supplementary Figure 10A). Target coding in this region was also moderately related to target coding in right dlPFC ($r = 0.57$), which was in turn related to performance

(Figure 10B), suggesting our target model indexed a process in which these regions were involved. Further suggesting this redundancy, in multiple regression model selection, only target coding in IPS was selected by the best model. These results suggest that task-relevant information encoded within these regions is an important mediator of Stroop interference.

Indeed, these regions are known to be tightly coupled to the implementation of top-down control (e.g., Buschman & Miller, 2007), via attentional sets (Corbetta & Shulman, 2002), task sets (Sakai, 2008), and in the representation of target information in a distractor-resistant format (e.g., Miller, Erickson, & Desimone, 1996; Jacob & Nieder, 2014; Qi, Elworthy, Lambert, & Constantinidis, 2014; Rademaker, Chunharas, & Serences, 2019). The observed brain–behavior correlations are also in line with the theoretical perspective that fidelity of target coding is a key locus of individual differences in cognitive control (e.g., Kane & Engle, 2002). The particular processes these relationships reflect, however, are less clear. With the present design, it is ambiguous whether target representations in IPS or dlPFC reflect stimulus or response-related information. Similarly, the downstream functional targets (e.g., IT, premotor cortex) of these representations are unclear. To shed light on these questions, future work could employ larger and more diverse stimulus sets, feature-based representational models (e.g., based on similarity in color space), and could examine inter-regional correlations between RSA model fits (e.g., to test "informational connectivity").

The observed brain–behavior correlations in target coding also formed one half of a predicted double dissociation with congruency coding in dmPFC. This dissociation rules out an alternative explanation of non-specific encoding (i.e., that in better-performing participants, any task variable would be more strongly encoded), and lends stronger support to our hypothesis that our representational models were specific in indexing reactive and proactive control processes. It is also notable that, relative to dmPFC, the other behavioral relationship with

congruency coding that we observed was more posterior in ventral visual cortex, and with opposite sign (negative). Further, within our model selection procedures, the suppression we observed between these two terms suggests that the dmPFC congruency response may contain two opposing components: one that indexes reactive control, another that is linked to the ventral visual response. Examining how brain–behavior correlations are modulated by control state (e.g., reactive, proactive) and contextual factors (e.g., list or item-level statistics) could shed light on these underlying functions.

We also observed a dissociation between distractor coding in sensory versus motor cortex: participants with a strong representation of distractor information in V1–V3, but weaker representations in somatomotor–mouth, had smaller Stroop effects. The SMMouth relationship was a clear prediction we derived from the assumption that participants with larger Stroop effects may be "closer" to articulating the stimulus word (e.g., sub-articulation). However, the early visual relationship was unpredicted. A speculative account is that stronger distractor representations may reflect better stimulus encoding. By making a further assumption that, at this relatively early level of vision, distractor (form) coding is somehow coupled to target (hue), this relationship could be explained. While tenuous, this assumption might not be implausible: hue and form information were spatially isomorphic in our task, thus distractor-correlated features such as word size may impact the strength of early target (hue) coding.

In contrast to these relationships, one association was directly contra-predicted: more precise distractor coding in left dlPFC was associated with smaller Stroop effects. The framework of top-down biased competition (e.g., Miller & Cohen, 2001) cannot account for this finding. Although we are unaware of an fMRI study that demonstrates distractor representations in dlPFC, neurons within this region do transiently encode distracting input in a variety of

control tasks (e.g., Jacob & Nieder, 2014; Mante, Sussillo, Shenoy, & Newsome, 2013). It is possible that this information is used in some form to guide subsequent-trial behavior. Future work could examine properties of these distractor representations and their correlation to behavior (e.g., timecourse, spatial distribution within dlPFC, sensitivity to feature-based models) to constrain plausible explanations. We discuss some of these potential directions in the next section.

4.2 Extending the RSA and multivariate framework to broader questions in cognitive control research

The general success of the present project in capturing dissociable representation of Stroop dimensions is highly encouraging for the utility of RSA and other multivariate techniques for addressing open questions in cognitive control. Here, we roughly sketch some selected examples of questions and methodologies that could shed light on them, involving not only Stroop, but also cued task-switching, and other cognitive control tasks.

4.2.1 Stroop.

The present study treated the Stroop effect as if it were a stationary phenomenon: a single behavioral readout was estimated per participant. In reality, however, the size of the Stroop effect depends greatly on the context in which the control system is embedded, that is, the statistics learned from trial history (Bugg & Crump, 2012). Learning and using these statistics to guide decisions is thought to be a central function of control systems centered over prefrontal and parietal cortices (e.g., Gold & Shadlen, 2007). Combining pattern analysis

with contextual manipulations might offer a useful window into how the brain mediates such adaptive control.

One well-researched modulation in the Stroop effect emerges from list-level manipulations of the proportion of congruent to incongruent stimuli (Gonthier et al., 2016; Logan & Zbrodoff, 1979). The learning mechanisms underlying this adaptation are relatively general (i.e., untethered to particular stimulus or response identities). But, the putative neural substrates and mechanisms mediating this process are less clear. A potentially fruitful approach to investigating these mechanisms is to extend the present design (estimating task-evoked target, distractor, and congruency representations) to lists with varying proportion congruency. This could enable within-subject tests of how these representations are modulated by list-level control across the cortical hierarchy.

Perhaps a more relevant analysis, however, would be in examining the time periods before trial onset. It has been hypothesized that list-wide adaptation effects are mediated by sustained and preparatory activity in dlPFC that reflects proactive coding of the task set. This mechanism of anticipatory task-set coding could account for the generality of this behavioral adaptation; however, a number of studies have failed to find sustained activity in dlPFC within mostly incongruent lists (Grandjean et al., 2012). However, this coding may not be evident in above-baseline activity, but could be present in a sub-threshold pattern — similar to notions of predictive coding in sensory systems (e.g., Kok, Mostert, & Lange, 2017), but for more abstract variables (i.e., congruency). Further, medial PFC may instead be involved in this prediction process, given its hypothesized role in using performance outcomes to generate expectations of abstract task-related variables (Alexander & Brown, 2011), and preliminary evidence that it may be activated within list-wide mostly incongruent

contexts (Wilk et al., 2012). In the Stroop task, these templates may take the form of "pre-activated" congruency representations (as this is the only information that can be predicted). Within the RSA or other multivariate decoding frameworks, this leads to a straightforward hypothesis: in mostly incongruent versus unbiased lists, are pre-trial activity patterns more similar to the conflict-evoked pattern?

Beyond list-wide modulations, research has demonstrated that the Stroop effect is modulated in a within-trial manner. Specifically, if stimulus locations or features (e.g., a blue hue) are made predictive of the congruency status of a trial (e.g., incongruent), individuals seem to take advantage of this information, as the Stroop effect is also reduced in these scenarios (Bugg, Jacoby, & Chanani, 2011). These item-specific proportion congruency effects (and others) are parsimoniously explained by an "event" or "episode file" framework of episodic memory and cognitive control (e.g., Egner, 2014; an extension of Hommel, 2004). Within this framework, representations of features of the task set (e.g., blue) are bound with co-occurring representations of internal control "settings" (e.g., decreased processing of distractor dimension) as an episodic trace. Presentation of any one feature leads to retrieval of the entire file, including re-instantiation of the associated control setting. This leads to a clear prediction that could be tested via multivariate methods and an appropriately designed Stroop task: within trials in which only one feature of a learned association is presented (a blue hue, without an accompanying distractor word), can the associated control condition (incongruent) be decoded from patterns of prefrontal activity (relative to when blue does not predict congruency)? In other words, when control is unnecessary , will there still be the obligatory retrieval and expression of the control state? Such a finding would provide the most direct neural evidence to-date of the existence of control-based event files.

4.2.2 Cued task-switching.

The cued-task switching paradigm is perhaps even more amenable to RSA decomposition than Stroop, as it orthogonalizes abstract task rules. This enables representational models to be simultaneously estimated at multiple levels of abstraction: from target or distractor stimuli or responses, to task rules, to effects of task-switching and task-rule congruency. Indeed, recent EEG studies have used this design to great effect (Hall-McMaster, Muhle-Karbe, Myers, & Stokes, 2019; Hubbard, Kikumoto, & Mayr, 2019) in tracing within-trial dynamics for each representational component. Within fMRI research, such a design could be useful for a variety of questions. For example, incorporating additional, 2nd or 3rd-order rules could enable novel tests of hierarchical theories of prefrontal cortex organization to be tested (Badre & Nee, 2018). Tracking how these representations are modulated by the effects of learning, or by performing the task under different instructions (e.g., to learn all S-R pairings, or to use a "hidden" task rule; Dreisbach & Haider, 2008) could inform the neural consequences of establishing a task set. Or, incorporating reward manipulation into the design (e.g., Hall-McMaster et al., 2019) could enable dissociating motivational from task representations, and examining their interaction. Findings from fMRI experiments could compliment the EEG work that has been conducted by localizing the representations to more focal anatomical areas.

