A representational similarity analysis of cognitive control during color-word Stroop

Michael C. Freund[1], Julie M. Bugg[1], & Todd S. Braver[1,2,3]

[1] Department of Psychological & Brain Sciences, Washington University in St. Louis, STL, MO 63130

[2,3] Departments of Radiology and Neuroscience, Washington University in St. Louis, School of Medicine, STL, MO 63110

Author Note

We have no conflicts of interest to disclose.

Michael C. Freund https://orcid.org/0000-0003-1701-0988

Julie M. Bugg http://orcid.org/0000-0002-0969-186X

Todd S. Braver https://orcid.org/0000-0003-1701-0988

The authors made the following contributions. Michael C. Freund: Conceptualization, Visualization, Formal analysis, Writing - original draft, Writing - review & editing; Julie M. Bugg: Supervision, Writing - review & editing; Todd S. Braver: Conceptualization, Supervision, Funding acquisition, Writing - review & editing.

Correspondence concerning this article should be addressed to Michael C. Freund, 1 Brookings Drive, St. Louis, Missouri, USA 63130. E-mail: m.freund@wustl.edu

Abstract

Progress in understanding the neural bases of cognitive control has been supported by the paradigmatic color-word Stroop task, in which a target response (color name) must be selected over a more automatic, yet potentially incongruent, distractor response (word). From this platform, models of control have postulated complementary coding schemes: dorsomedial frontal cortex (DMFC) is proposed to evaluate control demand via incongruency-related coding, whereas dorsolateral prefrontal cortex (DLPFC) is proposed to implement control via goal and target-related coding. But, mapping these theorized schemes to measured neural activity within this task has been challenging. Here, we tested for these coding schemes relatively directly, by decomposing an event-related color-word Stroop task via representational similarity analysis (RSA). Three coding models were fit to the similarity structure of multi-voxel patterns of fMRI activity, acquired from 66 healthy, young-adult male and female humans. Incongruency coding was predominant in DMFC, whereas both target and incongruency coding were present with indistinguishable strength in DLPFC. Distractor coding, in contrast, was not detected within any frontoparietal region, but was instead strongly encoded within early visual cortex. Further, these coding schemes were differentially related to behavior: individuals with stronger DLPFC (and lateral posterior parietal cortex) target coding, but weaker DMFC incongruency coding, exhibited less behavioral Stroop interference. These results highlight the utility of the RSA framework for investigating neural mechanisms of cognitive control and point to several promising directions to extend the Stroop paradigm.

Significance Statement

How the human brain enables cognitive control — the ability to override behavioral habits to pursue internal goals — has been a major focus of neuroscience research. This ability has been frequently investigated by using the Stroop color-word naming task. With the Stroop as a test-bed, many theories have proposed specific neuroanatomical dissociations, in which medial and lateral frontal brain regions underlie cognitive control by encoding distinct types of information. Yet providing a direct confirmation of these claims has been challenging. Here, we demonstrate that representational similarity analysis (RSA), which estimates and models the similarity structure of brain activity patterns, can successfully establish the hypothesized functional dissociations within the Stroop task. RSA may provide a useful approach for investigating cognitive control mechanisms.

Keywords: executive function, response conflict, dorsolateral prefrontal cortex, dorsal anterior cingulate cortex, multivariate pattern analysis, fMRI

Word count: 10817

A representational similarity analysis of cognitive control during color-word Stroop

Goals, held in mind, can be used to overcome behavioral habits. Understanding how the human brain enables this capacity, which is typically termed *cognitive control*, has been a major interest of both basic and translational cognitive neuroscience. Towards this end, the use of *response conflict* tasks has been instrumental (e.g., Botvinick, Braver, Barch, Carter, & Cohen, 2001; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004). Response conflict tasks involve trials in which a less-automatic, but goal-relevant course of action, the *target* response, must be selected in the face of a habitual, but goal-irrelevant alternative, the *distractor*. The paradigmatic example is the color-word Stroop task (MacLeod, 1991; Posner & Snyder, 1975; Stroop, 1935): on each trial, the hue of a word must be named, despite the word expressing a potentially conflicting, that is, *incongruent*, color (Figure 1, C). A major goal in cognitive neuroscience studies of response conflict tasks such as Stroop has been to use measures of task-evoked neural activity to test theoretical models of cognitive control.

One broad, neurocomputational-level model ascribes particular roles to different frontoparietal regions in overcoming response conflict (Miller & Cohen, 2001; Shenhav, Botvinick, & Cohen, 2013). Central to this view is the type of task information these regions encode. The dorsomedial frontal cortex (DMFC), including dorsal anterior cingulate (DACC) and pre-supplementary motor area (pre-SMA), is proposed to "evaluate" demand for cognitive control, via encoding of *incongruency-related information* (Figure 1, A, bottom row). This evaluative computation is proposed to signal when (and how) the current selection policy is suboptimal (Ridderinkhof et al., 2004; Shenhav et al., 2013). Such information, according to this view, is used by dorsolateral prefrontal cortex (DLPFC), in concert with lateral posterior parietal cortex (LPPC), to "implement" control, via encoding of goal and *target-related information* (Figure 1, A, top row). This implementational process is proposed to guide the selection of the target over the distractor response (Buschman &

Miller, 2007; Miller & Cohen, 2001). Collectively, this model predicts key dissociations between medial and lateral frontoparietal cortex in terms of the type of information each region encodes and its behavioral relevance (Figure 1, B; Kane & Engle, 2002; Braver, 2012; MacDonald, Cohen, Stenger, & Carter, 2000). Yet, although this view has been influential, directly establishing these dissociations during the performance of standard color-word Stroop tasks has been difficult.

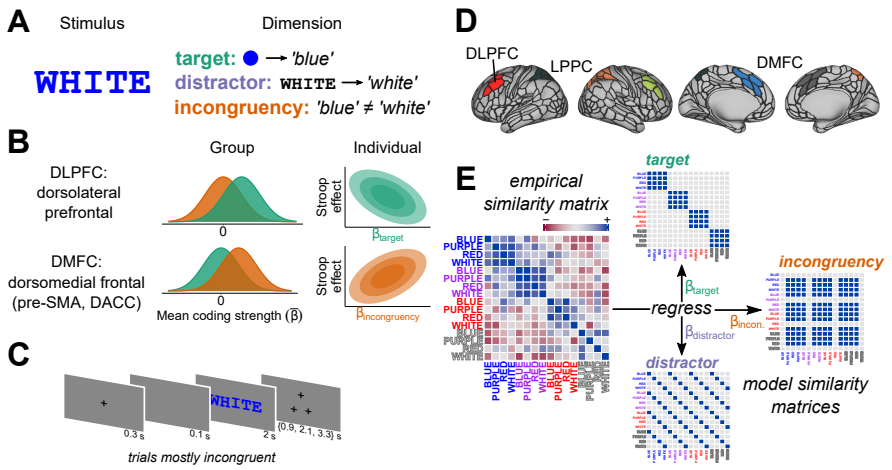

Figure 1. Schematic of framework and hypotheses. **A**, Conceptual framework. Cognitive theories and computational models of control have decomposed the classic color-word Stroop task into three *task dimensions*: the *target* (color-naming processes: the intentional mapping of the stimulus hue to the goal-relevant target response), the *distractor* (word-reading processes: the prepotent but goal-irrelevant mapping of the stimulus word to the non-target or distractor response), and *incongruency* (the potential match or mismatch between target and distractor dimensions). **B**, Hypotheses. Neuroscientific frameworks of cognitive control propose that representation of task dimensions is anatomically dissociated across medial and lateral frontoparietal cortices. Dorsomedial frontal cortex (DMFC), including dorsal anterior cingulate (DACC) and pre-supplementary motor area (pre-SMA), is proposed to "evaluate" demands for control, using information correlated with the *incongruency* dimension (**A**, bottom row), to determine when current attentional or action policies are suboptimal with respect to some criterion (e.g., a desired level of performance; Ridderinkhof et al., 2004; Shenhav et al., 2013). Conversely, dorsolateral PFC (DLPFC), in concert with lateral posterior parietal cortex (LPPC), is proposed to guide, or "implement", goal-driven attentional selection and mapping of hue–target-response processes, by way of representing information related to the goal-dependent target dimension (**A**, top row; Miller & Cohen, 2001; Buschman & Miller, 2007). Double dissociations are therefore predicted at multiple levels of analysis. At the group level, incongruency coding (orange univariate distributions) should predominate in DMFC, whereas target coding (green univariate distributions) should predominate in DLPFC. At the individual level, if the strength of target-related coding in DLPFC reflects the robustness of goal-driven selection, then subjects with stronger DLPFC target coding should resolve Stroop interference more efficiently (green bivariate distribution; Kane & Engle, 2002; Braver, 2012). Conversely, if the strength of incongruency-related coding in DMFC indicates a maladaptive selection policy, then subjects with stronger incongruency-related coding should resolve Stroop interference *less* efficiently (orange bivariate distribution; Braver, 2012; MacDonald et al., 2000). The β notation in axis titles corresponds to that used in **E** (and throughout this manuscript); $\bar{\beta}$ indicates mean over subjects. **C**–**E**, Analytic framework. Participants performed a color-word Stroop task while undergoing an fMRI scan (**C**). To derive neural correlates of the theorized task dimensions (in **A**–**B**), 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. We analyzed a mostly-incongruent trial condition, in order to obtain balanced estimates of each trial type. The Glasser et al. (2016) multi-modal atlas was used to parcellate cortex (**D**, light silver borders). Contiguous sets of parcels that tiled our regions of interest ("superparcels") were defined and treated as analytic units. Within each superparcel, linear correlations among response patterns from the sixteen stimuli were estimated to form an empirical similarity matrix (**E**, left). Through rank regression, these matrices were fit to three representational models (**E**, right), which corresponded to the three hypothesized dimensions of the Stroop task (**A**). The resulting β coefficients summarized the extent to which a parcel emphasized, within its distributed activity patterns, the representation of each unique task dimension. These β coefficients were used as the primary dependent variables in group-level analyses, and primary independent variables in individual-level analyses (e.g., as in **B**).

A source of this difficulty is due to the nature of the Stroop task itself. Testing these dissociations requires disentangling neural encoding of target and incongruency information from each other, and from the encoding of distractor information. But by design, these sources of information are perfectly spatiotemporally correlated in their presentation within standard Stroop paradigms. Under such conditions, classical functional neuroimaging techniques have limited ability to separately index these sources. While certain scalp-EEG methods have proven effective at disentangling early target versus distractor coding (Itthipuripat, Deering, & Serences, 2019), these lack the requisite anatomical precision. Conversely, the standard Stroop eludes more fine-grained methods of non-human primate neurophysiology, due to its linguistic nature and non-spatial manipulations of response prepotency.

To date, the most traction on this problem has been gained via fMRI designs that temporally dissociate presentation of task-rule and incongruency-related information, within hybrid task-switching–Stroop experiments (Floden, Vallesi, & Stuss, 2010; MacDonald et al., 2000), in which subjects were instructed prior to each Stroop trial about the task to perform (color-naming, word-reading). But, while these studies generally found supportive evidence for the key claims, results were subject to three notable limitations. First, these studies were likely underpowered for fMRI (e.g., N = 12 in Floden et al., 2010; N = 9 in MacDonald et al., 2000). This fact alone warrants a follow-up study. Second, it is unclear whether the hypothesized framework extends to the more-standard Stroop-task design, in which task rules are not explicitly instructed prior to each trial, but are instead internally maintained. Thus, it is unclear, for example, whether reliable goal-relevant DLPFC target coding is dependent upon explicit rule instruction prior to each trial. Third, the prior results do not speak to functional dissociations *within* a single Stroop trial, during which interference is actually experienced and resolved. It is therefore possible, for instance, that the role of DLPFC (or other frontoparietal regions) in Stroop is primarily preparatory, and is less critical during actual interference resolution.

To address this question, a neuroanatomically precise technique is needed that does not rely on temporal dissociations, but instead can read out multiple sources of task information that are simultaneously encoded within single, human brain regions of interest. Multivariate (multivoxel) pattern analysis (MVPA) of fMRI, in popular use for over a decade (Cox & Savoy, 2003; Edelman, Grill-Spector, Kushnir, & Malach, 1998; Haxby, Gobbini, Furey, & Ishai, 2001), accomplishes exactly this purpose. Surprisingly, however, these methods have not been brought to bear on the question of a functional dissociation between medial and lateral frontoparietal cortex in resolving Stroop conflict. We fill this gap in the literature by using Representational Similarity Analysis (RSA; Edelman et al., 1998; Kriegeskorte, 2008), a framework for MVPA, to test for dissociations in frontoparietal coding during Stroop-task performance (Figure 1, C–E). We conducted a retrospective analysis of data collected as part of the Dual Mechanisms of Cognitive Control project (Braver, Kizhner, Tang, Freund, & Etzel, 2020). Our primary goal was a proof-of-principle: to demonstrate that RSA has high potential for testing theorized distinctions in neural coding within cognitive control tasks such as the Stroop (Freund, Etzel, & Braver, 2020).

Materials and Methods

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2011).

Code, Data, and Task Accessibility

Code (R Core Team, 2019) and data to reproduce all analyses, in addition to supplementary analysis reports, are available on the first author's GitHub (https://github.com/mcfreund/stroop-rsa). As part of the planned data release of the Dual Mechanisms of Cognitive Control project, raw and minimally preprocessed fMRI data will be deposited on OpenNeuro in the near future. In the meantime, authors will share fMRI data directly upon reasonable request. Task scripts are available at the DMC project website (http://pages.wustl.edu/dualmechanisms/tasks). More detailed information regarding all aspects of the project can be found on the Project's OSF page (https://osf.io/xfe32/).

Participants

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 ($N = 66$; 39 women, 26 men, 1 "prefer not to answer") of these subjects: those with a full set of imaging and behavioral data from the Stroop task during a particular scanning session (the *proactive* session), selected for methodological reasons (see *Selection of Data*). Of the 66 subjects, 34 subjects were related as monozygotic twins (17 twin-pairs), and 32 were unrelated. We used this familial structure within our dataset to create two subsamples of subjects: a *primary analysis set* ($N = 49$; 27 women, 21 men, 1 "prefer not to answer"), which we used in all analyses, and a *validation set* ($N = 17$; 12 women, 5 men), which we used to assess the predictive accuracy of our selected brain–behavior model (see *Model Selection*). Within each subsample (*primary* and *validation*), all subjects were unrelated. The *primary analysis set* was composed of the 32 unrelated subjects and 17 co-twins. Two of these co-twins were selected for use in the *primary analysis set* as their respective co-twins had atypically high rates of response omission (>10%; >20% errors of any type); the remaining 15 co-twins were randomly selected. The *validation set* was composed of the 17 co-twins not selected for use in the primary analysis set.

As we performed this sorting of individuals into primary and validation sets only once and did not analyze the validation-set data except to assess predictive accuracy of the final selected model, the validation set provides an unbiased assessment of predictive accuracy, in the sense that no statistical "double-dipping" could have occurred. But because the sets are familially dependent, it is perhaps more accurate to consider validation-set analyses as assessing a kind of test–retest reliability (i.e., while eliminating the potential confound of practice effects), rather than providing an estimate of out-of-sample predictive accuracy.

Task

Participants performed the verbal color-word Stroop (1935) task. Names of colors were visually displayed in various hues, and participants were instructed to "say the name of the color, as fast and accurately as possible; do not read the word".

The set of stimuli consisted of two intermixed 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 design, forming 16 unique color-word stimuli within each set. The *mostly incongruent* set 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, bold Courier New font on a grey background (RGB = 191, 191, 191). Of stimuli within the *mostly incongruent* set, incongruent stimuli were presented to subjects more often than congruent stimuli (per block, proportion congruent = 0.25). *Unbiased* stimuli were presented with a balanced frequency (proportion congruent = 0.5). These manipulations of incongruency statistics are standard manipulations to elicit proactive control (Bugg, 2014; e.g., Gonthier, Braver, & Bugg, 2016) and were performed to investigate questions outside the scope of the current study.

Each trial (e.g., Figure 1, C) 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 (with uniform probability). Each of two scanning runs consisted of three blocks of 36 trials, intermixed with 4 resting fixation blocks, during which a fixation cross appeared for 30 s. This formed a mixed block-event design (Chawla, Rees, & Friston, 1999; Visscher et al., 2003; e.g., Dosenbach et al., 2006). Each of

the 16 *mostly incongruent* stimuli — that is, each unique colored word (e.g., "BLUE" displayed in red hue) — were presented in both runs. 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 fully randomized.

Selection of Data

We focused our fMRI pattern analyses solely on trials from the *mostly incongruent* stimulus set within a particular scanning session (the "proactive" session) of our Stroop task. This selection was made purely on the basis of methodological reasoning: these trials were the only set of trials within the larger Dual Mechanisms project in which each unique Stroop stimulus (i.e., one of the 16 color-word combinations) was presented an equal number of times (9) to each participant, constituting a balanced design.[1] Employing a balanced design is critical when contrasting linear correlations across conditions in low signal-to-noise ratio contexts (i.e., in event-related fMRI), as small differences in the amount of aggregation (number of trials) per condition can lead to large differences in the maximum correlation observable per condition. While other analytic strategies can handle unbalanced designs—namely, under-sampling conditions or using cross-validated measures of similarity—we preferred the current approach for the present study. These alternative strategies lead to higher variance similarity estimates, which could reduce statistical power.[2]

Display and Recording Systems

The experiment was programmed in EPrime 2.0 ("E-Prime 2.0," 2013), presented on a Windows 7 Desktop, and back-projected to a screen at the end of the bore for viewing via a mirror head-mount. Verbal responses were recorded for offline transcription and response-time (RT) estimation. The first 45 participants spoke into a MicroOptics

---

[1] 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.

[2] With only two scanning runs, using the cross-validated euclidean distance (which can be rendered comparable to linear correlation), for example, would have doubled the variance of our similarity measures (Diedrichsen, Berlot, Mur, Schütt, & Kriegeskorte, 2020).

MR-compatible electronic microphone (MicroOptics Technologies, Inc.); due to mechanical failure, however, we replaced this microphone with the noise-cancelling FOMRI III (OptoAcoustics, Ltd.), which subsequent participants used. A voice-onset processing script (from the MATLAB Audio Analysis Library) was used to derive response-time estimates on each trial via spectral decomposition (the accuracy of which was verified by manually coding response times from a subsample of subjects and ensuring the two methods gave similar estimates). Code for this algorithm is available within the Dual Mechanisms GitHub repository (https://github.com/ccplabwustl/dualmechanisms/tree/master/preparationsAndConversions/audio).

Importantly, we verified that the change in microphone did not induce confounding between-subject variance in response-time measures of interest. While response time estimates recorded via the Micro-Optics microphone tended to be slower ($b = 100.33, p = 0.01$) and more variable ($\chi^2_9 = 3720, p = 0$), the magnitude of the Stroop effect was not observably impacted by the microphone change ($b = -7.08, p = 0.58$).

Image Acquisition, Preprocessing, and General Linear Model

The fMRI data 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, no GRAPPA, ipat = 0), and subjected to the minimally pre-processed functional pipeline of the Human Connectome Project (version 3.17.0), outlined in (Glasser et al., 2013). More detailed information regarding acquisition can be found on the Project OSF site (https://osf.io/tbhfg/). All analyses were conducted in volumetric space; surface maps are displayed in figures only for ease of visualization.

After pre-processing, to estimate activation patterns, we fit a whole-brain voxel-wise general linear model (GLM) to blood-oxygen-level dependent (BOLD) timecourses in AFNI, version 17.0.00 (Cox, 1996). To build regressors of primary interest, we convolved with a hemodynamic response function [via AFNI's *BLOCK(1,1)*] 16 boxcar timecourses, each

coding for the initial second of presentation of a *mostly incongruent* stimulus that resulted in a correct response. We also included (1) two regressors [similarly created via *BLOCK(1,1)*] to capture signal associated with congruent and incongruent trials of non-interest (*unbiased* stimuli) that prompted correct responses, (2) an "error regressor" coding for any trial in which a response was incorrect or omitted (via *BLOCK*), (3) a sustained regressor coding for task versus rest (via *BLOCK*), (4) a transient regressor coding for task-block onsets [as a set of piecewise linear spline functions via *TENTzero(0,16.8,8)*], (5) six orthogonal motion regressors, (7) five polynomial drift regressors (order set automatically) for each run, and (8) an intercept for each run. These models were created via *3dDeconvolve* and solved via *3dREMLfit*. The data for each subject's model consisted of 2 runs $\times$ 3 blocks $\times$ 36 trials (144 from the *mostly incongruent* stimulus group, 72 from *unbiased*). Frames with FD $>$ 0.9 were censored.

Definition of Regions of Interest

Our primary hypotheses concerned a set of six anatomical regions — DMFC, DLPFC, and LPPC in each hemisphere. Consequently, our primary analyses employed a targeted ROI-based analysis approach. Rather than defining functional ROIs via a whole-brain searchlight, which has known issues (Etzel, Zacks, & Braver, 2013), we defined functional anatomic regions via a cortical parcellation atlas. We selected the MMP atlas (Glasser et al., 2016) for two reasons. (1) The atlas was developed recently via multi-modal imaging measures, and (2) individual MMP parcels are relatively interpretable, as they are heterogeneously sized and have been explicitly connected to a battery of cognitive tasks (Assem, Glasser, Van Essen, & Duncan, 2020), the canonical functional connectivity networks (Ji et al., 2019), and a large body of neuroanatomical research (Glasser et al., 2016). We used a volumetric version, obtained from https://figshare.com/articles/HCP-MMP1_0_projected_on_MNI2009a_GM_volumetric_in_NIfTI_format/3501911?file=5534024 (also available on the project GitHub repository; see *Code, Data, and Task Accessibility*). We then defined a set of six spatially contiguous

sets of MMP parcels (three in each hemisphere), which we refer to as "superparcels", that corresponded to each of our ROIs. For full superparcel definitions, see Figure A4. "DMFC" was defined as the four parcels covering SMA–pre-SMA and DACC. "DLPFC" was defined as the four parcels that cover middle frontal gyrus (i.e., mid-DLPFC). "LPPC" was defined as all parcels tiling IPS (both banks and fundus), from the occipital lobe to primary somatosensory cortex. The overwhelming majority of parcels that met these anatomical criteria were assigned, within a previous report, to the cinguloopercular (most of DMFC), frontoparietal (most of DLPFC), and dorsal-attention (most of LPPC) control networks (Ji et al., 2019). Further, these ROI definitions contain several parcels recently identified as corresponding to key nodes within the "multiple demand" network (Assem et al., 2020). To assess the robustness of our results to particular superparcel definitions, we additionally used alternative, more inclusive, superparcel definitions of DMFC and DLPFC (see Figure A4). For the brain–behavior model selection analysis (see *Model Selection*), we compiled a larger set of anatomically clustered MMP parcels, covering regions across the cortex (Table A8). Two additional, non-MMP ROIs were included in this set, in order to give better coverage of particular functional brain regions. A mask for ventral somatomotor cortex (the "SomatoMotor–Mouth" network) was obtained from the Gordon atlas (Gordon et al., 2016), as the MMP does not split somatomotor cortex into dorsal and ventral divisions. A mask for left ventral occipito-temporal cortex (encompassing the "visual word-form" area), was obtained using MNI coordinates $-54 < x < -30, -70 < y < -45, -30 < z < -4$, specified in a prior report (Twomey, Kawabata Duncan, Price, & Devlin, 2011). To remove cerebellar voxels from this ROI, we used the Deidricshen atlas (Diedrichsen, 2006) hosted by AFNI (https://afni.nimh.nih.gov/pub/dist/atlases/SUIT_Cerebellum/SUIT_2.6_1/).

Estimation of Coding Strength β

To estimate the regional strength of target, distractor, and incongruency coding, we used the representational similarity analysis framework (RSA; Kriegeskorte, 2008). The RSA framework consists of modeling the observed similarity structure of activation patterns with

a set of theoretically-specified model similarity structures (Figure 1, E). For a given subject and cortical region, fMRI GLM coefficients for each of the 16 conditions of interest (e.g., the word "WHITE" presented in blue hue) were assembled into a condition-by-voxel activity pattern matrix $\mathbf{B}$. The observed similarity structure was estimated as the condition-by-condition correlation matrix $\mathbf{R} = \mathrm{Cor}(\mathbf{B})$. Cell R_{ij} of this matrix gives the linear correlation observed between activity patterns evoked by conditions i and j. Model similarity structures were specified in this same correlation matrix form. The target model assumed that conditions (stimuli) with the same hue will evoke identical patterns, regardless of whether the words or congruency match (or mismatch). That is, if the hue of condition $i = j$, this model predicts $R_{ij} = 1$, otherwise 0. (In the "target" matrix in Figure 1, E, the only cells equal to 1 — i.e., blue cells — are those in which the stimulus hues match.) The distractor model assumed that conditions with the same word will evoke identical patterns, regardless of the hue or congruency. (If the word of condition $i = j$, $R_{ij} = 1$, otherwise 0. In the "distractor" matrix in Figure 1, E, the only cells equal to 1 are those in which the stimulus words match.) The incongruency model assumed that only conditions that were incongruent would evoke identical patterns, regardless of the hue or word. (If i and j are both incongruent, $R_{ij} = 1$, otherwise 0. In the "incongruency" matrix in Figure 1, E, the only cells equal to 1 are those which both stimuli are incongruent.) As a covariate of non-interest, we also included a model capturing similarity between congruent conditions (if i and j are both congruent, $R_{ij} = 1$, otherwise 0).

These four models were jointly fitted to the observed similarity structure from each region through multiple regression (ordinary least-squares), separately for each subject. The response vector $\mathbf{y}$ and design matrix of this regression were assembled in a series of steps. (1) The 120 unique off-diagonal elements of of each similarity matrix (one observed and four models) were extracted and unwrapped into vectors. (2) The four model similarity vectors were separately z-scored and assembled into columns. This formed the RSA design matrix. (3) The observed similarity vector was rank-transformed (Nili et al., 2014) then z-score

standardized, to form a vector $\mathbf{r}$. (4) The vector $\mathbf{r}$ was prewhitened to remove a specific nuisance component. 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 the first run versus six in the second (vice versa for the other half). As each scanning run contains a large amount of run-specific noise (Alink, Walther, Krugliak, Bosch, & Kriegeskorte, 2015; Mumford, Davis, & Poldrack, 2014), this imbalance across runs could lead to a bias in the resulting β coefficients, in which pattern similarity of stimuli that mostly occurred within the same run would be inflated. We formalized this component of "run bias" as another model similarity vector, $\mathbf{u}$, with elements equal to 1 if *the run in which condition i most frequently occurred = the run in which condition j most frequently occurred*, otherwise 0. The magnitude of this bias was estimated as the slope term b_1 in a linear regression $\mathbf{r} = \mathbf{u}b_1 + b_0 + \boldsymbol{\epsilon}$, where b_0 is the intercept coefficient and $\boldsymbol{\epsilon}$ is the residual vector. The model $\mathbf{u}$ was scaled by its magnitude then subtracted from $\mathbf{r}$, forming the RSA response vector $\mathbf{y} = \mathbf{r} - \mathbf{u}b_1$.

Thus the RSA regression yielded three β coefficients of interest: β_{target}, $\beta_{\text{distr.}}$, $\beta_{\text{incon.}}$. These coefficients can be understood as a (standardized) contrast on (rank-transformed) correlations of activity patterns, between conditions in which only one task dimension was shared (e.g., the target dimension for β_{target}), versus those in which no dimensions were shared (i.e., different levels of target, distractor, and congruency).

Dimensionality Reduction

We used non-metric multidimensional scaling (Kruskal, 1964), a flexible, non-parametric dimensionality reduction technique, to visualize the structures of activity patterns within selected regions (Figure 2): ventral somatomotor cortex (corresponding to "mouth" homunculi), primary visual cortex (V1), and our (left) DMFC superparcel. These parcels were selected to highlight coding of each task dimension. For each selected region, we

averaged observed similarity matrices across subjects, then subtracted these values from 1 to obtain a dissimilarity matrix. Before averaging, we applied Fisher's z-transform (inverse hyperbolic tangent, artanh), which was inverted after averaging (tanh). (In contrast to the RSA regression above, we did not rank-transform similarity matrices, as non-metric multidimensional scaling incorporates a monotonic regression). That is, for subject s in $1, \ldots, N$, $\bar{\mathbf{D}} = \mathbf{J} - \tanh[\sum_{s=1}^{N} \operatorname{artanh}(\mathbf{R}_s)/N]$, where $\mathbf{J}$ is a 16-by-16 matrix of all ones, and $\bar{\mathbf{D}}$ is the resulting mean dissimilarity matrix. Similar to our RSA, we regressed from each $\mathbf{R}_s$ the "run bias" model $\mathbf{u}$ prior to conducting this procedure (see step 4 in *Estimation of Coding Strength* β; here, however, an artanh transform was applied to the response vector, which was inverted with the prewhitened vector). Each $\bar{\mathbf{D}}$ was submitted to an implementation of Kruskal's non-metric multidimensional scaling, *vegan::metaMDS()* in R, to generate a 2-dimensional configuration (Oksanen et al., 2019).

Group-Level Dissociation Analysis

To test for regional dissociations in coding preferences, we fit a hierarchical linear model on RSA model fits (see Estimation of Coding Strength β) obtained from our three ROIs within each hemisphere, and for our three RSA models.

$$
\begin{aligned}
\beta_{srhm} =& [(\text{DLPFC} + \text{DMFC} + \text{LPPC})_r \cdot (\text{L} + \text{R})_h \cdot (\text{target} + \text{distr.} + \text{incon.})_m]\mathbf{b} + \\
& [(\text{DLPFC} + \text{DMFC} + \text{LPPC})_r \cdot (\text{target} + \text{distr.} + \text{incon.})_m]\mathbf{v}_s + e_{srhm} \\
\mathbf{v}_s \sim& \text{Normal}_{RM}(\mathbf{0}, \mathbf{\Sigma}_{RM \cdot RM}) \\
e_{srhm} \sim& \text{Normal}(0, \sigma^2)
\end{aligned}
$$

Lowercase subscripts indicate factors by which terms vary: s in $1, \ldots, N$ subjects, r in $1 \ldots, R$ ROIs, h in $1, 2$ hemispheres, and m in $1, \ldots, M$ models. Individual terms within brackets represent indicator variables and their interactions expand to form design matrices. Vector $\mathbf{b}$ indicates fixed-effect coefficients: across-subject means of each ROI×hemisphere×model term (here, $\mathbf{b}$ is of length $RHM = 18$). The set of vectors $\mathbf{v}_s$

indicate random-effect coefficients for each subject (i.e., conditional modes; here, each $\mathbf{v}_s$ is of length $RM = 9$). The full random effect covariance structure was estimated as $\mathbf{\Sigma}_{RM \times RM}$ (Barr, Levy, Scheepers, & Tily, 2013). Level-I reisduals indicated by e_{srhm}. This model was fit with *lme4::lmer()* in R (Bates, Maechler, Bolker, & Walker, 2014).

Planned contrasts on the fixed effects were performed to test our hypotheses. P-values were estimated using an asymptotic z-test, as implemented by the *multcomp::glht()* function in R (Hothorn, Bretz, & Westfall, 2008). We performed three types of contrasts: (1) to compare coding strengths within-region (e.g., DMFC: incon. − target), (2) to compare between regions, within-dimension (e.g., target: DLPFC − DMFC), and (3) to test their interaction [e.g., (target − incon.) · (DLPFC − DMFC)]. These contrasts were performed first by collapsing across hemisphere, then within each hemisphere separately. As we did not have any hypotheses regarding lateralization, p-values from hemisphere-specific contrasts were FDR corrected across hemispheres.

Comparison of Signal-to-Noise Ratios

To test for apparent differences in SNR (signal-to-noise ratios) between DMFC and DLPFC, we estimated "noise ceilings" within each region and contrasted them across regions. Noise ceilings indicate the maximum observable group-level effect size (RSA model fit) given the level of between-subject variability in similarity structure (Nili et al., 2014). Lower (smaller) average noise ceilings indicate poorer SNR for group-level tests. We used the "lower-bound" noise ceilings estimator of Nili et al. (2014).[3] For a given region, the lower-bound noise ceiling is defined for each subject s in $1, \ldots, N$ as $\text{Cor}(\mathbf{y}_s, \bar{\mathbf{y}}_{-s})$: the linear correlation between a subject s' observed similarity vector, $\mathbf{y}_s$, and the group-mean vector excluding subject s, $\bar{\mathbf{y}}_{-s}$. (See *Estimation of Coding Strength* β for definition of $\mathbf{y}$.) As each of the N estimates are interdependent, we used a percentile bootstrap to contrast noise

[3] We selected the lower-bound noise ceiling estimator because it yields a smaller-variance estimate, and therefore a more powerful contrast across regions, than the other noise ceiling estimator outlined in Nili et al. (2014), the "upper-bound" ceiling. This reduced variance is due to the lower-bound ceiling being a cross-validated estimator (which the upper-bound ceiling is not).

ceilings across regions. In each resample (of 10,000), noise ceilings were artanh transformed, contrasted across regions within-subject, then averaged across subjects. A two-sided p-value was provided by computing the proportion of resampled means greater than zero, p, then taking the minimum of $2p$ and $2(1-p)$. Finally, we conducted "two one-sided tests" for equivalence (Lakens, 2017; Schuirmann, 1987) to affirm a null hypothesis of no difference between regions in noise ceiling. This consists of defining a threshold effect size, the minimum effect size of interest, and testing whether the observed difference in noise ceilings is significantly less extreme than the threshold. For an objective threshold, we used the smallest standardized effect size at which our bootstrap procedure was expected to retain 80% power, Cohen's d = 0.35, determined via Monte Carlo simulation.