4.2.3 Across task and across timepoint analyses.

RSA also lends itself to testing theories regarding the format of dorsolateral prefrontal representations. One influential theory, "adaptive coding", posits that dlPFC has a high degree

of representational flexibility (Duncan, 2001; Stokes et al., 2013). This theory places limited constraints on long-term encoding stability of particular variables (e.g., the patterns on cortex that are evoked by a certain task rule), emphasizing instead a flexible, context-dependent organization. But, a key prediction this perspective makes is that, though the code itself is labile, the information contained within is stable. For example, in a test–retest cued task-switching design, dlPFC is expected to encode the same task-relevant features at test and retest, however the patterns on cortex that contain this information are liable to change. From an RSA framework, the stability at these two levels of analysis could be easily tested by assessing the test–retest correlation between the spatial activity patterns (encoding stability), and between the *correlation matrices* derived from these patterns (informational stability) (see, e.g., Kriegeskorte & Diedrichsen, 2019).

In contrast, alternative "compositional" frameworks propose that dlPFC stably encodes certain abstract functions, or "task primitives", that are combined to perform various tasks. The degree to which two tasks load on common primitives dictates the degree to which they drive similar activation patterns. To test this hypothesis, RSA models could be designed to evaluate, for example, the intercorrelations among conflict-driven activation patterns from a battery control tasks. Tasks that evoke similar activation patterns should tap similar underlying processes – and thus, should elicit similar behavioral performance across subjects. In other words, the similarity structure of activation patterns should predict that of behavioral performances. Demonstrating an isomorphism between brain and behavioral structures would suggest a compositional architecture underlies dlPFC function within these tasks.

4.3 Limitations

4.3.1 Task-correlated noise.

A general concern for multivariate based neuroimaging analyses is the degree to which fMRI BOLD activation patterns are reflective of non-neural contributions. Relative to regional synaptic activity, the BOLD signal is strongly dependent on non-neural, "nuisance" factors such as motion and respiration. Traditionally, these sources of noise are removed in fMRI analyses via regression, under the assumption that their timecourses are uncorrelated, or weakly correlated, with models of the timecourse of regional brain activity. Yet when this assumption is not met, interpretational problems can arise. The current design may be particularly vulnerable to this concern because of the use of overt verbal responding for the Stroop task.

This feature of the task design increases the difficulty of interpreting the results of our target coding model. Participants likely moved their heads or exhaled while articulating in a way that was to some degree unique to each target response.[8] In turn, this response-specific motion could have induced particular patterns of variance in the BOLD signal across voxels, which would inflate correlations to our target coding model. Although we used standard motion regressors and performed scrubbing, it is likely that our procedures were not completely effective. Thus, it is likely that the widespread target coding we observed (66% of parcels) reflects an overestimation of neural activity that was driven by task-correlated noise.

[8]While likely, this hypothesis could be tested by performing "RSA" on the timecourses of framewise displacement (i.e., the rigid-body motion estimates).: is target information decodable from these traces?

To mitigate this issue, we used model comparison. Model comparison is a characteristic advantage of the RSA framework over other multivariate decoding frameworks (MVPA), which typically do not furnish effect sizes (but rather decoding accuracies), and typically do not involve multiple models being compared in terms of their fit to the entire representational structure. For our purposes, RSA model comparison gave a principled way of narrowing the list of "target coding" parcels to those carrying representations less likely to be driven solely by noise. Encouragingly, these model comparisons led to a substantial reduction in the number of identified parcels; and, those that were identified included areas in line with previous research (see *Discussion* section *Specific implications: Neuroanatomical profiles*). Further, it is less clear how these sources of noise could account for the predicted and anatomically specific correlations we observed between the strength of target coding (in IPS and dlPFC) and the size of the behavioral Stroop effect.[9] Thus, while task-correlated motion may have inflated target decoding, our method of model comparison was to some degree effective at distinguishing neurally driven pattern representations.

Future fMRI work with verbal tasks should anticipate these sources of noise and use more aggressive motion removal procedures (either statistical or procedural). But, we note that the problem of "task-correlated noise", if conceptualized more generally to include correlated *feature spaces*, is not unique to fMRI. That is, neuroimagers and neurophysiologists alike must contend with the fact that the measured response of a region (or, e.g., neuron) may seem to reflect a particular hypothesized feature space, but in actuality encodes a correlated, yet fundamentally different, space. For example, rather than "conflict", dmPFC activity has been alternatively interpreted as encoding "time on task" (Grinband et al., 2011). This problem of correlated feature spaces is addressable via experimental design. In general, more

[9]To bolster this stance, additional correlational analyses could be performed that attempt to control for potential relationships between motion estimates and Stroop.

elaborate stimulus sets enable representational models with more fine-grained distinctions to be compared. Further, in combination with an expansive stimulus set, demonstrating specific modulation of a representation as a result of contextual manipulations (e.g., attention, proportion congruency) would afford stronger evidence for the model representation (Popov, Ostarek, & Tenison, 2018). In addressing these more fine-grained representational questions, future work would also likely attenuate issues with task-correlated "nuisance" factors.

4.3.2 Analytic decisions.

A general limitation of the current study was our choice of the particular statistical methods for RSA. While we followed the originally recommended default procedures (Nili et al., 2014), in recent years, several issues with these procedures have been highlighted and substantial improvements have made (Diedrichsen & Kriegeskorte, 2017).

In particular, interpretational problems arise from the use of the Pearson correlation to estimate pattern similarity. The Pearson correlation is statistically biased: with increasing noise, the expected value of r shrinks toward 0 (away from the "true" similarity). Because different ROIs have different signal-to-noise ratios (SNR), this bias makes it difficult to compare RSA model fits between regions. For example, we found that target coding in dlPFC (FEF) was not greater than dmPFC (Figure 3). Yet, it is hard to determine whether this was due to increased noise in dlPFC (e.g., from the scanner), or because of no "true" difference. The use of a biased statistic would also systematically impact individual differences analysis, as SNR likely also varies across participants. These issues can be entirely circumvented, however, by using unbiased estimators, which can be obtained in RSA through cross-validating patterns across scanning runs (Walther et al., 2016) or through empirical Bayesian methods (e.g.,

Diedrichsen, Ridgway, Friston, & Wiestler, 2011; Cai, Schuck, Pillow, & Niv, 2019; Friston, Diedrichsen, Holmes, & Zeidman, 2019). An added benefit of cross-run estimation is that it naturally affords an estimate of split-half pattern reliability. Reliability estimates could be used for several purposes: for example, to select ROIs that reliably encode some aspect of the task prior to RSA, or as a metric for unbiased methodological optimization (i.e., to find pre-processing procedures that maximize reliability).

4.4 Prospective design recommendations for RSA in Stroop.

A prospective design could address many limitations of the current study, in addition to many new questions. Here, we provide general suggestions for prospective studies.

4.4.1 Use unbiased measures of similarity.

Many interesting Stroop manipulations rely on changing the presentation frequency of certain stimuli or conditions to effect certain cognitive control processes. While creating a design that employs these manipulations, the current best practice is to use two sets of stimuli: those that induce the process, and those that diagnose the process (Braem et al., 2019). The inducer stimuli carry the frequency manipulations, while the diagnostic stimuli are "unbiased" (e.g., congruency status is uncorrelated with target or distractor identities, and cannot be predicted in advance). Thus, with equal proportion congruent:incongruent, the total number of the diagnostic stimuli will necessarily be unbalanced within each cell of the RSA matrix (i.e., incongruent stimuli will be presented less often than congruent). This

poses a problem for RSA, as patterns estimated from more numerous stimuli (i.e., more reliable estimates) will tend to have stronger correlations with other patterns. (In the case of diagnostic stimuli, this would inflate the correlations among congruent stimuli.)

Two solutions to this problem would be to either down-sample (exclude data), or to focus only on mostly incongruent lists (as we did here). A much more practical solution, however would be to simply to use an *unbiased measure* of similarity (see *Limitations*). Because the expected value of these measures is independent of the number of trials contributing to the pattern estimate, using such a measure would side-step this design issue.

4.4.2 Incorporate experimental control conditions or tasks.

Adding certain conditions to the paradigm would help to shed light on target and distractor representation. Including a reverse Stroop list, in which participants would be instructed to read the word (ignoring color), could be useful for several purposes. For example, this condition could used as a "negative" control, as any brain–behavior relationship that is hypothesized to depend on control should not be present (albeit, a null correlation would be qualified by the probable reduction in across-participant variance of the reverse Stroop effect). Also, examining the relative cortical distribution of target and distractor coding within reversed and "forward" Stroop tasks could highlight the impact that different contextual rules have on widespread cortical processing. Including a nonverbal paradigm would additionally be desirable to address concerns of task-correlated noise, although this comes with limitations, of course, of limited stimulus sets and arbitrary response mappings.

4.4.3 Densely sample the stimulus space.

RSA can be likened to casting a net over an invisible structure: the denser the net, the more detail will become visible. That is, with more diversity in the set of stimuli (or task conditions), more finer-grained representational models will be able to be teased apart. But, this type of "condition-rich" approach is incongruent with typical color-word Stroop investigations, which have typically used between 4–16 unique stimuli (whereas RSA studies often use upwards of 30). More diverse stimulus sets could be incorporated in Stroop in a number of ways.

One method for increasing the specificity of the results would be to incorporate condition-rich multivariate localizer tasks for different representations. That is, prior to the color-word Stroop experiment, participants could be presented with extensive lists of words and colors. A variety of models could be fit on these data to identify regions that preferentially represent certain stimulus features, which could be used as functional ROIs in the subsequent Stroop task.[10] For example, distractor models could be built to estimate representations of low-level visual form (by calculating the pixel-by-pixel overlap of images), orthography (e.g., via a model of open bigram similarity; Whitney, 2008), semantic features (via similarity in, e.g., word2vec embeddings; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), or phonology (e.g., Fischer-Baum, Bruggemann, Gallego, Li, & Tamez, 2017). Similarly, target models could be build to estimate hue (using similarity in a color space), semantic, or phonological representations. Further, manipulating the instructions (e.g., "name the color" versus "ignore the colors; perform another [irrelevant] task"), response versus perceptual processes could be (de)emphasized. When creating the stimuli lists, the to-be-tested models should

[10]This localizer analysis could be conducted via searchlight procedures, or a parcellation-based approach. Although searchlights may be appealing, care would have to be taken in establishing searchlight-based ROIs (Etzel et al., 2013)

be built and examined. An optimal stimulus list would result in orthogonal representational models. Combining these procedures with the color-word Stroop paradigm and proportion congruence manipulations could provide a greater level of representational specificity to the findings (e.g., "In mostly incongruent versus congruent lists, distractor representations within regions sensitive to orthographic features were attenuated.").

A stronger design would enrich the Stroop stimulus set itself to test for specific context-based modulation of representations. (This could be done in combination with a localizer task.) At least relative to the present design (of 4 colors $\times$ 4 words), the color-word stimulus set could be expanded (e.g., to 8 or 12 colors and words). Ultimately, however, the number of color words with high hue–name agreement is limited. Using Stroop variants that are not confined to the set of nameable colors (e.g., picture-word interference task) can address this issue.

Chapter 5

Conclusion

In the classic color-word Stroop task, distinct stimulus dimensions evoke corresponding components of fMRI activity patterns, with predictable dissociations in neuroanatomical distribution and behavioral relevance. With care, these representations can be revealed through simple representational similarity analysis. This neuroimaging approach opens the door for more sophisticated tests of cognitive control theory, as the language of many such theories is representational in nature.

References

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344. doi:10.1038/nn.2921

Alink, A., Walther, A., Krugliak, A., Bosch, J. J. F. van den, & Kriegeskorte, N. (2015). Mind the drift - improving sensitivity to fMRI pattern information by accounting for temporal pattern drift. *bioRxiv*, 032391. doi:10/gfsh5f

Assem, M., Glasser, M. F., Essen, D. C. V., & Duncan, J. (2019). A Domain-general Cognitive Core defined in Multimodally Parcellated Human Cortex. *bioRxiv*, 517599. doi:10/gftptd

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

Banich, M. T., Milham, M. P., Jacobson, B. L., Webb, A., Wszalek, T., Cohen, N. J., & Kramer, A. F. (2001). Chapter 29 Attentional selection and the processing of task-irrelevant information: Insights from fMRI examinations of the Stroop task. In *Progress in Brain Research* (Vol. 134, pp. 459–470). Elsevier. doi:10.1016/S0079-6123(01)34030-X

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). Lme4: Linear mixed-effects models using Eigen and S4. *R Package Version*, *1*(7), 1–23.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. doi:10/gfpkdx

Bonini, F., Burle, B., Ligeois-Chauvel, C., Rgis, J., Chauvel, P., & Vidal, F. (2014). Action Monitoring and Medial Frontal Cortex: Leading Role of Supplementary Motor Area. *Science*, *343*(6173), 888–891. doi:10/f5r6j5

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict Monitoring and Cognitive Control. *Psychological Review*, *108*(3), 624–652. doi:10.1037//0033-295X.I08.3.624

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring Adaptive Control in Conflict Tasks. *Trends in Cognitive Sciences*, S1364661319301640. doi:10/gf48x7

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. doi:10.1016/j.tics.2011.12.010

Brown, J. W., & Braver, T. S. (2005). Learned Predictions of Error Likelihood in the Anterior Cingulate Cortex. *Science; Washington*, *307*(5712), 1118–21. doi:10/c7qqpd

Bugg, J. M., & Crump, M. J. C. (2012). In Support of a Distinction between Voluntary and Stimulus-Driven Control: A Review of the Literature on Proportion Congruent Effects. *Frontiers in Psychology*, *3*. doi:10/gf39wh

Bugg, J. M., Jacoby, L. L., & Chanani, S. (2011). Why it is too early to lose control in accounts of item-specific proportion congruency effects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 844–859. doi:10/d6nd5h

Buschman, T. J., & Miller, E. K. (2007). Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices. *Science*, *315*(5820), 1860–1862. doi:10/dw85gh

Cai, M. B., Schuck, N. W., Pillow, J. W., & Niv, Y. (2019). Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLOS Computational Biology*, *15*(5), e1006299. doi:10/gf3cx4

Carter, C. S., Macdonald, A. M., Botvinick, M., Ross, L. L., Stenger, V. A., Noll, D., & Cohen, J. D. (2000). Parsing executive processes: Strategic vs. Evaluative functions of the anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *97*(4), 1944–1948. doi:10/bssf4g

Chawla, D., Rees, G., & Friston, K. J. (1999). The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neuroscience*, *2*(7), 671–676. doi:10/d6hpfk

Cieslik, E. C., Mueller, V. I., Eickhoff, C. R., Langner, R., & Eickhoff, S. B. (2015). Three key regions for supervisory attentional control: Evidence from neuroimaging meta-analyses. *Neuroscience & Biobehavioral Reviews*, *48*, 22–34. doi:10.1016/j.neubiorev.2014.11.003

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(3), 332–361. doi:10/cfn9t7

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. doi:10/brm459

Cox, R. W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. doi:10/ctwqf6

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. doi:10/gf65m8

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341. doi:10/btxznv

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, *13*. doi:10.1371/journal.pcbi.1005508

Diedrichsen, J., Ridgway, G. R., Friston, K. J., & Wiestler, T. (2011). Comparing the similarity and spatial structure of neural representations: A pattern-component model. *NeuroImage*, *55*(4), 1665–1678. doi:10.1016/j.neuroimage.2011.01.044

Dimsdale-Zucker, H. R., & Ranganath, C. (2018). Representational Similarity Analyses. In *Handbook of Behavioral Neuroscience* (Vol. 28, pp. 509–525). Elsevier. doi:10.1016/B978-0-12-812028-6.00027-6

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. doi:10.1016/j.neuron.2006.04.031

Dreisbach, G., & Haider, H. (2008). That's what task sets are for: Shielding against irrelevant information. *Psychological Research*, *72*(4), 355–361. doi:10/dwwxxt

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience*, *2*(11), 820–829. doi:10/b3t548

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. doi:10.1016/j.tics.2010.01.004

Duncan-Johnson, C. C., & Kopell, B. S. (1981). The Stroop effect: Brain potentials localize the source of interference. *Science*, *214*(4523), 938–940. doi:10/dsvr67

Egner, T. (2014). Creatures of habit (and control): A multi-level learning perspective on the modulation of congruency effects. *Frontiers in Psychology*, *5*(NOV), 1–11. doi:10.3389/fpsyg.2014.01247

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), 1784–1790. doi:10.1038/nn1594

E-Prime. (2016). Pittsburgh, PA: Psychology Software Tools, Inc.