Selection of Behavioral Measures for Individual-Level Analyses

Audio recordings of verbal responses were transcribed and coded for errors offline by two researchers independently. Discrepancies in coding were resolved by a third. Errors were defined as any non-target color word spoken by a subject prior to utterance of the correct response (e.g., including distractor responses, but not disfluencies) or as a response omission. Trials in which responses were present but unintelligible (e.g., due to high scanner noise or poor enunciation) were coded as such.

We fit two hierarchical models on these data, one on errors and one on response times (RT). Several observations were excluded from these models. From the error model, only trials with responses coded as "unintelligible" were excluded (60). From the response time (RT) model, several types of trials were excluded: Trials with RTs greater than 3000 ms (2) or less than 250 ms (92; 90 of which were equal to zero). A cluster of fast and unrealistically invariable RTs from two subjects (23/216, 26/216) that were likely due to an artifact of insufficient voice-onset signal within the recording. Trials with a residual RT that was more extreme than 3 inter-quartile ranges from an initial multi-level model fitted to all subjects data (of the structure of Equation 1 below; *a la* Baayen & Milin, 2010). All trials with

incorrect (184), unintelligible (60), or no response (90). In total, 319 trials were excluded (0–62 per subject), leaving 14256 trials for analysis (154–216 per subject).

The RT model was of the form:

$$\begin{aligned} \mathrm{RT}_{sk} =& (\text{congr.} + \text{stroop})_k \mathbf{b} + (\text{congr.} + \text{stroop})_k \mathbf{v}_s + e_{sk} \\ \mathbf{v}_s \sim& \mathrm{Normal}_2(\mathbf{0}, \mathbf{\Sigma}_{2\times 2}) \\ e_{sk} \sim& \mathrm{Normal}(0, c_s \sigma^2) \end{aligned} \tag{1}$$

for s in $1, \ldots, N$ subjects, k in $1 \ldots, K$ trials, where congr. is an all-ones vector (intercepts), stroop is an indicator variable for incongruent trials, and c_s is a participant-specific parameter by which their residual standard deviation was scaled. Under this parameterization, the b coefficient on the stroop term reflects the size of the group-level Stroop effect (i.e., the mean difference in RT on incongruent versus congruent trials). (For description of remaining notation, see model formula within *Group-Level Dissociation Analysis*.) These additional scaling parameters (c_s) relaxed the homogneneity of level-II variance assumption — a well-warranted complexity, given the vastly improved fit of the heterogeneous-variance model ($\chi^2_{48} = 4299, p = 0, \Delta\mathrm{BIC} = -3856, \Delta\mathrm{AIC} = -4203$). To accomplish this, the RT model was fitted with *nlme::lme()* in R (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2019). The behavioral error model had similar predictors, but assumed binomially distributed error, with logit link function (fitted in *lme4*).

This hierarchical modeling framework enabled us to to estimate the amount and reliability of individual variability in the Stroop effect in each measure while accounting for trial-level error (e.g., Haines et al., 2020). We used this framework to validate that our behavioral measures of the Stroop effect met prerequisite properties for individual differences analyses. To assess the amount of variability, a nested model comparison was conducted, in which the models fitted above were compared to a random-intercept model. Stroop effects differed significantly across subjects in response time

($\chi^2_2 = 87.53, p = 0.00; \Delta\text{BIC} = -69.08; \Delta\text{AIC} = -83.53$), but not in accuracy ($\chi^2_2 = 0.74, p = 0.69; \Delta\text{BIC} = 18; \Delta\text{AIC} = 3$). We therefore did not further analyze accuracy. To assess split-half reliability (cross-run correlation), we fit a model with fixed and random-effect terms for $(\text{run1} + \text{run2})_k \cdot (\text{congr.} + \text{stroop})_k$ and a full $\mathbf{\Sigma}_{4\times4}$ covariance structure, which was used to obtain the cross-run correlation in Stroop effect. This analysis indicated Stroop effect RT estimates were highly consistent across scanning runs ($r = 0.95$). We therefore considered our measurements of RT to be adequate for individual differences analysis.

Individual-Level Dissociation Analysis

Similar to our *Group-Level Dissociation Analysis*, we tested our individual-level hypotheses within a hierarchical modeling framework. Preliminary analyses suggested error measures were inadequate for individual differences analyses (see *Materials and Methods*, *Selection of Behavioral Measures for Individual-Level Analyses*), so we focused solely on RT measures.

We began with the RT model described in Equation 1. However now, for a given RSA model and region of interest, we incorporated into the model each subject's estimated coding strength, βs, as subject-level fixed-effects. This was done by interacting the coding terms with the fixed effects: $[\beta_s \cdot (\text{congr.} + \text{stroop})_k]\mathbf{b}$. The model now estimates a cross-level, continuous-by-categorical fixed-effect interaction, $\beta_s \cdot \text{stroop}_k$, that describes how the Stroop effect (level-I) was modulated across subjects as a function of coding strength (level-II). To test our hypotheses, we performed contrasts on coefficients from these $\beta_s \cdot \text{stroop}_k$ interactions. These contrasts are outlined within the *Results* describing Figure 3. In short, the first contrast tested for differences in the magnitude of the interaction coefficients between target and incongruency coding models, within a given region: $(\beta_{\text{target}} - \beta_{\text{incon.}}) \cdot \text{stroop}$. The second tested for differences between regions, within a given coding model: for example, $(\beta_{\text{DLPFC}} - \beta_{\text{DMFC}}) \cdot \text{stroop}$.

Model Selection

To complement our hypothesis-driven brain–behavior analyses, we used a more data-driven model-selection approach. An expanded set of 28 superparcels (in addition to our six ROIs) was defined (see Table A8 for definitions). Some superparcels were included as regions of interest, others were included as negative controls (i.e., that should not be important for explaining behavioral performance). By-subject conditional modes of the Stroop effect from the behavior-only RT model were extracted and used as the response vector (i.e., in Equation 1, elements of $\mathbf{v}_s$ that correspond to the Stroop contrast). RSA was conducted on each superparcel (see *Estimation of Coding Strength* β), furnishing three β coefficients per superparcel. These 84 measures were fitted to the response vector via elastic net regression, implemented via *glmnet::glmnet()* in R (Friedman, Hastie, & Tibshirani, 2010). Parameter α was set to 0.5 (balancing lasso and ridge penalties). The parameter λ was tuned via 10-fold cross-validation (default) using the "1-SE rule": the smallest λ greater than one SE of the minimum λ across folds was saved (Hastie, Tibshirani, & Friedman, 2009); this routine was repeated 1,000 times and the minimum mode (there were ties) of the saved λs was selected. We selected the minimum mode because the maximum suppressed all variables from the model.

To assess validation-set accuracy, the selected model coefficients were applied to a design matrix from validation-set subjects, generating a predicted Stroop effect vector. The linear correlation was estimated between this predicted Stroop effect and observed Stroop effects (estimated as conditional modes via a hierarchical model separately fitted to validation-set data). The significance of this correlation was assessed by randomly permuting the training-set response vector, re-fitting the model, generating new predicted validation-set values, and re-estimating the predicted–observed correlation 10,000 times. The p-value was given by the proportion of resamples in which the null correlation was greater than the observed correlation.

Exploratory Whole-Cortex RSA

The RSA-model fitting procedure, as outlined in *Estimation of Coding Strength* β was separately conducted on each MMP cortical parcel. Inferential statistics followed those suggested by Nili et al. (2014). One-sided signed-rank tests were conducted on the resulting βs (>0). P-values were FDR-corrected over all 360 parcels, separately for each task dimension (target, distractor, incongruency).

RSA on Head Motion Estimates

As a negative control analysis for our exploratory whole-cortex RSA, we attempted to decode task variables (target, distractor, incongruency) via RSA from framewise estimates of head motion. The 6 motion regressors that were used in the fMRI GLM as nuisance covariates (corresponding to translation and rotation in 3 dimensions) were regressed upon the design matrix containing predicted BOLD timecourses of our 16 conditions of interest. The coefficient matrix resulting from this regression was then submitted to the RSA procedures described in *Estimation of Coding Strength* β and *Exploratory Whole-Cortex RSA*. To check whether more aggressive movement denoising within the fMRI GLM was warranted (i.e., in addition to the 6 nuisance regressors), we conducted this same movement-based RSA, however using 12 motion regressors (the 6 bases and their temporal derivatives). RSA model fits between the 6 and 12-basis motion-based RSAs were compared via paired-sample signed-rank test.

Results

Influential theories of cognitive control have proposed specific dissociations in the type of task information encoded by human medial and lateral frontoparietal cortex (Figure 1, A–B). But, previous studies have largely approached this question indirectly, by employing tasks designed to recruit these regions differentially in time, then testing for temporal dissociations in regional-mean levels of fMRI activity. Here, we used a more direct approach, by analyzing the similarity structure of neural activity patterns evoked within these regions in terms of their informational content. In particular, through RSA, we compared neural

coding of three distinct types of Stroop-task information — target, distractor, and incongruency (Figure 1, D–E) — within each region of interest (ROI) simultaneously, while Stroop interference was being experienced and resolved.

We describe three sets of analyses. First, we examined group-level effects, to test for neuroanatomical dissociations in encoding of task information (Figure 1, B, center). Second, we examined individual-level effects (i.e., individual differences), to test for dissociations in brain–behavior relationships (Figure 1, B, right). These two analyses were ROI-based and primarily focused on dorsomedial frontal and lateral frontoparietal regions (Figure 1, D), but also included sensorimotor ROIs for comparison purposes. The last set of analyses was conducted in whole-cortex exploratory fashion, to provide a more comprehensive picture of the anatomical profile of each task dimension.

Group

Dorsomedial frontal and dorsolateral prefrontal cortex exhibit distinct coding profiles. Primary group-level results are summarized in Figure 2. Statistical estimates corresponding to results outlined within this section are contained in Tables A1–A3.

The DMFC has been strongly associated with the coding of incongruency information in response conflict tasks such as the Stroop. RSA approaches were used to directly test the specificity of that hypothesis. This region was indeed found to encode the incongruency dimension of the Stroop task (left: $\bar{\beta} = 0.11, p = 0.00$, right: $\bar{\beta} = 0.08, p = 0.00$; Table A1). Further, within left DMFC, a preference for this dimension was observed: incongruency information was encoded more strongly than either target or distractor information (vs. target: $\Delta\bar{\beta} = 0.07, p = 0.05$; vs. distractor: $\Delta\bar{\beta} = 0.10, p = 0.00$; p-values adjusted across hemisphere; Figure 2, A; Table A2). This DMFC preference was prominent enough that it could be seen in its overall similarity structure within a low-dimensional embedding (Figure 2, B). Further, incongruency coding was neuroanatomically dissociated, as this coding

scheme was reflected more strongly in left DMFC than in DLPFC activity patterns ($\Delta\bar{\beta} = 0.06, p = 0.02$; Table A3).

We then focused on lateral frontoparietal regions and the coding of target information. DLPFC indeed encoded target information (left: $\bar{\beta} = 0.03, p = 0.01$ right: $\bar{\beta} = 0.04, p = 0.00$; Table A1), and more strongly than distractor information ($\Delta\bar{\beta} = 0.04, p = 0.00$; Table A2). But, we did not find a preference for target over incongruency information in DLPFC ($\Delta\bar{\beta} = -0.01, p = 0.62$), nor did we observe significantly stronger target coding than in DMFC ($\bar{\beta} = 0.00, p = 0.92$; Table A3). A qualitatively similar pattern of results was observed in lateral posterior parietal cortex: significant target and incongruency coding, significantly enhanced target versus distractor coding, but no observed preference among target or incongruency nor detectable difference from DMFC target coding (Tables A1–A3).

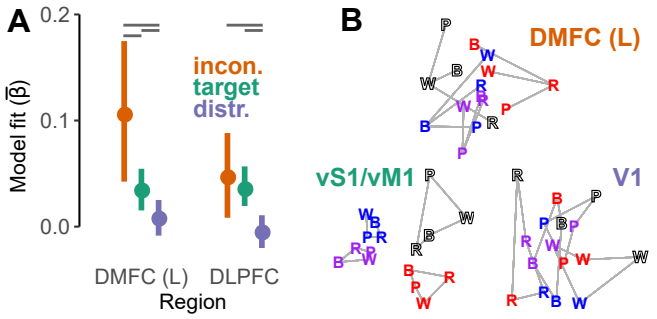

Figure 2. Group-level results. **A**, A single dissociation between dorsolateral versus dorsomedial frontal cortex (DLPFC, DMFC). For simplicity of display, DLPFC estimates are averaged across hemisphere (see Table A1 for per-hemisphere means for each region). Error bars are bootstrapped 95% bias-corrected and accelerated confidence intervals of between-subject variability. Horizontal grey lines at the top indicate significance of within-subject significance tests (see Table A2 for contrast estimates). While all ROIs encoded target and incongruency information, and did so more strongly than distractor information (coding of which was not detected), incongruency coding was stronger in left DMFC versus DLPFC. **B**, Across-voxel activity patterns from three select regions, embedded within a two-dimensional space (via non-metric multidimensional scaling). These geometries exemplify regional dissocations in coding of Stroop-task dimensions. Letters represent the first letter of the corresponding distractor word (BLUE, PURPLE, RED, WHITE), and colors represent the target hue (hollow letters indicate white-hue stimuli). Connecting lines are drawn to highlight the various structures present. The geometry of (left) dorsomedial frontal cortex (DMFC) is marked by a radial separation of patterns evoked by congruent and incongruent conditions (lines connect congruent and incongruent stimuli of each target color), such that incongruent patterns tend to be located more centrally (i.e., more similar to each other), whereas congruent patterns peripherally diverge (i.e., less similar to each other). This suggests that incongruent trials drove a common component of activation in this region, regardless of target or distractor features. In contrast, within primary ventral somatomotor cortices (vS1/vM1), patterns strongly cluster by target level (color), whereas in primary visual cortex (V1), they tend to cluster by distractor level (word).

Thus, at the group-level, we observed a single rather than double dissociation between medial and lateral frontoparietal cortex, in the form of enhanced sensitivity to incongruency relative to target coding in left DMFC. In all ROIs, however, these two sources of task

information were more strongly encoded than distractor information.

Sensitivity and control analyses. Because of the failure to observe a full double-dissociation, we tested a series of supplementary hypotheses.

First, we conducted sensitivity analyses of our results to the particular ROI definitions used. In one analysis, we tested more expansive ROI definitions, by using alternatively-defined superparcels (Figure A4). These definitions included additional, more rostral PFC parcels (1 in DMFC, 3 in DLPFC), which begin to encroach into ventromedial PFC and frontopolar cortex (e.g., the rostral DMFC parcel was assigned to the Default-Mode network within the Cole-Anticevic divisions). Nevertheless, the previously observed dissociations were robust to these more liberal definitions (Table A4). In the other analysis, we examined whether the overall superparcel coding profiles were representative of individual parcels. While DMFC and DLPFC results generally reflected that of constituent parcels (Figure A2, A, B), interestingly, there was substantial heterogeneity within left LPPC (Figure A2, C). Similar to left DMFC, a collection of left LPPC regions spanning the length of intraparietal sulcus strongly encoded the incongruency dimension (i.e., IP1, IP2, IPS1, AIP, LIP, MIP).

Second, we conducted a positive control analysis to bolster confidence in the statistical power of our method. With this analysis, we sought to determine whether our methods could detect dissociations in task coding that are strongly expected to exist. In particular, we focused on primary somatomotor and visual cortical ROIs, the responses of which can be assumed to reflect, relatively selectively, response-related (i.e., motoric) and visual form-related coding. As the distractor (word) defines the visual form of the Stroop stimulus, coding of form-related features should be captured by our distractor model. In parallel, as our analysis included only correct-response trials (i.e., in which the target response was spoken), coding of motoric features should be captured by our target model. Consistent with this logic, within early visual cortex, we found clear evidence of preferential distractor coding

(Figure A1, red line), whereas within primary ventral somatomotor cortices (encompassing the "mouth" homunculi), we found a relatively selective pattern of target coding (Figure A1, blue line). Further, coding of these dimensions was strong enough to clearly predominate the overall structure of patterns within a low-dimensional embedding (Figure 2, A, green and purple circles). Thus, in primary visual and somatomotor regions, a relatively clear-cut group-level double-dissociation emerged. This suggests that our models were adequately powered, at least in primary sensorimotor cortices, to detect functional distinctions, and that the failure to observe distractor coding in DMFC and DLPFC was not due to a general deficiency in our distractor coding model.

Third, we examined whether the lack of observed discrimination between target and incongruency coding dimensions within DLPFC could be explained by increased error variance potentially present in fMRI activity patterns within this region. Prior work has suggested that PFC regions might be particularly susceptible to this confound (Bhandari, Gagne, & Badre, 2018). It is possible that we might have observed a dissociation in left DMFC but not DLPFC due to differential levels of statistical power across the two regions. We therefore derived an SNR (signal-to-noise ratio) analysis to determine whether this was a viable explanation (see Method). A paired-sample bootstrap test did not indicate a systematic difference between DLPFC versus left DMFC group-level SNR ($\Delta z = -0.01, p = 0.59$). However, the SNRs in these regions were also not confirmably similar, as indicated by an equivalence test (which provides a confirmatory test of the null hypothesis; $p = 0.10$). Therefore, while we cannot rule out the possibility that the single dissociation observed between left DMFC and DLPFC was driven by better group-level SNR in DMFC, any potential SNR differences between the two regions were not substantial enough for our methods to detect.

Individual

Primary individual-level results are summarized in Figure 3. Statistical estimates corresponding to results outlined within this section are contained in Tables A5–A7.

Better-performing subjects have stronger lateral frontoparietal target coding. The fidelity of target-related information in lateral frontoparietal cortex — DLPFC, in particular — is thought to be closely linked to the efficiency with which an individual resolves response conflict in tasks such as Stroop. By using subject-level target-coding estimates (β_{target}) to model behavioral performance (RT), we tested this fundamental prediction relatively directly. Indeed, subjects with stronger target coding in both right DLPFC and right LPPC (Figure A3) resolved Stroop interference effects more quickly ($\beta_{\text{target}} \cdot \text{stroop}$, DLPFC: $b = -239.13, p = 0.00$; LPPC: $b = -235.34, p = 0.00$; p-values corrected across hemisphere; Figure 3, A; Figure A3; Table A5). Subjects' target coding estimates were moderately correlated ($r = 0.53$) between these two lateral frontoparietal regions, suggesting these measures indexed a common component of individual variability. In neither of these lateral regions was incongruency coding significantly related to the Stroop interference effect ($\beta_{\text{incon.}} \cdot \text{stroop}$, DLPFC: $b = 38.42, p = 0.25$; LPPC $b = -32.40, p = 0.31$; p-values uncorrected), and within right DLPFC, target coding was a considerably stronger explanatory variable than incongruency coding ($\Delta b = -299.21, p = 0.00$; Table A6; A3).

Conversely, incongruency-related responses of DMFC are thought to be positively associated with maladaptive policies of response selection. In line with this notion, subjects with stronger incongruency coding in left DMFC tended to exhibit greater Stroop interference, although this interaction was non-significant ($\beta_{\text{incon.}} \cdot \text{stroop}$: $b = 47.66, p = 0.13$; p-value corrected across hemisphere). Similarly, incongruency coding was numerically more strongly associated with Stroop interference effects than target coding in this region, though this statistical difference was also non-significant ($\Delta b = -20.92, p = 0.83$; Table A6; Figure A3). While weak, this DMFC–Stroop association

notably emerged within the same hemisphere that displayed a group-level preference for incongruency information (Figure 2, B).

Considering the collective pattern of results, we conclude here that our findings support the hypothesis that target coding in (right) DLPFC and in LPPC were linked, in some manner, to a common process of implementing control, and that this association reflects a process discrepant from (1) that in which (left) DMFC appeared to be involved and (2) that indexed by (right) DLPFC incongruency coding.

Model selection affirms a lateral–medial dissociation and identifies unexpected relationships. Because the preceding hypothesis-driven analysis exclusively focused on a limited set of regions, important brain–behavior relationships may have been missed. A more accurate model may even omit target and incongruency coding from DLPFC, DMFC and LPPC altogether. Consequently, we conducted a more comprehensive test to identify regions and task dimensions that could better account for Stroop performance variability across individuals.

A data-driven model selection analysis was conducted to address this question (see Method). We defined an expanded set of 28 cortical regions (superparcels), including the six defined and used earlier, which covered various areas that may be important for performing the Stroop task (Table A8 and Figure A5). Conducting RSA on each superparcel furnished three coding estimates (one per coding model) per superparcel. These 84 estimates were then used as features in a cross-validated model selection procedure.

Strikingly, the selected model contained all three hypothesized measures: (right) DLPFC and LPPC target coding, and (left) DMFC incongruency coding. (See also Table A9 for the largest 20 bivariate brain–behavior correlations observed across the full set of predictor variables.) In addition, two unexpected measures were identified, both with negative slopes: left DLPFC distractor coding and left ventral visual incongruency coding (bivariate correlations shown in Figure 3, B). We tested the predictive accuracy of the

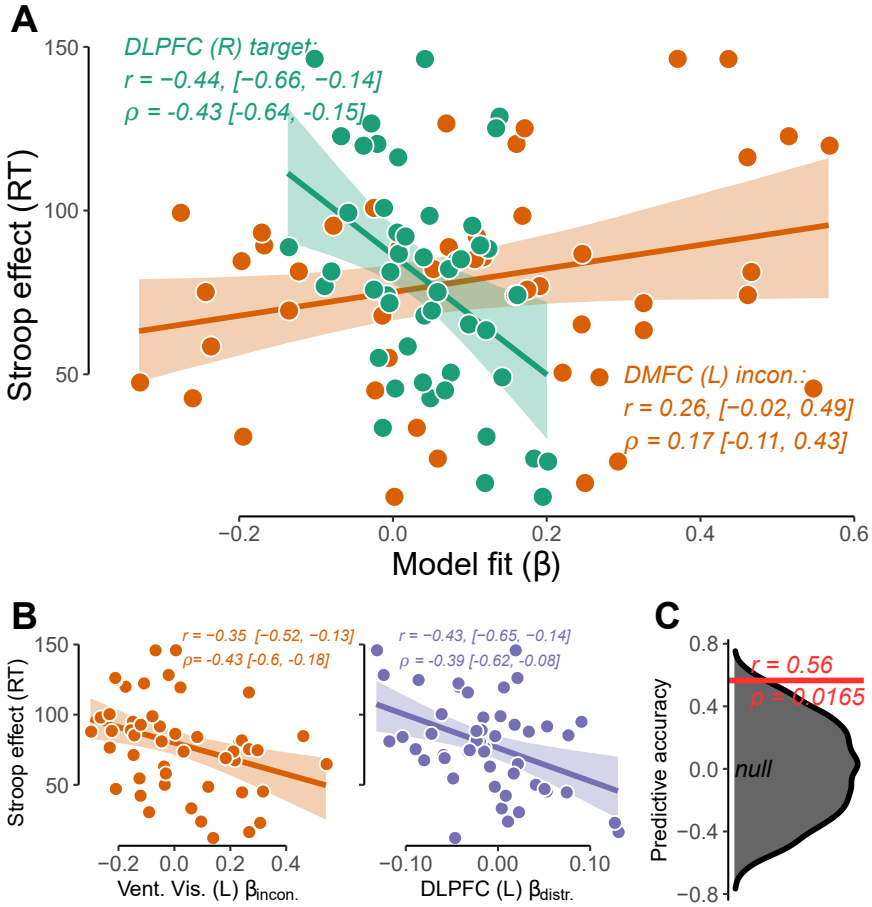

Figure 3. Individual-level results. **A**, The coding strength of task variables (β) within right dorsolateral prefrontal cortex (DLPFC) and left dorsomedial frontal cortex (DMFC) accounted for individual differences in the size of the Stroop interference effect (response time, RT) in hypothesized ways. (See also the hypothesized brain–behavior association observed within right lateral posterior parietal cortex, LPPC, in Figure A3). Inset text displays correlation coefficients (r, linear; ρ, rank) and bootstrapped, bias-corrected and accelerated 95% confidence intervals. Lines indicate a regression line of RT onto Model Fit (β), and confidence bands indicate 95% confidence intervals of predicted values from this regression obtained via percentile bootstrap. For display of relationships between target and incongruency coding in all ROIs (DLPFC, DMFC, LPPC) and Stroop effect, see Figure A3. **B**, Additional brain–behavior associations identified via a cross-validated model selection procedure conducted on a more inclusive feature set (84 features = 28 cortical regions × 3 coding measures). Left ventral visual incongruency coding (left panel) and left DLPFC distractor coding (right panel) were selected — in addition to right DLPFC target, right LPPC target, and left DMFC incongruency. See caption **A** for description of inset text, lines, and bands. **C**, Estimate of selected model prediction accuracy. Within a validation set of 17 subjects (monozygotic co-twins of subjects within the analysis sample), observed Stroop effects were moderately correlated to those fitted from the selected (held-out) model (red line). A permutation test indicated the selected model significantly explained validation-set variance, suggesting the selected model results are relatively stable.

selected model by using a held-out validation set of 17 subjects (co-twins of the primary analysis set). The selected model explained significant variance in Stroop interference effects within the validation set (Figure 3, B). While not fully independent, these validation set data were from a true hold-out and obtained from distinct individuals. This result therefore bolsters claims regarding the stability of the selected model.

Results from this model selection analysis affirm a functional dissociation across the medial–lateral axis of frontoparietal cortex, and further demonstrate that Stroop-task representations within DLPFC and DMFC hold relatively privileged relationships with behavior.

Exploratory

In a final, exploratory analysis, we conducted a whole-cortex RSA, separately for each MMP parcel, to determine more comprehensively how target, distractor, and incongruency coding are distributed across cortex. These three task dimensions were encoded across cortex according to different neuroanatomical profiles. Target coding was widespread (observed in 207/360 parcels), covering substantial portions of the frontal and temporal lobes, including many perisylvian regions (Figure 4, top; Table A10). Notably, the strongest target coding was observed within regions that receive strong sensory and (or) motor-related input (Figure A6). Contrastingly, incongruency coding was detected predominantly within prefrontal and intraparietal sulcal parcels— including DMFC, but also left LPPC, bilateral superior frontal gyrus, and left lateral frontopolar cortex (rostral DLPFC)—but additionally within left retrosplenial and right lateral occipital cortex (Figure 4, middle; Table A12). Aside from this occipital area, these incongruency-coding parcels notably belonged to control networks (frontoparietal, cinguloopercular, and dorsal attention; Table A12). In a third, distinct profile, distractor coding was only observed within early visual cortex (left V1 and V2; Figure 4, bottom; Table A11).

Finally, as a negative control analysis, we tested whether we could decode these three

task variables (target, distractor, incongruency) from framewise head motion estimates, using the same RSA procedures as above (see Table A13). No task variable was significantly encoded within patterns of head movements (bs < 0.01, ps > 0.36). Using a larger basis set of motion estimates (12 vs 6) did not yield significant decoding (bs < 0.01, ps > 0.17), nor any increased sensitivity to task variables (bs < 0.00, ps > 0.17), suggesting that our motion removal procedures (scrubbing and 6 motion regressors) were adequate.

Collectively, these exploratory results confirm and extend our prior findings. (1) As with the reported brain–behavior associations, target coding was emphasized relative to coding of other task dimensions. (2) Yet despite this emphasis, important and expected dissociations in anatomical profiles were identified across our three coding models, further suggesting that these models were successful in indexing neural correlates of distinct cognitive processes (Figure 1, A).

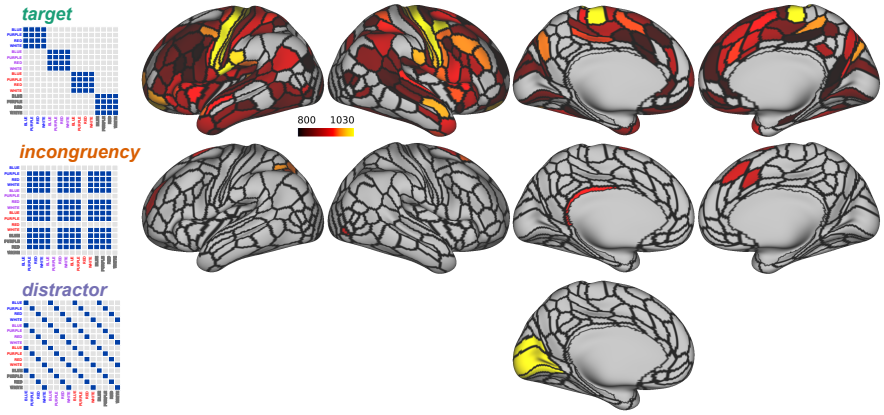

Figure 4. Cortical distributions of target (top), incongruency (middle) and distractor (bottom) coding, identified via an exploratory, whole-brain analysis. Color bar indicates test statistic from one-sided sign-ranked test.

Discussion

We analyzed the similarity structure, or representational geometry (Kriegeskorte & Kievit, 2013), of frontoparietal activity patterns evoked during a color-word Stroop task, as subjects selected target responses amidst prepotent and potentially distracting alternatives. In left DMFC, incongruency coding predominated. While the geometries of DLPFC and LPPC encoded both target and incongruency-related information, distractor coding was not

detected, but was instead reliably identified in early visual cortex. Further, these neural coding estimates were important and specific indicators of individual differences in magnitude of the behavioral Stroop interference effect. Individuals with stronger target coding in right DLPFC and right LPPC, but weaker incongruency coding in left DMFC, exhibited enhanced cognitive control, in terms of a reduced Stroop effect. In fact, when we more comprehensively examined a wide set of various cortical regions, coding measures specifically from lateral frontoparietal and dorsomedial frontal regions were privileged in their link to behavior.

On one level, this study is a specific extension of research that has drawn dissociations between control-related functions of DLPFC and DMFC (Floden et al., 2010; MacDonald et al., 2000). Most prominently, MacDonald et al. (2000) used a modified Stroop-task design, in which task rules (delivered via pre-cue) randomly alternated between color naming and word reading across trials, to demonstrate that DLPFC and DMFC encoded different types of information during cognitive control engagement. In particular, DLPFC was selectively recruited following cues for the more demanding color-naming task, whereas DMFC was instead driven by incongruent color-naming trials. This pattern of recruitment suggested that DLFPC encodes task-set and rule-related information in a preparatory manner, whereas DMFC encodes incongruency and conflict-related information in a stimulus-evoked manner. Here, however, by leveraging the high spatial dimensionality of fMRI, we tested whether this functional dissociation can be observed *within* a common time window of response selection, and further with a more traditional Stroop-task design, which does not involve task cues or switches. Our findings reinforce the conclusions of these relatively low-powered studies (N = 12 in Flooden et al., 2010; N = 9 in MacDonald et al., 2000) and indicate that the dissociations were not dependent upon the use of cued task-switching designs. In synthesis with these prior findings and more recent perspectives (Mante, Sussillo, Shenoy, & Newsome, 2013; Rigotti et al., 2013; Stokes et al., 2013), our present findings suggest a continuity in the putative role of DLPFC throughout the course of selection events. Rather than exclusively

contributing to preparation, DLPFC coding may evolve from representing abstract rule or set-related information, towards more concrete targets and behavioral choices, as evidence becomes available within the environment and is integrated into the code.