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, *78*, 261–269. doi:10.1016/j.neuroimage.2013.03.041

Fischer-Baum, S., Bruggemann, D., Gallego, I. F., Li, D. S. P., & Tamez, E. R. (2017). Decoding levels of representation in reading: A representational similarity approach. *Cortex*, *90*, 88–102. doi:10/gf7hhm

Friston, K. J., Diedrichsen, J., Holmes, E., & Zeidman, P. (2019). Variational representational similarity analysis. *NeuroImage*, 115986. doi:10/gf5hv5

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. doi:10/dbh75h

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

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*(7615), 171–178. doi:10.1038/nature18933

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., ... Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124. doi:10/f46nj4

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574. doi:10/bd7gvx

Gonthier, C., Braver, T. S., & Bugg, J. M. (2016). Dissociating proactive and reactive control in the Stroop task. *Memory & Cognition*, *44*(5), 778–788. doi:10.3758/s13421-016-0591-1

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., & Petersen, S. E. (2016). Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cerebral Cortex*, *26*(1), 288–303. doi:10.1093/cercor/bhu239

Grandjean, J., D'Ostilio, K., Phillips, C., Balteau, E., Degueldre, C., Luxen, A., ... Collette, F. (2012). Modulation of Brain Activity during a Stroop Inhibitory Task by the Kind of Cognitive Control Required. *PLOS ONE*, *7*(7), e41513. doi:10/f357x3

Grinband, J., Savitsky, J., Wager, T. D., Teichert, T., Ferrera, V. P., & Hirsch, J. (2011). The Dorsal Medial Frontal Cortex is Sensitive to Time on Task, Not Response Conflict or Error Likelihood. *NeuroImage*, *57*(2), 303–311. doi:10.1016/j.neuroimage.2010.12.027

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. doi:10/gcp4v6

Hall-McMaster, S., Muhle-Karbe, P., Myers, N., & Stokes, M. (2019). Reward boosts neural coding of task rules to optimise cognitive flexibility. *bioRxiv*, 578468. doi:10/gfw6sc

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Oxford, England: Wiley.

Henriksson, L., Khaligh-Razavi, S.-M., Kay, K., & Kriegeskorte, N. (2015). Visual representations are dominated by intrinsic fluctuations correlated between areas. *NeuroImage*, *114*, 275–286. doi:10/f7dgzv

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences*, *8*(11), 494–500. doi:10/dvbnjx

Hubbard, J., Kikumoto, A., & Mayr, U. (2019). EEG Decoding Reveals the Strength and Temporal Dynamics of Goal-Relevant Representations. *Scientific Reports*, *9*(1), 1–11. doi:10/gf4v7t

Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance Monitoring by the Anterior Cingulate Cortex During Saccade Countermanding. *Science*, *302*(5642), 120–122. doi:10/bbmc8c

Jackson, J., & Woolgar, A. (2018). Adaptive coding in the human brain: Distinct object features are encoded by overlapping voxels in frontoparietal cortex. *Cortex*, (August), 2–11. doi:10.1016/j.cortex.2018.07.006

Jacob, S. N., & Nieder, A. (2014). Complementary Roles for Primate Frontal and Parietal Cortex in Guarding Working Memory from Distractor Stimuli. *Neuron*, *83*(1), 226–237. doi:10/f6dfxn

Jiang, J., & Egner, T. (2014). Using neural pattern classifiers to quantify the modularity of conflict-control mechanisms in the human brain. *Cerebral Cortex*, *24*(7), 1793–1805. doi:10.1093/cercor/bht029

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671. doi:10/bwh9mt

Kok, P., Mostert, P., & Lange, F. P. de. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, 201705652–201705652. doi:10.1073/pnas.1705652114

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 1–28. doi:10.3389/neuro.06.0

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. *Annual Review of Neuroscience*, *42*(1), 407–432. doi:10/gf56s3

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. doi:10.1016/j.tics.2013.06.007

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. doi:10/dk9pcw

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 219–234. doi:10/d42zj4

Logan, G. D. (1980). Attention and automaticity in Stroop and priming tasks: Theory and data. *Cognitive Psychology*, *12*(4), 523–553. doi:10/b862f6

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, *7*(3), 166–174. doi:10/dxn324

MacDonald, A. W., Cohen, J. D., Andrew Stenger, V., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835–1838. doi:10.1126/science.288.5472.1835

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203. doi:10.1037/0033-2909.109.2.163

Mair, P., & Wilcox, R. (2018). *WRS2: Wilcox robust estimation and testing*.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, *503*(7474), 78–84. doi:10/f5gdn6

Maronna, R. A., & Yohai, V. J. (1995). The Behavior of the Stahel-Donoho Robust Multivariate Estimator. *Journal of the American Statistical Association*, *90*(429), 330–341. doi:10/cpjmtj

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc.

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. doi:10.1146/annurev.neuro.24.1.167

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *The Journal of Neuroscience*, *16*(16), 5154–5167. doi:10/gfvsq9

Munakata, Y., Herd, S. A., Chatham, C. H., Depue, B. E., Banich, M. T., & O'Reilly, R. C. (2011). A unified framework for inhibitory control. *Trends in Cognitive Sciences*, *15*(10), 453–459. doi:10.1016/j.tics.2011.07.011

Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: Insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(1), 1–17. doi:10/ff7z4n

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). doi:10.1371/journal.pcbi.1003553

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2018). *Nlme: Linear and Nonlinear Mixed Effects Models*.

Pinheiro, J., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, *174*(November 2017), 340–351. doi:10.1016/j.neuroimage.2018.03.041

Qi, X.-L., Elworthy, A. C., Lambert, B. C., & Constantinidis, C. (2014). Representation of remembered stimuli and task information in the monkey dorsolateral prefrontal and posterior parietal cortex. *Journal of Neurophysiology*, *113*(1), 44–57. doi:10/f6vbp3

Rademaker, R. L., Chunharas, C., & Serences, J. T. (2019). Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience*, 1. doi:10/gf4kkp

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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. doi:10.1038/nature12160

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

Sarafyazd, M., & Jazayeri, M. (2019). Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, *364*(6441), eaav8911. doi:10/gf2tzj

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, *55*, 103–111. doi:10/gfxhm8

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron*, *91*(6), 1402–1412. doi:10/f889vm

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. *Neuron*, *79*(2), 217–240. doi:10.1016/j.neuron.2013.07.007

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. doi:10.1016/j.neuron.2013.01

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. doi:10.1037/h0054651

Tanji, J., & Hoshi, E. (2008). Role of the Lateral Prefrontal Cortex in Executive Behavioral Control. *Physiological Reviews*, *88*(1), 37–57. doi:10/bkw4t3

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. doi:10.1016/j.neuroimage.2015.12.012

Weiner, K. S., Barnett, M. A., Lorenz, S., Caspers, J., Stigliani, A., Amunts, K., ... Grill-Spector, K. (2017). The Cytoarchitecture of Domain-specific Regions in Human High-level Visual Cortex. *Cerebral Cortex*, *27*(1), 146–161. doi:10.1093/cercor/bhw361

Whitney, C. (2008). Supporting the serial in the SERIOL model. *Language and Cognitive Processes*, *23*(6), 824–865. doi:10/fk7krw

Wilcox, R., Rousselet, G., & Pernet, C. (2018). Improved Methods for Making Inferences About Multiple Skipped Correlations. *arXiv:1807.05048 [Stat]*.

Wilk, H. A., Ezekiel, F., & Morton, J. B. (2012). Brain regions associated with moment-to-moment adjustments in control and stable task-set maintenance. *NeuroImage*, *59*(2), 1960–1967. doi:10/dtk5g9

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, *4*(1), 58–73. doi:10/cvx2tx

Tables

Table 1: Task-relevant selective parcels ($\mathbf{t} > \{0, \mathbf{d}\}$, $\mathbf{d} \not> 0$).

parcel	neighborhood	rho(t)	p(t¿0)	rho(d)	p(d¿0)	rho(c)	p(c¿0)
d32_r	aCC and mPFC	0.04	0.00	0.00	0.48	0.04	0.07
a24_r	aCC and mPFC	0.02	0.02	-0.01	0.92	0.02	0.34
9m_l	aCC and mPFC	0.02	0.01	-0.01	0.88	0.03	0.17
STSda_r	aud. assoc.	0.04	0.00	0.00	0.79	0.00	0.63
STSdp_l	aud. assoc.	0.02	0.01	-0.01	0.92	0.02	0.27
V6A_r	d vis.	0.03	0.00	-0.01	0.96	-0.01	0.86
p9-46v_r	dlPFC	0.05	0.00	0.01	0.37	0.04	0.10
8C_r	dlPFC	0.04	0.00	0.00	0.29	0.03	0.12
9p_l	dlPFC	0.04	0.00	-0.01	0.76	0.03	0.23
a9-46v_r	dlPFC	0.04	0.00	0.00	0.71	0.01	0.17
8Ad_r	dlPFC	0.04	0.00	0.00	0.59	0.04	0.08
8BL_r	dlPFC	0.03	0.00	-0.01	0.83	0.05	0.04
8Ad_l	dlPFC	0.03	0.00	-0.01	0.94	0.04	0.08
46_l	dlPFC	0.03	0.01	-0.01	0.85	0.01	0.44
9-46d_r	dlPFC	0.03	0.02	-0.01	0.85	0.04	0.15
8Av_l	dlPFC	0.03	0.01	-0.01	0.88	0.02	0.25
LBelt_l	early aud.	0.05	0.00	0.01	0.10	0.04	0.17
RI_l	early aud.	0.04	0.00	0.01	0.10	0.04	0.05
MBelt_r	early aud.	0.04	0.00	-0.01	0.77	0.01	0.55
A1_l	early aud.	0.03	0.01	-0.01	0.78	0.03	0.31
IFSp_r	iFC	0.04	0.00	0.00	0.39	0.03	0.14
45_l	iFC	0.04	0.00	-0.01	0.84	0.03	0.19
IFJa_l	iFC	0.04	0.00	0.00	0.58	0.06	0.00
47l_l	iFC	0.03	0.00	0.01	0.49	0.00	0.72
47l_r	iFC	0.03	0.03	-0.01	0.77	0.04	0.23
p47r_r	iFC	0.03	0.03	-0.01	0.92	0.01	0.40
IFSp_l	iFC	0.02	0.01	-0.01	0.82	0.05	0.06
FOP4_r	insular and FO	0.05	0.00	0.00	0.32	0.06	0.04
MI_r	insular and FO	0.05	0.00	0.01	0.38	0.05	0.03

FOP5_r	insular and FO	0.05	0.00	0.01	0.32	0.04	0.20
FOP3_l	insular and FO	0.04	0.00	0.01	0.38	0.03	0.09
Ig_l	insular and FO	0.04	0.00	0.00	0.75	0.05	0.03
FOP3_r	insular and FO	0.04	0.00	0.01	0.61	0.03	0.17
PoI2_l	insular and FO	0.04	0.00	-0.01	0.81	0.00	0.50
PI_r	insular and FO	0.03	0.01	0.00	0.41	0.03	0.13
MI_l	insular and FO	0.03	0.00	-0.01	0.92	0.02	0.18
AVI_r	insular and FO	0.03	0.00	-0.01	0.91	0.03	0.32
IP2_r	iP	0.04	0.00	-0.01	0.79	0.07	0.01
PFop_r	iP	0.04	0.00	0.00	0.47	0.01	0.35
PGi_l	iP	0.03	0.00	0.00	0.53	0.03	0.23
IP0_l	iP	0.02	0.02	-0.01	0.76	0.05	0.04
PGp_l	iP	0.02	0.02	-0.01	0.92	0.05	0.04
TGd_r	lT	0.04	0.00	0.00	0.73	0.01	0.30
TGd_l	lT	0.03	0.00	0.00	0.57	0.00	0.47
TE2p_l	lT	0.03	0.00	-0.01	0.85	0.01	0.51
TE1a_l	lT	0.03	0.01	-0.01	0.93	-0.01	0.76
TE1a_r	lT	0.03	0.01	-0.01	0.91	-0.01	0.80
PHT_l	lT	0.03	0.00	0.00	0.55	0.05	0.03
TF_r	lT	0.02	0.02	-0.01	0.90	-0.01	0.84
3b_r	M1 and S1	0.06	0.00	0.02	0.06	0.04	0.13
4_r	M1 and S1	0.06	0.00	0.00	0.39	0.02	0.24
2_l	M1 and S1	0.04	0.00	-0.01	0.89	0.03	0.13
1_l	M1 and S1	0.03	0.01	-0.01	0.88	0.02	0.22
PeEc_l	mT	0.04	0.00	0.00	0.80	0.01	0.51
FST_r	MT+	0.04	0.00	0.01	0.46	0.07	0.01
V4t_l	MT+	0.04	0.00	0.00	0.71	0.02	0.30
LO3_r	MT+	0.03	0.01	0.00	0.68	0.01	0.40
p10p_l	oFC and pFC	0.05	0.00	0.00	0.43	0.04	0.03
a47r_l	oFC and pFC	0.04	0.00	0.01	0.28	0.02	0.15
OFC_r	oFC and pFC	0.03	0.00	-0.01	0.81	0.00	0.52
OFC_l	oFC and pFC	0.02	0.01	-0.01	0.93	0.03	0.18
47s_r	oFC and pFC	0.02	0.03	-0.01	0.77	0.01	0.43
23c_l	pCC	0.05	0.00	0.00	0.62	0.03	0.12
PCV_l	pCC	0.02	0.02	-0.01	0.87	0.05	0.03
SCEF_r	PL and mid cing.	0.05	0.00	0.01	0.26	0.05	0.09
24dv_l	PL and mid cing.	0.03	0.02	0.00	0.62	0.01	0.37
FEF_r	PM	0.05	0.00	0.01	0.27	0.03	0.12
6v_r	PM	0.05	0.00	-0.01	0.86	0.01	0.37
55b_l	PM	0.05	0.00	0.00	0.69	0.04	0.10
55b_r	PM	0.05	0.00	-0.01	0.77	0.04	0.05

6d_l	PM	0.04	0.00	0.00	0.38	0.03	0.10
FEF_l	PM	0.04	0.00	-0.01	0.94	0.03	0.23
6v_l	PM	0.03	0.00	-0.02	0.96	0.00	0.45
6a_r	PM	0.03	0.00	0.00	0.72	0.06	0.01
OP4_l	pOperc	0.06	0.00	-0.01	0.88	0.05	0.03
43_r	pOperc	0.04	0.00	0.01	0.19	0.00	0.71
PFcm_l	pOperc	0.04	0.00	0.01	0.22	0.03	0.17
OP4_r	pOperc	0.03	0.00	-0.01	0.83	0.00	0.78
OP1_l	pOperc	0.03	0.00	0.00	0.45	0.03	0.13
43_l	pOperc	0.03	0.00	0.00	0.48	0.02	0.34
FOP1_r	pOperc	0.03	0.00	-0.01	0.84	0.02	0.38
AIP_r	sP	0.05	0.00	0.00	0.38	0.03	0.21
AIP_l	sP	0.04	0.00	0.00	0.35	0.09	0.01
7PC_l	sP	0.04	0.00	0.00	0.67	0.02	0.21
LIPv_r	sP	0.03	0.01	0.00	0.64	0.05	0.01
MIP_r	sP	0.03	0.01	0.00	0.66	0.08	0.01
LIPd_r	sP	0.02	0.02	-0.01	0.83	0.03	0.26
PSL_r	TPOJ	0.03	0.01	0.00	0.51	0.03	0.17
V8_l	v vis.	0.04	0.00	0.01	0.39	0.03	0.39
VVC_l	v vis.	0.04	0.00	0.00	0.72	0.03	0.24
VVC_r	v vis.	0.04	0.00	0.00	0.37	0.02	0.39
FFC_l	v vis.	0.03	0.01	0.00	0.52	0.03	0.26

Note. This list does not include target-selective parcels (see Table 2), or task-relevant–congruency selective parcels (see Table 1). Anterior cingulate and medial prefrontal cortex (aCC and mPFC); auditory association cortex (aud. assoc.); dorsal visual cortex (d vis.); dorsolateral prefrontal cortex (dlPFC); early auditory cortex (early aud.); inferior frontal cortex (iFC); insular and frontal operculum (insular and FO); inferior parietal lobule (iP); lateral temporal cortex (lT); mT (middle temporal); orbital and polar frontal cortex (oFC and pFC); posterior cingulate cortex (pCC); premotor cortex (PM); posterior operculum (pOperc); superior parietal lobule (sP); temporo-parietal-occipital junction (TPOJ); ventral visual (v vis.).