More broadly, the results of this study highlight the utility of RSA and the general representational geometric framework for investigating cognitive control. While previous work has used MVPA decoding to study the impact of control demand on posterior representations (e.g., Banich et al., 2019), we used the RSA framework to explicitly model control-related frontoparietal representations. Our results suggest that the RSA framework holds promise in furnishing measures of frontoparietal coding that can map relatively straightforwardly onto theorized neurocognitive mechanisms of control. In fact, a major motivation of the current study was to assess the validity of this method–theory mapping. For this purpose, the medial–lateral functional dissociation in frontoparietal cortex was a useful test-bed, as it features in several theoretical accounts (Botvinick et al., 2001; Miller & Cohen, 2001; Ridderinkhof et al., 2004; Shenhav et al., 2013). In line with these accounts, the prime coding measures we obtained from these regions were DLPFC target coding and DMFC incongruency coding. At the group level, each coding scheme was enhanced relative to other schemes (target over distractor in DLPFC, incongruency over target in DMFC); likewise, between individuals, variation in the strength of these schemes was associated with enhanced Stroop performance, in opposite but expected directions. These results join with a growing body of research (see Freund et al., 2020, for a review) in suggesting that RSA and the associated representational geometric approach can serve as an effective and relatively direct analytic framework from which to use neural measures to test cognitive control theory.

Nevertheless, the current work represents only an initial step in using this framework to investigate cognitive control within Stroop-like tasks. As such, our study raises a number of unaddressed questions. But promisingly, there are ample opportunities for improving and extending the RSA framework highlighted here. In particular, a key limitation of the current

study was the finding of widespread coding of the target dimension, suggesting a lack of specificity in this RSA model. This is perhaps not surprising, however, as the model would capture not only coding of attentional-template and choice-related information, but also hue and response-related information. We mitigated this issue by demonstrating that target coding was selectively related to behavior within our DLPFC and LPPC. But, this limitation could be addressed more powerfully by experimental design. For example, a design could be employed in which participants respond to the present set of stimuli, but under both color-naming and word-reading conditions (*a la* MacDonald, et al. 2000). Via RSA, rule-dependent coding of target features could now be isolated from rule-independent coding of stimulus hue, word, and response-related information, in addition to coding of abstract task rules (Figure 5, A, B; see also Hall-McMaster, Muhle-Karbe, Myers, & Stokes, 2019). By incorporating an additional manipulation of response modality, this response-related information could be further decomposed into abstract choice (effector independent) versus motoric coding (Figure 5, A, C; *a la* Minxha, Adolphs, Fusi, Mamelak, & Rutishauser, 2020). In these expanded designs, the interplay of the additional distinct coding schemes — within or between brain regions, with more classical neural correlates of control (e.g., incongruency response), or with behavioral measures — could then be properly modeled, perhaps jointly (Turner, Forstmann, & Steyvers, 2019) at the between-subjects or single-trial (Kikumoto & Mayr, 2020) levels. Thus, with appropriate experimental design, RSA and the representational geometric framework can provide an explicit, comprehensive framework for testing neurocognitive theories of cognitive and attentional control.

Future work could also address some of the complexities revealed by our data that were not entirely accounted for by the theoretical frameworks we used. For one, although theoretically-predicted coding profiles emerged in some frontoparietal regions of interest, all regions encoded target and incongruency information. One speculative interpretation is that, during the relatively late phase of response selection and execution, control networks may lose modular structure as the circuitry collectively converges on a behavioral choice. This

interpretation accords with the fact that "choice axes" are encoded within multiple key nodes of frontoparietal decision circuitry (e.g., in macaque LIP: Roitman & Shadlen, 2002; in macaque caudal DLPFC: Mante et al., 2013; in human DACC and pre-SMA: Minxha et al., 2020). This account could be addressed using the enriched experimental design described above, identifying when and where choice coding is emphasized over the course of a trial.

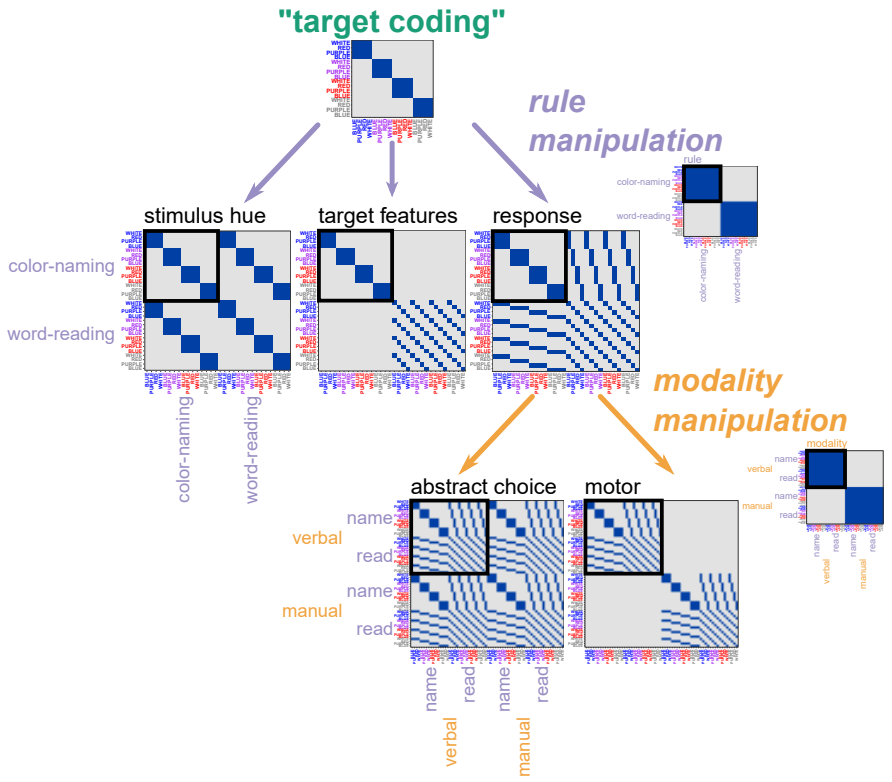

Figure 5. Expanded experimental designs. The current target coding model (top, green text; see also Figure 1) conflates distinct coding schemes, including those associated with relatively "early" vision, relatively "late" motoric commands, or more "central" rule-dependent schemes. By adding a factorial manipulation of task-rule (color naming or word reading; light purple arrows), the target coding model can be expanded into three more precise models (middle row of matrices), which are identical to our "target coding" model within the top-left quadrant (black square; naming–naming), but are now distinguished among the other quadrants (reading–reading, reading–naming). "Stimulus hue" and "response" models would identify coding that depends upon features of either the stimulus or response, independent of the task rules. Conversely, "target features" would identify a more flexible "attentional template" coding scheme, which depends critically on the current task goals (note that, the lower-right quadrant resembles our "distractor" coding model in Figure 1, as word features become the target in the word-reading condition). Importantly, too, abstract coding of task rule would also now be distinguishable ("rule" matrix on right). For an elegant example of a similar design within cued-task switching, see Hall-McMaster et al. (2019). The "response" model could be further elaborated by incorporating another manipulation of response modality (verbal, manual; orange arrows). Now, coding of true abstract choice options or decision variables, independent of effector, could be separated from "lower-level" motor-related (effector-dependent) coding. For an elegant example of a similar manipulation, see Minxha et al. (2020).

Another unexpected finding was a relatively robust negative relationship between the strength of incongruency coding in left ventral visual cortex and the magnitude of the Stroop effect (Figure 3, B, left panel). One interpretation of this finding is provided by the biased

competition framework, as an effect of selective visual attention. Prior work has demonstrated that certain ventral visual regions, those which are strongly tuned to target features, activate as a function of Stroop incongruency (Egner & Hirsch, 2005). In our task, mid-ventral stream areas may have received a volley of biasing input, selectively on incongruent trials. This biasing signal may have enhanced stimulus-related coding and communication with downstream regions. Using the expanded RSA design sketched above, we might expect that such an effect would be limited to color naming conditions, when selective attention processes would be most prominent. Further, the strength of visual incongruency coding might track, on a trial-by-trial basis or across subjects, the strength of frontoparietal "target feature" coding (see Figure 5).

Perhaps our most surprising result was the robust negative correlation observed between left DLPFC distractor coding and the behavioral Stroop effect (Figure 3, B, right panel). At face, accounting for this finding within the framework of top-down biased competition is difficult. But, given the statistics of our task, in which incongruent trials were frequent and congruent rare, distractor information could have been used to facilitate performance. An association between distractor features and incongruency could have been learned and used to influence response selection — for example, by retrieving and implementing a stimulus-appropriate attentional setting (Bugg & Crump, 2012; Melara & Algom, 2003). In fact, subjects clearly do exploit these associations, as indicated by reduced Stroop effects for stimuli that are "mostly incongruent", as in the robust "item-specific proportion congruency" effect (ISPC; Bugg, Jacoby, & Chanani, 2011; see also Crump & Milliken, 2009). The prediction that distractor coding might reflect ISPC effects could be tested by varying ISPC levels across different stimuli. For stimuli in which the specific color or word is not predictive of congruency, the relationship between DLPFC distractor coding and improved Stroop performance should not be present.

Finally, the present study sets the stage for using RSA to test the dual-mechanisms

framework of cognitive control (Braver, 2012). This framework explains much within and between-individual variability in cognitive control function by the existence of two operational "modes" of cognitive control: proactive and reactive. These modes are proposed to have dissociable signature neural coding schemes. Proactive control should rely heavily on task-relevant coding schemes that originate prior to target-stimulus onset in LPFC as abstract rule or context coding, but which may morph into target coding post-stimulus onset. In contrast, reactive control should rely on an incongruency-based coding-scheme, including coding of whichever task dimensions are predictive of incongruency, originating post-target onset, with potential loci in DMFC or subcortical structures (Chiu, Jiang, & Egner, 2017; Jiang, Brashier, & Egner, 2015). As suggested here, it may be possible to measure correlates of these neural coding schemes via RSA. Experimental factors that push subjects to adopt one mode over another (e.g., strategy training, expectancy of difficulty) should correspondingly shift frontoparietal coding schemes along these proactive and reactive dimensions. Further, their behavioral relevance should predictably change, as well: for example, in task contexts in which a proactive control mode is theoretically maladaptive, subjects with stronger proactive coding should perform worse. Thus, the dual-mechanisms framework suggests a broad range of hypotheses amenable to testing with RSA methods (see, e.g., Hall-McMaster et al., 2019). Such hypotheses can be addressed in the broader Dual Mechanisms of Cognitive Control dataset (Braver et al., 2020), of which the data used here are a small subset.

Acknowledgements

Funding support for this work was provided to TSB through NIH grant R37 MH066078.

We thank all former and current team members of the Dual Mechanisms of Cognitive Control Project for their efforts, Jo Etzel for general methodological wisdom, Atsushi Kikimoto for useful thoughts on our RSA models, Jeff Zacks for comments on an earlier

version of this manuscript, and members of the Barch, Braver, Bugg, and Kool laboratories at WUSTL for their support and suggestions. Mensh and Kording (2017) was a useful resource for organizing an initial draft of this manuscript.

Computations were performed in part using the facilities of the Washington University Center for High Performance Computing, which were partially funded by NIH grants 1S10RR022984-01A1 and 1S10OD018091-01. Several software applications were heavily used in this study. Software used to implement key analytic methods (cited in Materials and Methods). Several R libraries were heavily used that were not cited in the methods: *data.table*, (Dowle & Srinivasan, 2019), *doParallel* (Corporation & Weston, 2019), *dplyr* (Wickham, François, Henry, & Müller, 2020), *foreach* (Microsoft & Weston, 2019)„ *oro.nifti* (Whitcher, Schmid, & Thornton, 2011), *WRS2* (Mair & Wilcox, 2020). Figures were prepared with *ggplot2* (Wickham, 2016), *Inkscape* (Inkscape Project, 2020), and *Connectome Workbench* (Marcus et al., 2011). The original copy of this manuscript was drafted with *papaja* (Aust & Barth, 2020) and *knitr* (Xie, 2015).

References

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. https://doi.org/10/gfsh5f

Assem, M., Glasser, M. F., Van Essen, D. C., & Duncan, J. (2020). A Domain-General Cognitive Core Defined in Multimodally Parcellated Human Cortex. *Cerebral Cortex*, *30*(8), 4361–4380. https://doi.org/10/ghbh24

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12. https://doi.org/10/gddhjp

Banich, M. T., Smolker, H. R., Snyder, H. R., Lewis-Peacock, J. A., Godinez, D. A., Wager, T. D., & Hankin, B. L. (2019). Turning down the heat: Neural mechanisms of cognitive control for inhibiting task-irrelevant emotional information during adolescence. *Neuropsychologia*, *125*, 93–108.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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.

Bhandari, A., Gagne, C., & Badre, D. (2018). Just above Chance: Is It Harder to Decode Information from Prefrontal Cortex Hemodynamic Activity Patterns? *Journal of Cognitive Neuroscience*, *30*(10), 1473–1498. https://doi.org/10.1162/jocn_a_01291

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. https://doi.org/10.1037//0033-295X.I08.3.624

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

Braver, T. S., Kizhner, A., Tang, R., Freund, M. C., & Etzel, J. A. (2020). The Dual Mechanisms of Cognitive Control (DMCC) Project. *bioRxiv*, 2020.09.18.304402. https://doi.org/10/ghdrbt

Bugg, J. M. (2014). Conflict-triggered top-down control: Default mode, last resort, or no such thing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 567.

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*. https://doi.org/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. https://doi.org/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. https://doi.org/10/dw85gh

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

https://doi.org/10/d6hpfk

Chiu, Y.-C., Jiang, J., & Egner, T. (2017). The caudate nucleus mediates learning of stimulus–control state associations. *The Journal of Neuroscience*, *37*(4), 1028–1038. https://doi.org/10.1523/JNEUROSCI.0778-16.2017

Corporation, M., & Weston, S. (2019). *DoParallel: Foreach parallel adaptor for the 'parallel' package*. Retrieved from https://CRAN.R-project.org/package=doParallel

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2), 261–270. https://doi.org/10/d2xbcz

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

Crump, M. J. C., & Milliken, B. (2009). The flexibility of context-specific control: Evidence for context-driven generalization of item-specific control settings. *Quarterly Journal of Experimental Psychology*, *62*(8), 1523–1532. https://doi.org/10.1080/17470210902752096

Diedrichsen, J. (2006). A spatially unbiased atlas template of the human cerebellum. *NeuroImage*, *33*(1), 127–138. https://doi.org/10/dgzq5s

Diedrichsen, J., Berlot, E., Mur, M., Schütt, H. H., & Kriegeskorte, N. (2020). Comparing representational geometries using the unbiased distance correlation. *arXiv:2007.02789 [Stat]*.

Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., . . . Petersen, S. E. (2006). A Core System for the Implementation of Task Sets.

Neuron, *50*(5), 799–812. https://doi.org/10.1016/j.neuron.2006.04.031

Dowle, M., & Srinivasan, A. (2019). *Data.table: Extension of 'data.frame'*. Retrieved from https://CRAN.R-project.org/package=data.table

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309–231.

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

E-Prime 2.0. (2013). 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. https://doi.org/10.1016/j.neuroimage.2013.03.041

Floden, D., Vallesi, A., & Stuss, D. T. (2010). Task Context and Frontal Lobe Activation in the Stroop Task. *Journal of Cognitive Neuroscience*, *23*(4), 867–879. https://doi.org/10/b7wrdt

Freund, M. C., Etzel, J. A., & Braver, T. (2020). *Neural coding of cognitive control: The representational similarity analysis approach*. PsyArXiv. https://doi.org/10.31234/osf.io/ctw7k

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01/

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. https://doi.org/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. https://doi.org/10/f46nj4

Gonthier, C., Braver, T. S., & Bugg, J. M. (2016). Dissociating proactive and reactive control in the Stroop task. *Memory & Cognition*, *44*(5), 778–788. https://doi.org/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. https://doi.org/10.1093/cercor/bhu239

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., ... Turner, B. (2020). *Learning from the Reliability Paradox: How Theoretically Informed Generative Models Can Advance the Social, Behavioral, and Brain Sciences* (preprint). PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

Hall-McMaster, S., Muhle-Karbe, P. S., Myers, N. E., & Stokes, M. G. (2019). Reward boosts neural coding of task rules to optimize cognitive flexibility. *Journal of Neuroscience*, *39*(43), 8549–8561. https://doi.org/10/ggcf34

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* (2nd ed.). New York: Springer-Verlag.