Table 2: Target-selective parcels ($\mathbf{t} > \{0, \mathbf{d}, \mathbf{c}\}$, $\{\mathbf{d}, \mathbf{c}\} \not> 0$).

parcel	neighborhood	rho(t)	p(t¿0)	rho(d)	p(d¿0)	rho(c)	p(c¿0)
10r_r	aCC and mPFC	0.02	0.02	0.00	0.56	-0.01	0.71
STSda_l	aud. assoc.	0.05	0.00	-0.01	0.81	0.00	0.45
STGa_l	aud. assoc.	0.04	0.00	0.01	0.35	0.00	0.68
A4_r	aud. assoc.	0.04	0.00	-0.01	0.86	-0.01	0.88
A5_r	aud. assoc.	0.02	0.03	-0.01	0.86	-0.02	0.92
V6_r	d vis.	0.04	0.00	0.01	0.08	0.00	0.64
PBelt_r	early aud.	0.05	0.00	0.00	0.49	0.01	0.62
PFop_l	iP	0.05	0.00	-0.01	0.89	0.00	0.52
TGv_r	lT	0.04	0.00	0.01	0.22	0.01	0.55
TE2a_l	lT	0.03	0.01	-0.01	0.80	-0.02	0.87
3b_l	M1 and S1	0.08	0.00	-0.01	0.86	0.00	0.46
4_l	M1 and S1	0.06	0.00	-0.01	0.85	-0.01	0.79
3a_l	M1 and S1	0.06	0.00	-0.01	0.85	-0.02	0.95
11l_l	oFC and pFC	0.04	0.00	0.01	0.12	0.01	0.51
11l_r	oFC and pFC	0.04	0.00	0.01	0.29	0.00	0.64
POS2_r	pCC	0.04	0.00	0.00	0.65	0.00	0.70
5mv_r	PL and mid cing.	0.05	0.00	0.02	0.10	0.02	0.30
5m_l	PL and mid cing.	0.04	0.00	0.00	0.35	-0.01	0.67
OP2-3_l	pOperc	0.03	0.00	0.00	0.49	0.00	0.76

Note. See Table 1 for "neighborhood" abbreviations.

Table 3: Distractor coding parcels ($\mathbf{d} > 0$).

parcel	neighborhood	rho(t)	p(t¿0)	rho(d)	p(d¿0)	rho(c)	p(c¿0)
V2_l	early vis.	0.03	0.00	0.05	0.00	0.04	0.05
V3_r	early vis.	0.03	0.02	0.04	0.00	0.06	0.03
V1_l	V1	0.03	0.05	0.05	0.00	0.06	0.03

Note. See Table 1 for "neighborhood" abbreviations.

Table 4: Congruency coding parcels ($\mathbf{c} > 0$).

parcel	neighborhood	rho(t)	p(t¿0)	rho(d)	p(d¿0)	rho(c)	p(c¿0)
8BM_r	aCC and mPFC	0.05	0.00	0.00	0.47	0.08	0.00
p32pr_r	aCC and mPFC	0.04	0.00	0.01	0.41	0.07	0.00
s6-8_r	dlPFC	0.03	0.01	0.00	0.37	0.08	0.00
i6-8_l	dlPFC	0.03	0.00	0.02	0.08	0.07	0.02
9-46d_l	dlPFC	0.01	0.29	-0.02	0.97	0.06	0.00
p9-46v_l	dlPFC	0.04	0.00	-0.01	0.92	0.06	0.01
SFL_l	dlPFC	0.04	0.00	0.02	0.05	0.06	0.01
IFJp_l	iFC	0.03	0.01	0.01	0.17	0.07	0.00
44_l	iFC	0.03	0.01	-0.01	0.93	0.06	0.01
FOP4_l	insular and FO	0.03	0.00	0.00	0.46	0.08	0.01
IP1_l	iP	0.02	0.02	0.01	0.37	0.13	0.00
IP0_r	iP	0.05	0.00	0.03	0.02	0.09	0.00
LO2_r	MT+	0.01	0.18	0.00	0.72	0.08	0.00
RSC_l	pCC	0.03	0.00	0.01	0.06	0.08	0.00
SCEF_l	PL and mid cing.	0.04	0.00	0.01	0.07	0.09	0.00
6ma_l	PL and mid cing.	0.02	0.04	0.02	0.04	0.08	0.00
6ma_r	PL and mid cing.	0.05	0.00	0.02	0.07	0.07	0.00
PEF_l	PM	0.04	0.00	0.01	0.11	0.07	0.01
MIP_l	sP	0.02	0.02	0.02	0.09	0.12	0.00
LIPd_l	sP	0.02	0.01	0.02	0.03	0.11	0.00

Note. See Table 1 for "neighborhood" abbreviations.

Table 5: Congruency coding parcels that were better explained by an "incongruency" coding model.

parcel.hemi	rho(incon-con)	p
IP1_l	0.03	0.00
MIP_l	0.03	0.01
SFL_l	0.02	0.01
6ma_l	0.02	0.01
6ma_r	0.02	0.02
SCEF_l	0.02	0.04
LIPd_l	0.02	0.04

Note. Incongruency model fit (i); Congruency model fit (c). See Table 1 for "neighborhood" abbreviations.

Table 6: "Super-parcels" defined for brain–behavior analysis.

region	abbreviation	hemi	parcels	nvox
V1–V3	V1–V3	l	V1, V2, V3	1986
V1–V3	V1–V3	r	V1, V2, V3	1918
ventral occipito-temporal	VOT	l	FFC, VVC, V8, VMV3	721
ventral occipito-temporal	VOT	r	FFC, VVC, V8, VMV3, VMV2	695
intra parietal sulcal	IPS	l	IP0, IP1, LIPd, VIP, LIPv, AIP, 7PC	678
intra parietal sulcal	IPS	r	IP0, IP1, IP2, IPS1, MIP, LIPd, LIPv, AIP, 7PC	754
dorsolateral prefrontal	dlPFC	l	p9-46v, 8C, 8Av, i6-8	636
dorsolateral prefrontal	dlPFC	r	p9-46v, 8C, 8Av, i6-8	721
dorsomedial prefrontal	dmPFC	l	a32pr, p32pr, 8BM, SCEF	624
dorsomedial prefrontal	dmPFC	r	a32pr, p32pr, 8BM, SCEF	618
frontal insular	fIns	l	FOP4, FOP5, FOP3	278
frontal insular	fIns	r	FOP4, FOP5, FOP3	260
inferior frontal cortex	IFC	l	44, 45, IFSa, IFSp, p47r, p47l	723
inferior frontal cortex	IFC	r	44, 45, IFSa, IFSp, p47r, p47l	649
somato-motor–mouth	SMmouth	bil.		1114

Note. All parcels are from Glasser's atlas (Glasser et al., 2016), except *SMmouth*, which was obtained from Gordon et al. (2016).