Haxby, J. V., Gobbini, M. I., Furey, M. L., & Ishai, A. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science; Washington*,

293(5539), 2425–2430. https://doi.org/10/cffn26

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Inkscape Project. (2020). Inkscape (Version 0.92.5). Retrieved from https://inkscape.org

Itthipuripat, S., Deering, S., & Serences, J. T. (2019). When Conflict Cannot be Avoided: Relative Contributions of Early Selection and Frontal Executive Control in Mitigating Stroop Conflict. *Cerebral Cortex*, *29*(12), 5037–5048. https://doi.org/10/gfxb7d

Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., & Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, *185*, 35–57. https://doi.org/10/ggd8dm

Jiang, J., Brashier, N. M., & Egner, T. (2015). Memory meets control in hippocampal and striatal binding of stimuli, responses, and attentional control states. *Journal of Neuroscience*, *35*(44), 14885–14895. https://doi.org/10.1523/JNEUROSCI.2957-15.2015

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. https://doi.org/10/bwh9mt

Kikumoto, A., & Mayr, U. (2020). Conjunctive representations that integrate stimuli, responses, and rules are critical for action selection. *Proceedings of the National Academy of Sciences*, *117*(19), 10603–10608. https://doi.org/10/ggx4sq

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

https://doi.org/10.3389/neuro.06.004.2008

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/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. https://doi.org/10/dk9pcw

Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10/gbf8nt

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835–1838. https://doi.org/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. https://doi.org/10.1037/0033-2909.109.2.163

Mair, P., & Wilcox, R. (2020). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, *52*, 464–488.

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. https://doi.org/10/f5gdn6

Marcus, D., Harwell, J., Olsen, T., Hodge, M., Glasser, M., Prior, F., ... Van Essen, D. (2011). Informatics and data mining tools and strategies for the human connectome project. *Frontiers in Neuroinformatics*, *5*, 4.

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of stroop

effects. *Psychological Review*, *110*(3), 422.

Mensh, B., & Kording, K. (2017). Ten simple rules for structuring papers. *PLOS Computational Biology*, *13*(9), e1005619. https://doi.org/10/ckqp

Microsoft, & Weston, S. (2019). *Foreach: Provides foreach looping construct*. Retrieved from https://CRAN.R-project.org/package=foreach

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

Minxha, J., Adolphs, R., Fusi, S., Mamelak, A. N., & Rutishauser, U. (2020). Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science*, *368*(6498). https://doi.org/10/gg4btz

Mumford, J. A., Davis, T., & Poldrack, R. A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, *103*, 130–138. https://doi.org/10/f6qvfq

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). https://doi.org/10.1371/journal.pcbi.1003553

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... Wagner, H. (2019). *Vegan: Community ecology package*. Retrieved from https://CRAN.R-project.org/package=vegan

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2019). *nlme: Linear and nonlinear mixed effects models*. Retrieved from https://CRAN.R-project.org/package=nlme

Posner, M. I., & Snyder, C. R. R. (1975). Attention and Cognitive Control. In *In R. Solso (Ed.) Information processing and cognition: The Loyola Symposium* (pp. 669–682).

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The Role of the Medial Frontal Cortex in Cognitive Control. *Science*, *306*(5695), 443–447. https://doi.org/10/dzbqvx

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Roitman, J. D., & Shadlen, M. N. (2002). Response of Neurons in the Lateral Intraparietal Area during a Combined Visual Discrimination Reaction Time Task. *The Journal of Neuroscience*, *22*(21), 9475–9489. https://doi.org/10.1523/JNEUROSCI.22-21-09475.2002

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10/bsdb55

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. https://doi.org/10.1016/j.neuron.2013.07.007

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10/bxbw3c

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

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

Turner, B. M., Forstmann, B. U., & Steyvers, M. (2019). *Joint Models of Neural and Behavioral Data*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-03688-1

Twomey, T., Kawabata Duncan, K. J., Price, C. J., & Devlin, J. T. (2011). Top-down modulation of ventral occipito-temporal responses during visual word recognition. *NeuroImage*, *55*(3), 1242–1251. https://doi.org/10/b2dccr

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

Whitcher, B., Schmid, V. J., & Thornton, A. (2011). Working with the DICOM and NIfTI data standards in R. *Journal of Statistical Software*, *44*(6), 1–28. Retrieved from http://www.jstatsoft.org/v44/i06/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida:

Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/

Appendix

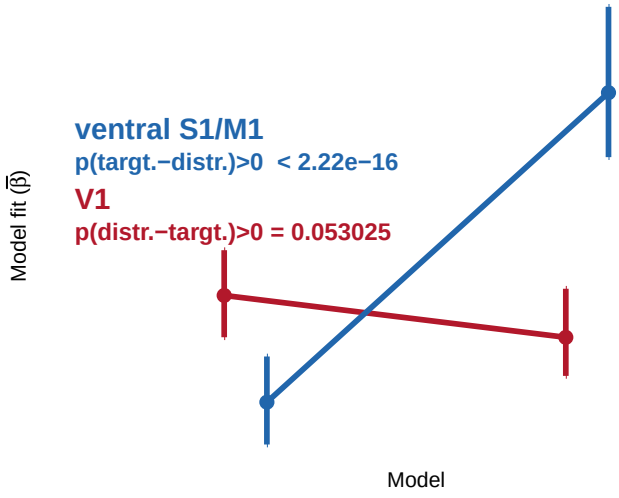

Figure A1. Expected double dissociation between ventral somatomotor cortex target coding and V1 distractor coding. Each region was bilateral.

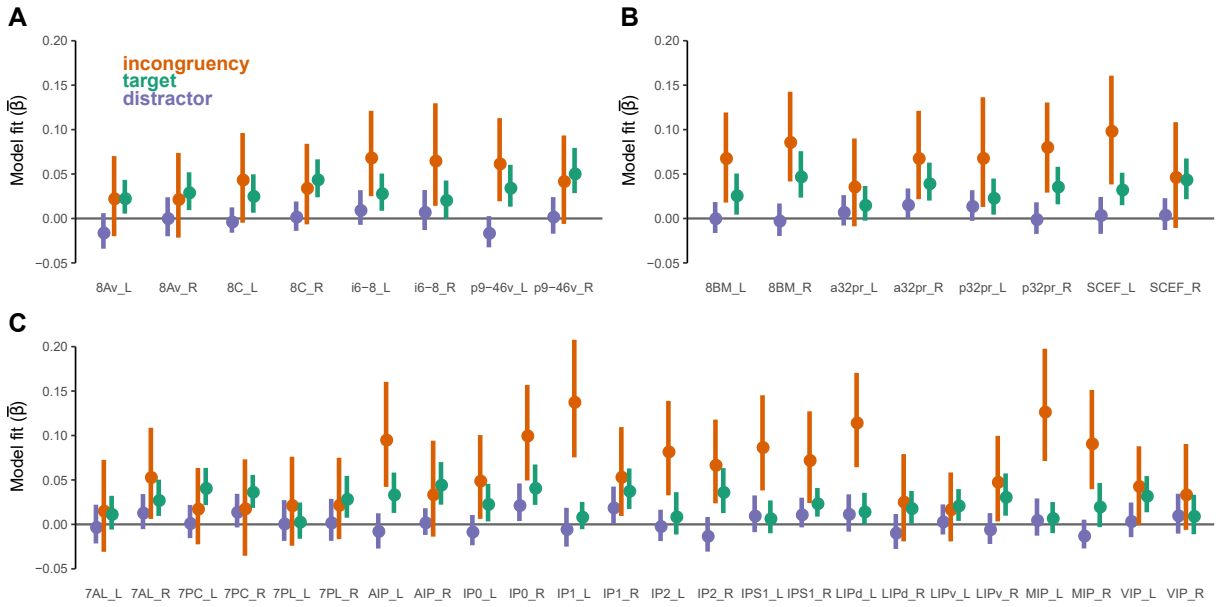

Figure A2. Mean RSA model fits in each MMP parcel within our frontoparietal ROIs. (**A**) DLPFC, (**B**) DMFC, (**C**) LPPC. Hemispheres displayed separately. Error-bars represent 95% confidence intervals on between subject variability from bias-corrected and accelerated bootstrap (10,000 resamples).

Table A1

Condition-wise mean estimates from the group-level RSA. Unless noted, means were averaged across hemisphere.

contrast	$\bar{\beta}$	σ	t	p
target DLPFC (L)	0.03	0.01	2.56	0.01058
target LPPC (L)	0.03	0.01	2.74	0.00617
target DMFC (L)	0.04	0.01	2.61	0.00918
target DLPFC (R)	0.04	0.01	3.18	0.00146
target LPPC (R)	0.05	0.01	4.38	0.00001
target DMFC (R)	0.05	0.01	3.94	0.00008
distr. DLPFC (L)	-0.01	0.01	-0.98	0.32577
distr. LPPC (L)	0.00	0.01	0.28	0.77731
distr. DMFC (L)	0.01	0.01	0.78	0.43652
distr. DLPFC (R)	0.00	0.01	0.31	0.75805
distr. LPPC (R)	0.02	0.01	1.85	0.06368
distr. DMFC (R)	0.00	0.01	0.19	0.84902
incon. DLPFC (L)	0.04	0.02	2.04	0.04127
incon. LPPC (L)	0.07	0.02	3.07	0.00211
incon. DMFC (L)	0.11	0.03	3.65	0.00026
incon. DLPFC (R)	0.05	0.02	2.39	0.01667
incon. LPPC (R)	0.06	0.02	2.64	0.00818
incon. DMFC (R)	0.08	0.03	2.82	0.00484

Table A2

Within-region contrast estimates from the group-level RSA. Unless noted, contrasts were averaged across hemisphere. For contrasts performed separately in each hemisphere, p-values were FDR-adjusted for two comparisons.

contrast	$\bar{\beta}$	σ	t	p
target − distr.: DLPFC	0.04	0.01	3.48	0.00050
target − distr.: LPPC	0.03	0.01	2.61	0.00917
target − distr.: DMFC	0.04	0.01	3.26	0.00111
target − incon.: DLPFC	-0.01	0.02	-0.50	0.61509
target − incon.: LPPC	-0.03	0.02	-1.13	0.25832
target − incon.: DMFC	-0.05	0.03	-1.69	0.09038
distr. − incon.: DLPFC	-0.05	0.02	-2.55	0.01062
distr. − incon.: LPPC	-0.06	0.02	-2.41	0.01609
distr. − incon.: DMFC	-0.09	0.03	-3.14	0.00167
target − distr.: DLPFC (L)	0.04	0.02	2.76	0.01148
target − distr.: LPPC (L)	0.03	0.02	1.87	0.06216
target − distr.: DMFC (L)	0.03	0.02	1.63	0.10233
target − distr.: DLPFC (R)	0.04	0.02	2.33	0.01994
target − distr.: LPPC (R)	0.03	0.02	1.95	0.06216
target − distr.: DMFC (R)	0.05	0.02	3.18	0.00297
target − incon.: DLPFC (L)	-0.01	0.02	-0.46	0.65768
target − incon.: LPPC (L)	-0.04	0.03	-1.62	0.21241
target − incon.: DMFC (L)	-0.07	0.03	-2.26	0.04725
target − incon.: DLPFC (R)	-0.01	0.02	-0.44	0.65768
target − incon.: LPPC (R)	-0.01	0.03	-0.43	0.66714
target − incon.: DMFC (R)	-0.03	0.03	-0.91	0.36027
distr. − incon.: DLPFC (L)	-0.06	0.02	-2.41	0.03193
distr. − incon.: LPPC (L)	-0.07	0.03	-2.74	0.01223
distr. − incon.: DMFC (L)	-0.10	0.03	-3.22	0.00253
distr. − incon.: DLPFC (R)	-0.05	0.02	-2.09	0.03654
distr. − incon.: LPPC (R)	-0.04	0.03	-1.63	0.10353
distr. − incon.: DMFC (R)	-0.08	0.03	-2.64	0.00819

Table A3

Between-region contrast and and interaction estimates from the group-level RSA. Unless noted, contrasts were averaged across hemisphere.

contrast	$\bar{\beta}$	σ	t	p
incon.: DMFC (L) – DLPFC	0.06	0.02	2.38	0.01739
incon.: DMFC (L) – LPPC	0.04	0.03	1.47	0.14101
target: DMFC (L) – DLPFC	0.00	0.01	-0.09	0.92458
target: DMFC (L) – LPPC	-0.01	0.01	-0.56	0.57859
[target – incon.] · [DLPFC – DMFC (L)]	0.06	0.03	2.20	0.02786
[target – incon.] · [DLPFC – LPPC (L)]	0.03	0.03	1.06	0.29001

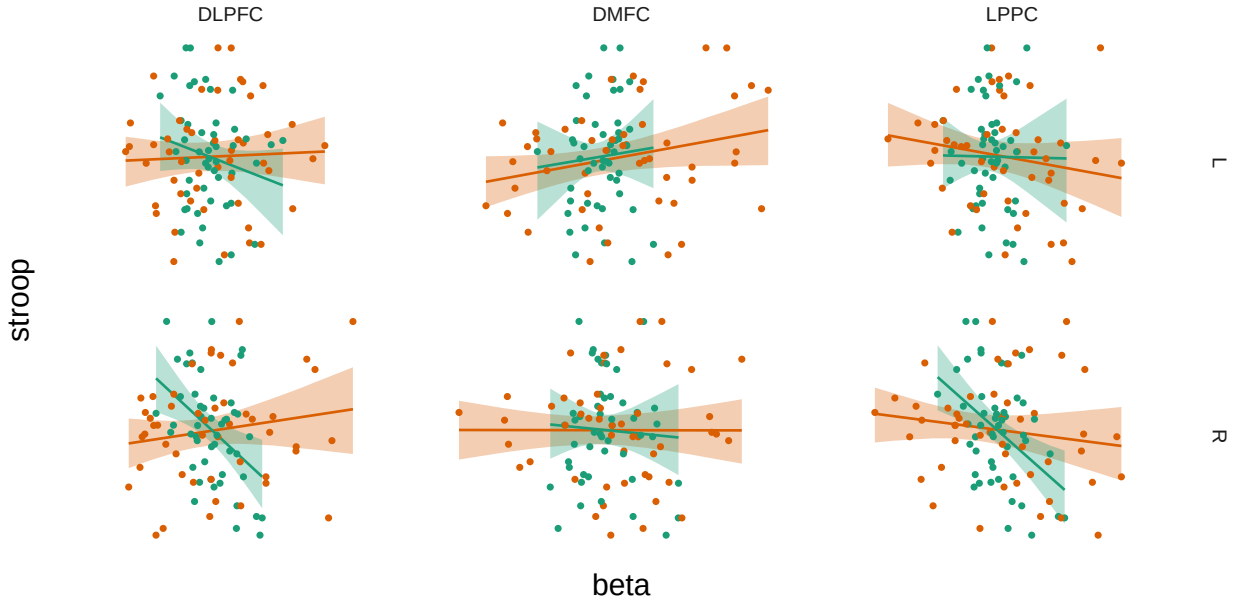

Figure A3. Associations between subjects' coding strength β and Stroop effect (response time) separately displayed within each hemisphere, ROI, and coding model (target, incongruency). Green, target coding model; orange; incongruency coding model. Stroop effects estimated as conditional modes within a hierarchical model (see Method). Confidence bands indicate 95% confidence interval from percentile bootstrap (10,000 resamples).

Table A4

Parameter estimates from a sensitivity test of the group-level RSA, which used alternative (more inclusive) definitions of DLPFC and DMFC (see Figure A4). This table contains statistics analogous to Tables A1-A3. Unless noted, contrasts were collapsed across hemisphere. For contrasts performed separately in each hemisphere, p-values were FDR-adjusted for multiple comparisons.

contrast	$\bar{\beta}$	σ	t	p
target DLPFC (L)	0.03	0.01	2.08	0.01058
target DMFC (L)	0.03	0.01	2.38	0.00918
target DLPFC (R)	0.05	0.01	3.17	0.00146
target DMFC (R)	0.05	0.01	3.65	0.00008
distr. DLPFC (L)	-0.01	0.01	-0.70	0.32577
distr. DMFC (L)	0.01	0.01	0.82	0.43652
distr. DLPFC (R)	0.00	0.01	0.27	0.75805
distr. DMFC (R)	0.00	0.01	0.05	0.84902
incon. DLPFC (L)	0.05	0.02	2.34	0.04127
incon. DMFC (L)	0.10	0.02	4.16	0.00026
incon. DLPFC (R)	0.04	0.02	2.18	0.01667
incon. DMFC (R)	0.08	0.02	3.16	0.00484
target – distr.: DLPFC	0.04	0.01	3.23	0.00050
target – distr.: DMFC	0.04	0.01	3.02	0.00111
target – incon.: DLPFC	-0.01	0.02	-0.36	0.61509
target – incon.: DMFC	-0.05	0.02	-2.05	0.09038
distr. – incon.: DLPFC	-0.05	0.02	-2.24	0.01062
distr. – incon.: DMFC	-0.08	0.02	-3.89	0.00167
target – distr.: DLPFC (L)	0.04	0.02	2.24	0.02501
target – distr.: DMFC (L)	0.02	0.02	1.33	0.18277
target – distr.: DLPFC (R)	0.04	0.02	2.39	0.02501
target – distr.: DMFC (R)	0.05	0.02	2.99	0.00551
target – incon.: DLPFC (L)	-0.02	0.03	-0.68	0.95934
target – incon.: DMFC (L)	-0.07	0.03	-2.64	0.01636
target – incon.: DLPFC (R)	0.00	0.03	0.05	0.95934
target – incon.: DMFC (R)	-0.03	0.03	-0.97	0.33044
distr. – incon.: DLPFC (L)	-0.06	0.02	-2.28	0.04530
distr. – incon.: DMFC (L)	-0.09	0.02	-3.67	0.00049
distr. – incon.: DLPFC (R)	-0.04	0.02	-1.63	0.10409
distr. – incon.: DMFC (R)	-0.08	0.02	-3.11	0.00188
incon.: DMFC (L) – DLPFC	0.06	0.02	3.41	0.01739
incon.: DMFC (L) – LPPC	0.03	0.02	2.01	0.14101
target: DMFC (L) – DLPFC	0.00	0.02	-0.17	0.92458
target: DMFC (L) – LPPC	-0.01	0.02	-0.51	0.57859
[target – incon.] · [DLPFC – DMFC (L)]	0.06	0.02	2.84	0.02786
[target – incon.] · [DLPFC – LPPC (L)]	0.03	0.02	1.57	0.29001

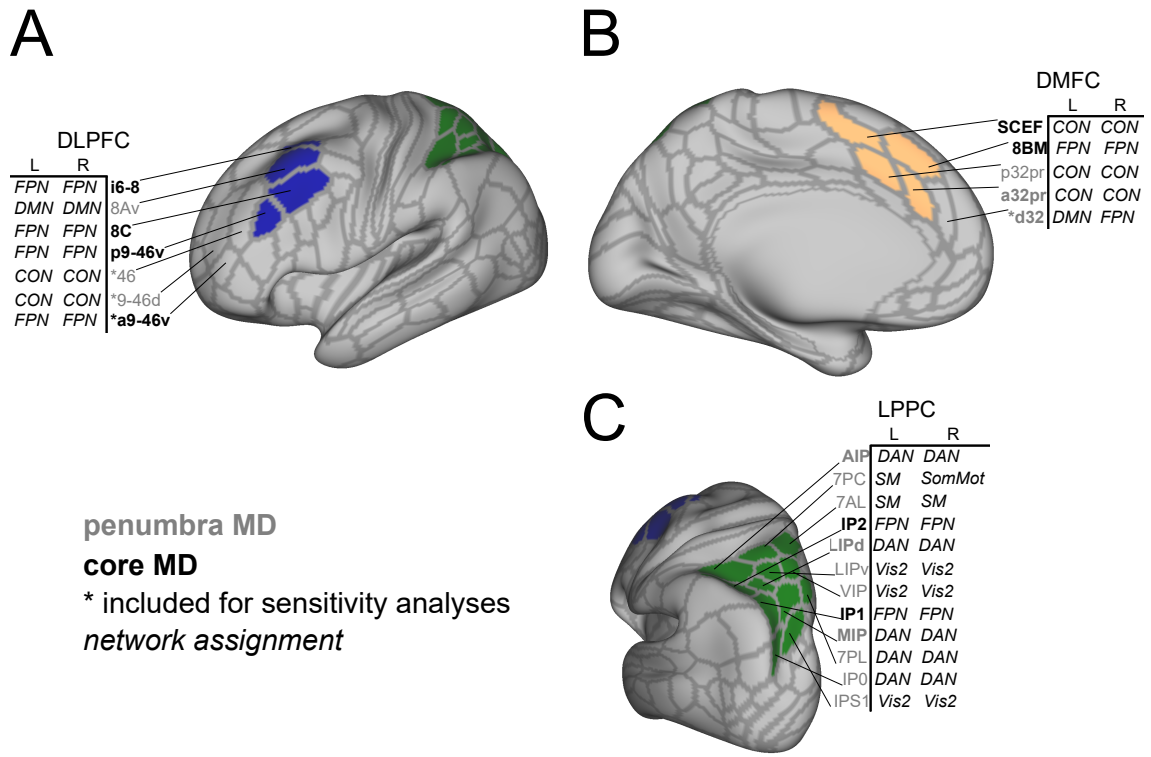

Figure A4. Regions of interest, or "superparcels", for primary analyses. Colored parcels were included in these superparcels. For DLPFC and DMFC, we additionally conducted sensitivity analyses with more inclusive definitions (i.e., including parcels with asterisked labels). Network assignments were provided in Ji et al. (2019). Parcel names that are in bold were identified as likely being situated over nodes of the "multiple demand" network (Assem, et al. 2020), a collection of regions that is commonly recruited by a variety of demanding tasks. Of these, those in grey belonged to the penumbra (recruited somewhat task-selectively) and those in black belonged to the core (robustly recruited regardless of task). DLPFC, dorsolateral prefrontal cortex; DMFC, dorsomedial frontal cortex; LPPC, lateral posterior parietal cortex; MD, multiple demand; FPN, frontoparietal network; CON, cinguloopercular network; DMN, default mode network; DAN, dorsal attention network; SM, somatomotor network; Vis2, 2nd visual network.

Table A5

Parameter estimates from separate brain–behavior models. Each estimate corresponds to the two-way interaction, stroop $\cdot\ \beta$), for a given region and coding model (e.g., incongruency). Incon., Incongruency.

model	b	σ	t	p
target DLPFC (R)	-239.13	71.68	-3.34	0.00171
target DLPFC (L)	-105.93	78.34	-1.35	0.17637
target LPPC (R)	-235.34	71.23	-3.30	0.00191
target LPPC (L)	-1.90	93.65	-0.02	0.98385
target DMFC (R)	-24.04	63.75	-0.38	0.70613
target DMFC (L)	48.40	90.63	0.53	0.70613
incon. DLPFC (R)	38.42	33.66	1.14	0.50735
incon. DLPFC (L)	14.08	41.17	0.34	0.73227
incon. LPPC (R)	-32.40	32.22	-1.01	0.31467
incon. LPPC (L)	-48.95	38.13	-1.28	0.31467
incon. DMFC (R)	-0.10	33.15	0.00	0.99762
incon. DMFC (L)	47.66	25.97	1.84	0.13293

Table A6

Contrasts on parameter estimates from separate brain–behavior models fitted using RSA model fits from each region (right DLPFC, right LPPC, left DMFC). Each estimate corresponds to the three-way interaction, stroop $\cdot\ (\beta_{target} - \beta_{incon.})$, for a given region. This interaction amounts to a contrast on slopes: within each region, does the slope between subjects' coding strengths and Stroop effects differ with the information encoded (target vs. incongruency)? Incon., Incongruency.

model	b	σ	t	p
DLPFC (R)	-299.21	80.62	-3.71	0.00021
LPPC (R)	-204.56	79.94	-2.56	0.01050
DMFC (L)	-20.92	97.23	-0.22	0.82962

Table A7

Contrasts on parameter estimates from brain–behavior models fitted using RSA model fits from pairs of regions. Each estimate corresponds to a between-region contrast [e.g., stroop $\cdot$ $(\beta_{DMFC\ (L)} - \beta_{DLPFC\ (R)})$*], for a given coding model (e.g, incongruency). Incon., Incongruency.*

contrast	b	σ	t	p
incon.: DMFC (L) – DLPFC (R)	22.85	52.92	0.43	0.66587
incon.: DMFC (L) – LPPC (R)	134.19	49.85	2.69	0.00710
target.: DMFC (L) – DLPFC (R)	500.63	139.14	3.60	0.00032
target.: DMFC (L) – LPPC (R)	468.99	137.46	3.41	0.00065

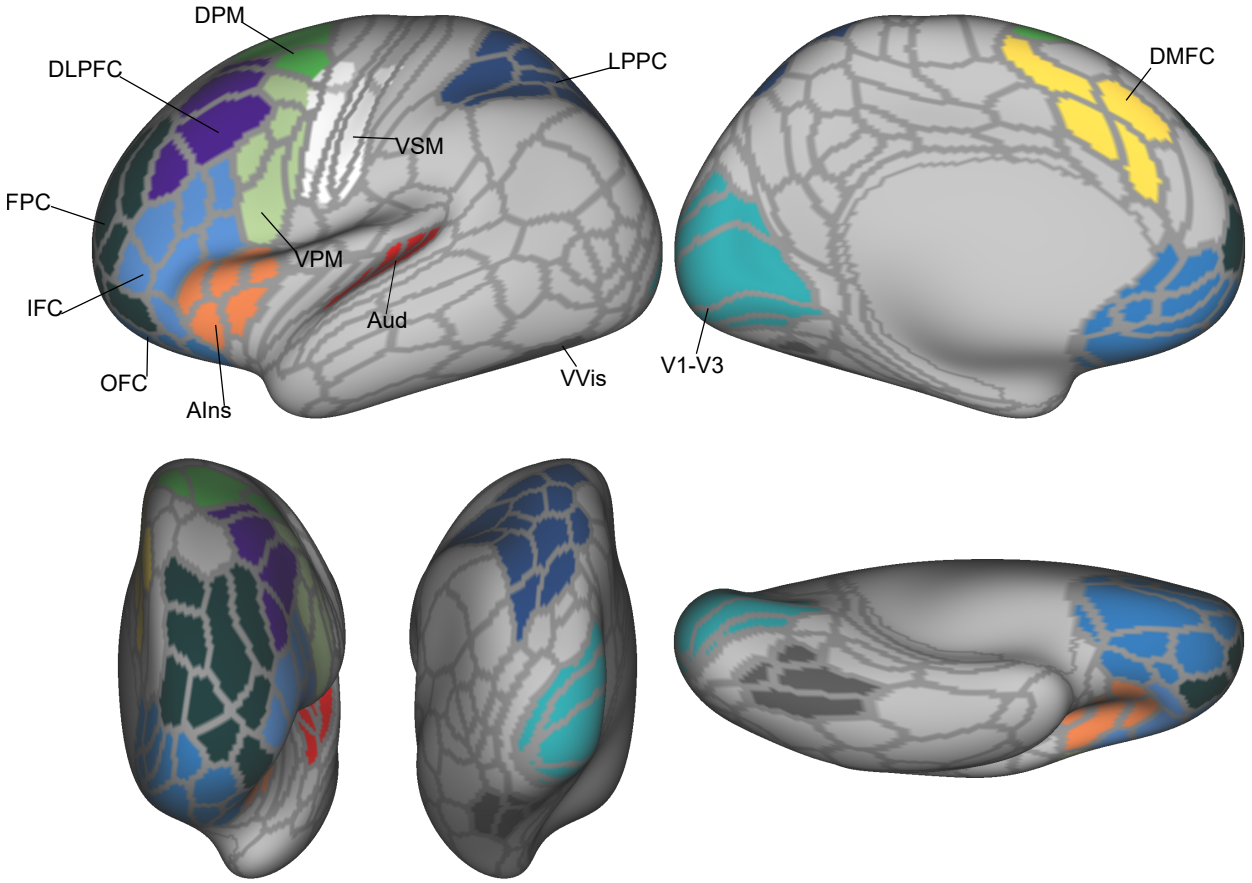

Figure A5. Superparcels defined for model-selection analysis. DPM, dorsal premotor cortex; VPM, ventral premotor cortex; VVis, ventral visual cortex; AIns, anterior (frontal) insular cortex; IFC, inferior frontal cortex; OFC, orbitofrontal cortex; FPC, frontoparietal cortex; Aud, early auditory cortex; V1–V3, visual cortex, VSM, ventral somatomotor cortex. Not displayed: left ventral occipito-temporal cortex ("visual word-form area").

Table A10

MMP parcels with significant target model fits. P-values are FDR-corrected across all 360 cortical parcels. $\bar{\beta}$, Across-subject mean of RSA model fits. v, Wilcoxon sum of signed-rank test statistic. network, Cole-Anticevic network assignemnt. md, Multiple-Demand assignment (Assem, et al. 2020).

parcel	$\bar{\beta}$	v	p	network	md
3b_L	0.08	1108	0.00001	Somatomotor	
4_L	0.07	1097	0.00001	Somatomotor	
FEF_R	0.05	1063	0.00010	Cingulo-Opercular	
4_R	0.06	1049	0.00016	Somatomotor	
3b_R	0.06	1047	0.00016	Somatomotor	
OP4_L	0.06	1033	0.00031	Somatomotor	
3a_L	0.07	1009	0.00088	Somatomotor	
PFop_L	0.05	1008	0.00088	Cingulo-Opercular	
11l_L	0.04	1004	0.00096	Frontoparietal	penumbra
11l_R	0.04	1000	0.00106	Frontoparietal	penumbra
3a_R	0.04	993	0.00135	Somatomotor	
PFop_R	0.04	989	0.00139	Cingulo-Opercular	
STSda_R	0.04	989	0.00139	Default	
a47r_L	0.04	985	0.00156	Frontoparietal	penumbra
V8_L	0.04	982	0.00164	Visual2	
V6_R	0.04	981	0.00164	Visual2	
LBelt_L	0.04	978	0.00165	Auditory	
p10p_L	0.05	978	0.00165	Frontoparietal	penumbra
5mv_R	0.05	977	0.00165	Cingulo-Opercular	
55b_L	0.05	976	0.00165	Language	
8C_R	0.04	974	0.00172	Frontoparietal	core
p9-46v_R	0.05	971	0.00188	Frontoparietal	core
POS2_L	0.05	970	0.00188	Frontoparietal	penumbra
6v_R	0.05	969	0.00188	Somatomotor	
PFcm_L	0.04	966	0.00206	Cingulo-Opercular	
55b_R	0.04	965	0.00207	Language	
43_R	0.04	963	0.00209	Cingulo-Opercular	
STGa_L	0.04	963	0.00209	Language	
7PC_R	0.04	957	0.00252	Somatomotor	
MI_R	0.04	957	0.00252	Cingulo-Opercular	
SCEF_L	0.03	956	0.00255	Cingulo-Opercular	core
23c_L	0.05	953	0.00280	Cingulo-Opercular	
FOP5_L	0.05	952	0.00283	Cingulo-Opercular	penumbra
TGv_R	0.04	951	0.00286	Language	
a32pr_R	0.04	950	0.00289	Cingulo-Opercular	penumbra
6d_L	0.04	949	0.00293	Somatomotor	
7PC_L	0.04	947	0.00310	Somatomotor	
A4_R	0.04	945	0.00319	Auditory	
AIP_R	0.05	945	0.00319	Dorsal-Attention	penumbra
IP0_R	0.04	943	0.00328	Dorsal-Attention	
MBelt_R	0.04	943	0.00328	Auditory	
STSda_L	0.05	942	0.00334	Language	
MI_L	0.03	941	0.00339	Cingulo-Opercular	
FOP4_R	0.05	938	0.00371	Cingulo-Opercular	
SCEF_R	0.04	937	0.00371	Cingulo-Opercular	core
TGd_R	0.04	937	0.00371	Default	
PBelt_R	0.05	936	0.00377	Auditory	
47l_L	0.03	935	0.00377	Default	
VVC_L	0.04	935	0.00377	Visual2	
5m_L	0.04	934	0.00384	Somatomotor	
6ma_R	0.04	933	0.00384	Cingulo-Opercular	
V3B_L	0.04	933	0.00384	Visual2	
6r_R	0.04	932	0.00391	Cingulo-Opercular	penumbra
6v_L	0.04	931	0.00399	Somatomotor	
6a_R	0.03	930	0.00407	Dorsal-Attention	
a10p_R	0.03	928	0.00424	Frontoparietal	penumbra
POS2_R	0.04	928	0.00424	Frontoparietal	penumbra
2_L	0.04	927	0.00433	Somatomotor	
DVT_R	0.04	923	0.00494	Visual1	
PGi_L	0.03	920	0.00543	Default	
9p_L	0.04	919	0.00554	Default	
p32pr_R	0.04	917	0.00587	Cingulo-Opercular	
FOP4_L	0.03	916	0.00599	Cingulo-Opercular	
IP1_R	0.04	915	0.00611	Frontoparietal	core
8BM_R	0.05	912	0.00660	Frontoparietal	core
PFm_R	0.03	912	0.00660	Frontoparietal	core
OP2-3_L	0.03	910	0.00698	Somatomotor	
6mp_L	0.04	909	0.00713	Somatomotor	
FOP3_R	0.04	908	0.00717	Cingulo-Opercular	
V4t_L	0.03	908	0.00717	Visual2	
OP4_R	0.04	907	0.00733	Somatomotor	
d32_R	0.04	906	0.00738	Frontoparietal	penumbra
VVC_R	0.04	906	0.00738	Visual2	
43_L	0.03	905	0.00744	Cingulo-Opercular	
TGd_L	0.04	905	0.00744	Default	