Table 7: Correlations with $R^2 > 0.01$ between the strength of task-dimension coding and Stroop effect (RT) across individuals.

region	dimension	r [95% CI]	r^2	ρ [95% CI]	ρ^2
IPS (r)	target	-0.50 [-0.663 -0.275]	0.25	-0.45 [-0.667 -0.183]	0.20
dlPFC (l)	distractor	-0.39 [-0.645 -0.072]	0.15	-0.32 [-0.597 -0.010]	0.10
dlPFC (r)	target	-0.39 [-0.615 -0.096]	0.15	-0.35 [-0.596 -0.049]	0.12
VOT (l)	congruency	-0.32 [-0.507 -0.121]	0.10	-0.38 [-0.580 -0.139]	0.15
SMmouth	distractor	0.29 [0.064 0.507]	0.09	0.32 [0.040 0.564]	0.10
dmPFC (l)	congruency	0.26 [-0.018 0.495]	0.07	0.16 [-0.145 0.427]	0.03
V1–V3 (r)	distractor	-0.25 [-0.484 0.008]	0.06	-0.24 [-0.492 0.034]	0.06
V1–V3 (r)	congruency	-0.25 [-0.494 0.019]	0.06	-0.30 [-0.543 -0.009]	0.09
V1–V3 (l)	congruency	-0.24 [-0.526 0.085]	0.06	-0.31 [-0.562 -0.015]	0.09
V1–V3 (r)	target	-0.23 [-0.495 0.068]	0.05	-0.28 [-0.537 0.004]	0.08
VOT (r)	congruency	-0.22 [-0.507 0.070]	0.05	-0.32 [-0.570 -0.041]	0.11
IPS (r)	congruency	-0.20 [-0.459 0.092]	0.04	-0.19 [-0.460 0.119]	0.04
IFC (r)	target	-0.20 [-0.483 0.110]	0.04	-0.21 [-0.496 0.102]	0.04
dlPFC (r)	congruency	0.18 [-0.139 0.453]	0.03	0.10 [-0.195 0.373]	0.01
VOT (r)	target	-0.17 [-0.473 0.162]	0.03	-0.22 [-0.493 0.072]	0.05
dlPFC (l)	target	-0.16 [-0.402 0.089]	0.03	-0.17 [-0.429 0.121]	0.03
IPS (l)	congruency	-0.15 [-0.364 0.057]	0.02	-0.23 [-0.461 0.016]	0.05
SMmouth	congruency	0.14 [-0.146 0.411]	0.02	0.09 [-0.211 0.375]	0.01
fIns (l)	congruency	-0.13 [-0.376 0.127]	0.02	-0.16 [-0.425 0.137]	0.02
IPS (r)	distractor	-0.13 [-0.421 0.174]	0.02	-0.11 [-0.404 0.194]	0.01
dmPFC (l)	target	0.12 [-0.181 0.434]	0.02	0.13 [-0.182 0.428]	0.02
VOT (l)	distractor	-0.12 [-0.417 0.181]	0.01	-0.18 [-0.446 0.111]	0.03
SMmouth	target	-0.11 [-0.377 0.218]	0.01	-0.07 [-0.361 0.237]	0.00
fIns (r)	congruency	0.10 [-0.188 0.361]	0.01	0.04 [-0.271 0.329]	0.00
fIns (l)	target	-0.10 [-0.355 0.166]	0.01	-0.06 [-0.336 0.234]	0.00

Note. 95% confidence intervals obtained through bootstrap resampling (10,000 replicates). r = Pearson's correlation coefficient; ρ = Spearman's correlation coefficient; IPS = intra-parietal sulcus; dlPFC = dorsolateral prefrontal cortex; VOT = ventral occipito-temporal cortex; fIns = frontal insular; IFC = inferior frontal cortex; SMmouth = somato-motor–mouth.

Table 8: Parameter estimates from the selected brain–behavior model.

term	b	se	t	p	CPD
distractor, dlPFC (l)	-163.95	60.48	-2.71	0.01	0.08
target, IPS (r)	-153.48	44.09	-3.48	0.00	0.13
congruency, mPFC (l)	63.18	17.07	3.70	0.00	0.15
congruency, v-vis. (l)	-57.42	20.23	-2.84	0.01	0.09

Note. CPD = coefficient of partial determination, the unique variance explained by a given term.

Supplemental Figures

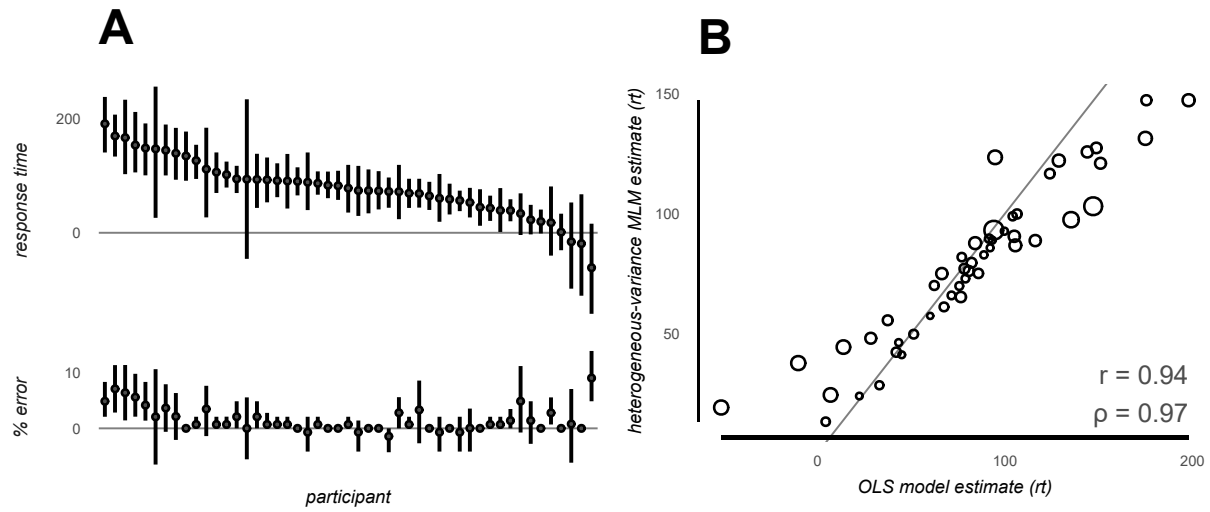

Figure 7: Stroop effects estimated with simple linear models. **A**, Participants' mean Stroop effects, estimated through simple linear contrast. These point-estimates were calculated the "standard" way: the mean RT or error rate on incongruent trials minus congruent, independently for each participant. Error bars represent percentile bootstrapped 95% confidence intervals. Two things are notable: three participants have negative Stroop effects, and there is substantial heterogeneity in variance across individuals' estimates. Notably, the individuals with negative Stroop effects (which are generally thought not to exist in the population) either have relatively increased error rates, larger variances, or a combination thereof. These factors — heterogeneity of variance, and negative Stroop effects — are undesireable, but can be taken into account by mixed-level models, which furnish "posterior" estimates of participants' stroop effects (i.e., after shrinking each effect toward the mean in proportion to its reliability). **B**, Plot of estimated Stroop effects from homogeneous versus heterogeneous-variance mixed-level models. The x-axis displays estimates from a homogeneous-variance model (note that these estimates were virtually identical to those estimated by simple linear contrast in **A**, with an R^2 of 0.88). The y-axis displays the same coefficients, estimated through a model that additionally estimates a separate residual variance parameter per participant. The area of each circle is proportional to the estimated residual variance (arbitrary scale). The horizontal deviations from the grey (unity) line indicate that the heterogeneous-variance model shrunk the participant coefficients more towards their mean. Notably, the larger circles (individuals with higher variance) tend to be more deviant from the line, illustrating that the less precise estimates tended to be more extreme.

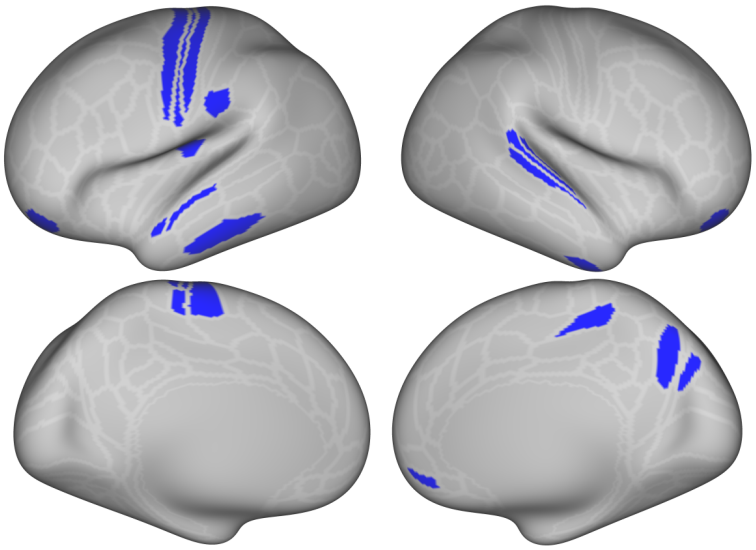

Figure 8: Target-selective parcels. Areas that displayed significant coding only for the target dimension and greater target coding than both distractor or congruency dimensions [$\mathbf{t} > (0, \mathbf{d}, \mathbf{c})$, $(\mathbf{d}, \mathbf{c}) \not> 0$]. Parcels within this set are within left somatomotor strip (4, 3a, 3b), bilateral STL (STGa–*l*, STSda–*l*, A4–*r*, PBelt–*r*), bilateral IT (TE2a–*l*, TGv–*r*), in addition to OFC (11-l and a10p–*r*) and vmPFC (10r–*r*), left rostral IPL (PFop–*l*), left posterior opercular (OP2-3–*l*), right posterior paracentral lobular (5mv–*r*), right precuneus (POS2–*r*) and medial extrastriate (V6–*r*).

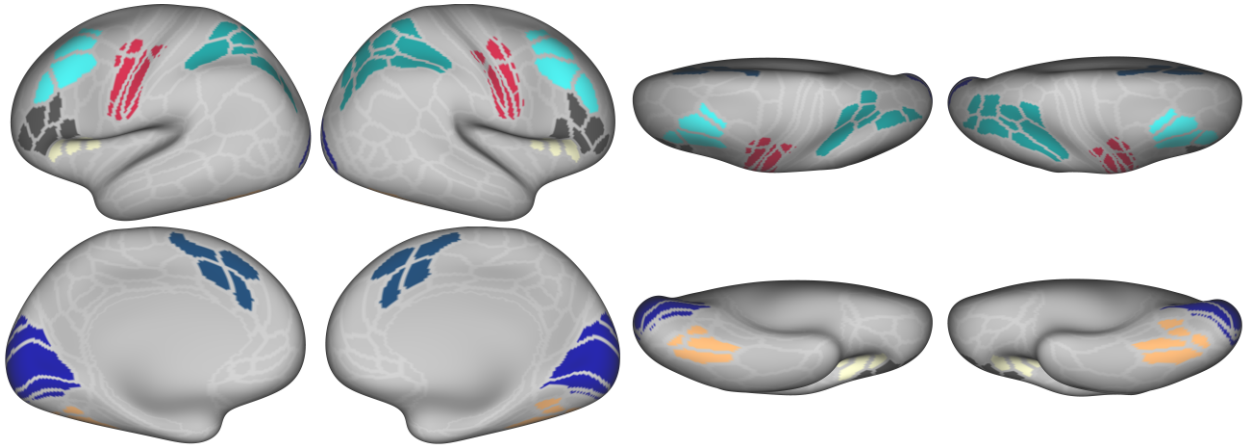

Figure 9: "Super-parcels" defined for brain–behavior analysis.

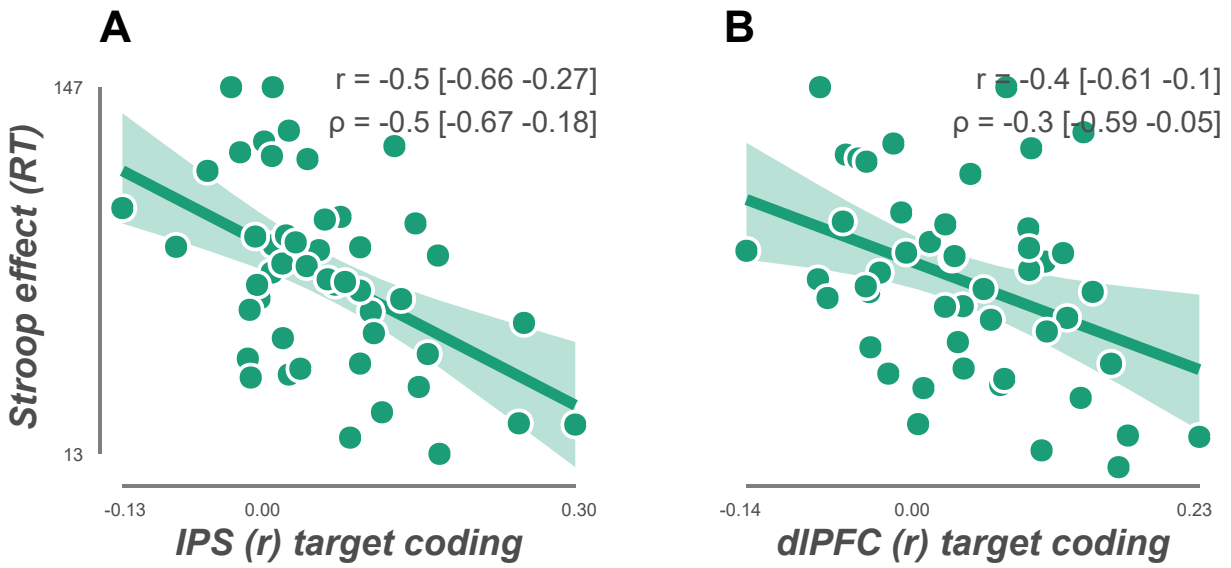

Figure 10: Bivariate relationships between target coding in IPS–r and dlPFC–r and behavior.

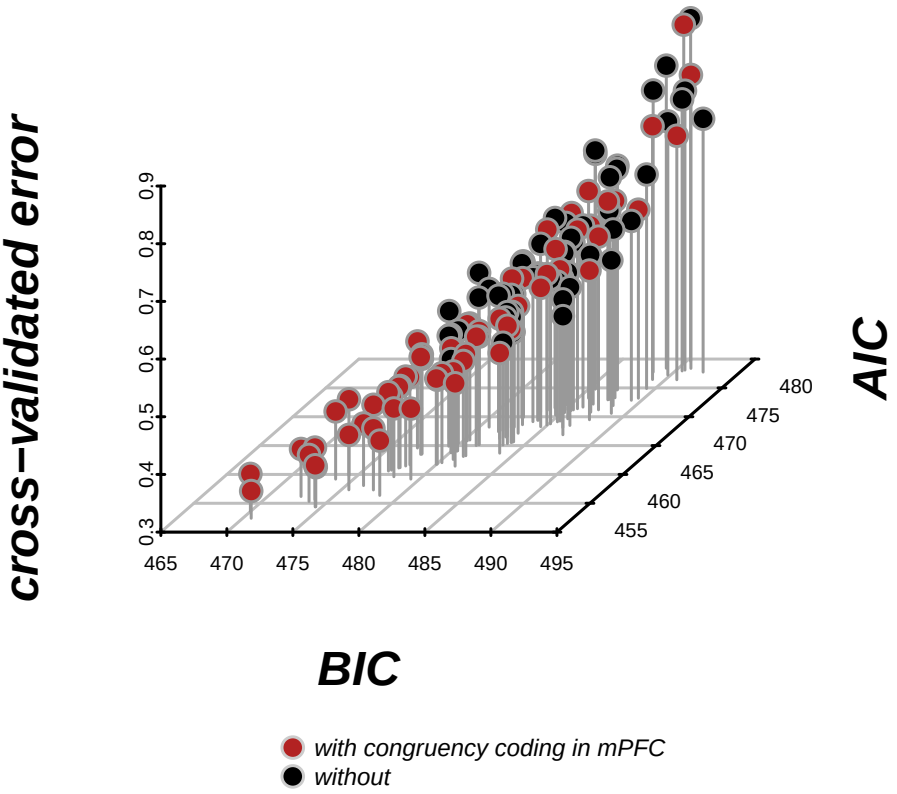

Figure 11: Fit statistics from brain–behavior regression model selection. Each point is a model: a unique combination of any number of the 7 explanatory variables, with Stroop effect in RT as the response variable. Cross-validated error (y-axis) refers to the error obtained in a leave-one-out cross-validation procedure (i.e., $1 - r_{CV}$; see *Method* section *Brain–behavior model fitting and evaluation*). Among the best models, the statistics agree: a general structure with congruency coding in left mPFC (with $+\beta$ coefficient), congruency coding in left ventral visual ($-$), target coding in right IPS ($-$), distractor coding in left dlPFC ($-$) yeilds the lowest BIC, and the second lowest CV-error and AIC. (The model that minimized these latter statistics additionally included a term for distractor coding in bilateral somato-motor–mouth ($+$); however, the change in AIC and CV-error associated with this addition was relatively small. Further, the best 17% of these models (with lowest BIC) contain coding of congruency in both mPFC (red points), suggesting that this term, despite its weak bivariate correlation (Figure 5D), is an important explanatory variable of individual variability in Stroop RT.