Table A10 continued

parcel	$\bar{\beta}$	v	p	network	md
IFSp_R	0.04	904	0.00760	Frontoparietal	
47s_L	0.04	903	0.00776	Default	
PBelt_L	0.04	902	0.00793	Auditory	
45_L	0.04	899	0.00847	Language	
TA2_L	0.03	899	0.00847	Auditory	
V6A_R	0.03	899	0.00847	Visual2	
a24pr_R	0.03	897	0.00895	Cingulo-Opercular	
d32_L	0.03	896	0.00915	Default	penumbra
44_R	0.03	895	0.00935	Frontoparietal	
FEF_L	0.04	893	0.00976	Cingulo-Opercular	
IFJa_R	0.04	893	0.00976	Language	
VIP_L	0.03	892	0.00997	Visual2	
24dd_L	0.04	891	0.01008	Somatomotor	
47m_R	0.03	891	0.01008	Default	
AVI_R	0.03	890	0.01030	Frontoparietal	core
31pv_R	0.04	889	0.01041	Default	
FOP5_R	0.05	889	0.01041	Cingulo-Opercular	penumbra
TE2a_L	0.03	886	0.01136	Default	
FOP1_R	0.03	885	0.01136	Cingulo-Opercular	
FOP3_L	0.04	885	0.01136	Cingulo-Opercular	
TPOJ1_L	0.03	885	0.01136	Language	
24dv_R	0.03	884	0.01162	Somatomotor	
AVI_L	0.03	883	0.01166	Frontoparietal	core
PoI2_L	0.04	883	0.01166	Cingulo-Opercular	
FOP1_L	0.03	882	0.01166	Cingulo-Opercular	
TA2_R	0.02	882	0.01166	Auditory	
TE1a_R	0.04	882	0.01166	Default	
TE2p_L	0.03	882	0.01166	Dorsal-Attention	
V7_L	0.03	880	0.01231	Visual2	
10d_L	0.03	879	0.01235	Default	
8Ad_R	0.03	879	0.01235	Default	
AIP_L	0.03	879	0.01235	Dorsal-Attention	penumbra
FST_L	0.03	878	0.01251	Visual2	
SFL_L	0.04	878	0.01251	Language	
a9-46v_R	0.04	877	0.01279	Frontoparietal	core
STV_L	0.02	876	0.01308	Language	
IP2_R	0.04	875	0.01325	Frontoparietal	core
IPS1_R	0.02	875	0.01325	Visual2	
10pp_L	0.03	874	0.01343	Default	
8Ad_L	0.03	874	0.01343	Default	
FFC_L	0.03	872	0.01393	Visual2	
FST_R	0.04	872	0.01393	Visual2	
PeEc_L	0.04	872	0.01393	Ventral-Multimodal	
OFC_R	0.03	871	0.01424	Frontoparietal	
RI_L	0.04	870	0.01456	Auditory	
PHT_L	0.03	869	0.01476	Dorsal-Attention	
TE1a_L	0.03	869	0.01476	Default	
a10p_L	0.03	868	0.01509	Frontoparietal	penumbra
PGi_R	0.02	867	0.01543	Default	
IFJp_R	0.03	865	0.01625	Frontoparietal	core
8Av_R	0.03	864	0.01635	Default	
i6-8_L	0.03	864	0.01635	Frontoparietal	core
p47r_L	0.03	864	0.01635	Frontoparietal	penumbra
5mv_L	0.03	862	0.01709	Cingulo-Opercular	
FFC_R	0.02	862	0.01709	Visual2	
8BL_R	0.03	858	0.01892	Default	
PI_R	0.03	858	0.01892	Cingulo-Opercular	
PEF_L	0.03	857	0.01904	Dorsal-Attention	
STV_R	0.03	857	0.01904	Posterior-Multimodal	
V2_L	0.03	857	0.01904	Visual2	
OFC_L	0.02	856	0.01904	Default	
PSL_R	0.03	856	0.01904	Cingulo-Opercular	
STSdp_R	0.03	856	0.01904	Language	
V3_L	0.03	856	0.01904	Visual2	
24dd_R	0.03	855	0.01919	Somatomotor	
46_L	0.03	855	0.01919	Cingulo-Opercular	
7AL_R	0.03	855	0.01919	Somatomotor	
IFSa_L	0.02	854	0.01934	Frontoparietal	
V2_R	0.02	854	0.01934	Visual2	
V6A_L	0.03	854	0.01934	Visual2	
1_L	0.03	853	0.01976	Somatomotor	
A5_L	0.02	851	0.02049	Language	
OP1_L	0.03	851	0.02049	Somatomotor	
OP1_R	0.02	851	0.02049	Somatomotor	
10r_R	0.03	850	0.02067	Default	
6r_L	0.03	850	0.02067	Cingulo-Opercular	penumbra
PFcm_R	0.02	850	0.02067	Cingulo-Opercular	
PH_L	0.03	848	0.02172	Visual2	
1_R	0.03	847	0.02190	Somatomotor	
A5_R	0.03	847	0.02190	Language	
p9-46v_L	0.04	847	0.02190	Frontoparietal	core
Ig_L	0.03	845	0.02272	Somatomotor	
PCV_R	0.03	845	0.02272	Posterior-Multimodal	
VMV3_R	0.03	845	0.02272	Visual2	
LO2_L	0.02	844	0.02320	Visual2	
MT_R	0.02	842	0.02421	Visual2	
TF_R	0.02	842	0.02421	Ventral-Multimodal	

Table A10 continued

parcel	$\bar{\beta}$	v	p	network	md
6mp_R	0.02	840	0.02525	Somatomotor	
TGv_L	0.03	840	0.02525	Language	
OP2-3_R	0.03	837	0.02721	Somatomotor	
LIPv_L	0.02	836	0.02745	Visual2	
LO3_R	0.03	836	0.02745	Visual2	
PEF_R	0.03	836	0.02745	Cingulo-Opercular	
LIPv_R	0.03	835	0.02802	Visual2	
a24_L	0.03	834	0.02827	Default	
DVT_L	0.03	834	0.02827	Visual1	
IFJa_L	0.03	834	0.02827	Language	
a47r_R	0.03	832	0.02963	Frontoparietal	penumbra
6d_R	0.03	831	0.02990	Somatomotor	
A4_L	0.03	831	0.02990	Auditory	
TE1m_L	0.02	831	0.02990	Default	penumbra
SFL_R	0.03	830	0.03051	Language	
8Av_L	0.02	829	0.03114	Default	
8C_L	0.03	827	0.03260	Frontoparietal	core
TPOJ1_R	0.03	826	0.03327	Language	
24dv_L	0.03	824	0.03481	Somatomotor	
9a_R	0.02	823	0.03531	Default	
V8_R	0.03	823	0.03531	Visual2	
9m_R	0.03	822	0.03582	Default	
ProS_R	0.03	822	0.03582	Visual1	
8BM_L	0.03	821	0.03634	Frontoparietal	core
H_R	0.02	821	0.03634	Default	
AAIC_R	0.02	820	0.03707	Orbito-Affective	
10v_R	0.02	818	0.03834	Default	
44_L	0.02	818	0.03834	Language	
p32pr_L	0.02	818	0.03834	Cingulo-Opercular	
a24pr_L	0.03	817	0.03910	Cingulo-Opercular	
a24_R	0.02	815	0.04064	Default	
PHA3_L	0.03	815	0.04064	Dorsal-Attention	
FOP2_L	0.02	814	0.04143	Somatomotor	
IFSp_L	0.02	811	0.04432	Language	
10v_L	0.02	810	0.04471	Default	
23c_R	0.02	810	0.04471	Cingulo-Opercular	
7PL_R	0.03	810	0.04471	Dorsal-Attention	
9a_L	0.02	808	0.04643	Default	
STSvp_R	0.03	808	0.04643	Default	
9-46d_R	0.03	806	0.04750	Cingulo-Opercular	
IFJp_L	0.02	806	0.04750	Frontoparietal	core
IP0_L	0.02	806	0.04750	Dorsal-Attention	
PGp_L	0.02	806	0.04750	Dorsal-Attention	
POS1_R	0.03	806	0.04750	Default	
s6-8_R	0.03	805	0.04839	Frontoparietal	penumbra

Table A8

Superparcels defined for model-selection analysis. All superparcels were lateralized (one in each hemisphere) except for relatively early sensory and motor regions ('Aud', 'V1–V3', 'VSM'). Ventral somatomotor cortex was defined using the Gordon et al. (2016) atlas 'somato-motor–mouth network', as MMP's somatomotor parcels encompass all of somatomotor strip (i.e., dorsal and ventral areas). Left ventral occipitotemporal cortex ("visual word-form area") was defined using coordinates from Twomey et al. (2011). DPM, dorsal premotor cortex; VPM, ventral premotor cortex; VVis, ventral visual cortex; AIns, anterior (frontal) insular cortex; IFC, inferior frontal cortex; OFC, orbitofrontal cortex; FPC, frontoparietal cortex; Aud, early auditory cortex; V1–V3, visual cortex, VSM, ventral somatomotor cortex. LVOT, left visual occipitotemporal cortex.

superparcel	constituent MMP parcels	# voxels (L, R)
DPM	6a, FEF, 6ma	508, 495
VPM	6v, PEF, 6r, 55b, IFJp	629, 572
VVis	FFC, VVC, V8, VMV3	648, 721
AIns	FOP4, FOP5, FOP3, AVI, MI	527, 536
IFC	IFJa, 44, 45, IFSa, p47r, IFSp, a47r, 47l	1065, 1263
OFC	47s, 47m, 11l, 13l, OFC, 10pp, 10v, 10r, s32, p32, a24, 25, pOFC	1927, 1810
FPC	46, 9-46d, a9-46v, a10p, p10p, a47r, 9a, 9p, 10d	1657, 1769
Aud	LBelt, A1, MBelt, PBelt	473
V1–V3	V1, V2, V3	3904
VSM	none; Gordon et al. (2016)	1114
LVOT	none; Twomey et al. (2011)	939

Table A9

*Largest 20% of brain–behavior correlations among all 84 superparcel*model combinations. r, linear correlation; ρ, rank correlation; subscripts l and u indicate lower and upper 95% confidence intervals from bias-corrected and accelerated bootstrapped distribution (10,000 resamples). DPM, dorsal premotor cortex; VPM, ventral premotor cortex; VVis, ventral visual cortex; AIns, anterior (frontal) insular cortex; IFC, inferior frontal cortex; OFC, orbitofrontal cortex; FPC, frontoparietal cortex; Aud, early auditory cortex; V1–V3, visual cortex, VSM, ventral somatomotor cortex. LVOT, left visual occipitotemporal cortex.*

region	model	r	r^2	r_l	r_u	ρ	ρ^2	ρ_l	ρ_u
DLPFC_R	target	-0.44	0.20	-0.66	-0.14	-0.39	0.15	-0.63	-0.08
LPPC_R	target	-0.44	0.19	-0.63	-0.18	-0.35	0.12	-0.59	-0.05
DLPFC_L	distractor	-0.43	0.18	-0.66	-0.14	-0.33	0.11	-0.60	-0.02
VVis_L	incongruency	-0.35	0.12	-0.52	-0.13	-0.43	0.18	-0.62	-0.19
IFC_R	target	-0.30	0.09	-0.55	0.00	-0.27	0.07	-0.52	0.02
VPM_R	target	-0.29	0.08	-0.54	-0.01	-0.17	0.03	-0.44	0.12
V1–V3_L	incongruency	-0.27	0.07	-0.52	0.05	-0.31	0.10	-0.56	-0.02
DMFC_L	incongruency	0.26	0.07	-0.02	0.49	0.17	0.03	-0.13	0.44
DPM_L	target	-0.24	0.06	-0.46	-0.01	-0.23	0.05	-0.47	0.05
VSM_L	distractor	0.23	0.05	-0.02	0.44	0.25	0.06	-0.04	0.50
VVis_R	incongruency	-0.22	0.05	-0.47	0.14	-0.31	0.09	-0.54	0.01
AIns_L	target	-0.19	0.04	-0.44	0.09	-0.19	0.04	-0.46	0.11
DLPFC_L	target	-0.19	0.04	-0.42	0.07	-0.16	0.03	-0.43	0.14
VSM_L	incongruency	0.18	0.03	-0.07	0.45	0.14	0.02	-0.16	0.42
LVOT	incongruency	-0.18	0.03	-0.40	0.08	-0.27	0.07	-0.50	0.02
LPPC_L	incongruency	-0.18	0.03	-0.40	0.14	-0.26	0.07	-0.49	0.04
FPC_L	target	-0.17	0.03	-0.43	0.09	-0.14	0.02	-0.41	0.15
DLPFC_R	incongruency	0.17	0.03	-0.14	0.45	0.10	0.01	-0.19	0.37
VSM_L	target	-0.16	0.03	-0.42	0.17	-0.14	0.02	-0.41	0.14
AIns_R	incongruency	0.15	0.02	-0.15	0.40	0.09	0.01	-0.20	0.37

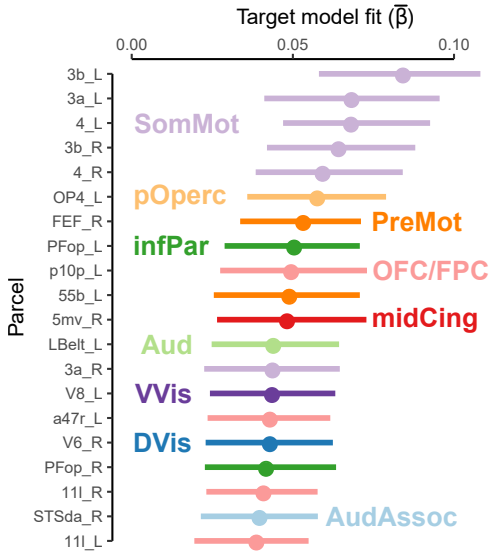

Figure A6. Largest 10% of target coding effect sizes observed in the whole-cortex exploratory analysis. Many of these regions recieve strong sensory or motor-related input.

Table A11
MMP parcels with significant distractor model fits. P-values are FDR-corrected across all 360 cortical parcels. $\bar{\beta}$, Across-subject mean of RSA model fits. v, Wilcoxon sum of signed-rank test statistic. network, Cole-Anticevic network assignemnt. md, Multiple-Demand assignment (Assem, et al. 2020).

parcel	$\bar{\beta}$	v	p	network	md
V2_L	0.05	1029	0.00144	Visual2	
V1_L	0.05	1025	0.00144	Visual1	

Table A12
MMP parcels with significant incongruency model fits. P-values are FDR-corrected across all 360 cortical parcels. $\bar{\beta}$, Across-subject mean of RSA model fits. v, Wilcoxon sum of signed-rank test statistic. network, Cole-Anticevic network assignemnt. md, Multiple-Demand assignment (Assem, et al. 2020).

parcel	$\bar{\beta}$	v	p	network	md
LIPd_L	0.12	978	0.01973	Dorsal-Attention	penumbra
MIP_L	0.13	967	0.01973	Dorsal-Attention	penumbra
s6-8_R	0.09	961	0.01973	Frontoparietal	penumbra
IP1_L	0.14	956	0.01973	Frontoparietal	core
6ma_L	0.09	937	0.02743	Cingulo-Opercular	
RSC_L	0.09	937	0.02743	Frontoparietal	
9-46d_L	0.07	934	0.02743	Cingulo-Opercular	
IP0_R	0.10	920	0.03821	Dorsal-Attention	
6ma_R	0.08	918	0.03821	Cingulo-Opercular	
LO2_R	0.09	914	0.03821	Visual2	
8BM_R	0.09	913	0.03821	Frontoparietal	core
p32pr_R	0.08	907	0.04334	Cingulo-Opercular	

Table A13

Results from conducting RSA on movement regressors. Subscripts indicate basis.

model	$\bar{\beta}_6$	p_6	$\bar{\beta}_{12}$	p_{12}	$\Delta\bar{\beta}_{12-6}$	p_{12-6}
distractor	0.000	0.64450	0.004	0.42151	0.004	0.17161
incongruency	0.009	0.42401	0.008	0.41154	-0.001	0.49025
target	0.010	0.36027	0.013	0.17059	0.003	0.17770