A Representational Similarity Analysis of Cognitive Control during Color-Word Stroop

Michael C. Freund,† Julie M. Bugg,† and Todd S. Braver‡§

†Department of Psychological & Brain Sciences, Washington University in St. Louis, St. Louis, Missouri 63130, ‡Department of Radiology, Washington University in St. Louis School of Medicine, St. Louis, Missouri 63110, and §Department of Neuroscience, Washington University in St. Louis School of Medicine, St. Louis, Missouri 63110

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). For this paradigm, models have postulated complementary coding schemes: dorsomedial frontal cortex (DMFC) is proposed to evaluate the demand for control via incongruency-related coding, whereas dorsolateral PFC (DLPFC) is proposed to implement control via goal and target-related coding. Yet, 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. Three neural coding models were fit to the similarity structure of multivoxel patterns of human fMRI activity, acquired from 65 healthy, young-adult males and females. Incongruency coding was predominant in DMFC, whereas both target and incongruency coding were present with indistinguishable strength in DLPFC. In contrast, distractor information was 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 representational similarity analysis framework for investigating neural mechanisms of cognitive control and point to several promising directions to extend the Stroop paradigm.

Key words: dorsal ACC; dorsolateral PFC; executive function; fMRI; multivariate pattern analysis; response conflict

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, which estimates and models the similarity structure of brain activity patterns, can successfully establish the hypothesized functional dissociations within the Stroop task. Representational similarity analysis may provide a useful approach for investigating cognitive control mechanisms.

Introduction
Goals, held in mind, can be used to overcome behavioral habits. Understanding how the human brain enables such cognitive control has been a fundamental interest of both basic and translational cognitive neuroscience. Toward this end, the use of response conflict tasks has been instrumental (e.g., Botvinick et al., 2001; Ridderinkhof et al., 2004). These 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 (Stroop, 1935; Posner and Snyder, 1975; MacLeod, 1991): on each trial, the hue of a word must be named, despite the word expressing a potentially conflicting, that is, incongruent, color (see Fig. 1C). A major goal in
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 (the goal-relevant mapping of the stimulus hue to the target response, i.e., color naming), the distractor (the prepotent but goal-irrelevant mapping of the stimulus word to the non-target or distractor response, i.e., word reading), and incongruency (whether the target and distractor responses match or mismatch). B, Hypotheses. Neuroscientific frameworks of cognitive control propose that representation of task dimensions is anatomically dissociated across medial and lateral frontoparietal cortices. DMFC, including dorsal anterior cingulate (dACC) and pre-supplementary motor area (pre-SMA), is proposed to implement goal-driven attentional selection and mapping of hue-target-response processes, by way of representing information related to the goal-dependent target dimension (green univariate distributions; Macdonald et al., 2000; MacDonald et al., 2000; 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 univariate distribution; Kane and Engle, 2002; Braver, 2012). C, 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 16 unique Stroop stimuli (e.g., “WHITE” displayed in blue hue) independently for each voxel. We analyzed a mostly-incongruent trial condition to obtain balanced estimates of each trial type. D, 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 16 unique Stroop stimuli (e.g., “WHITE” displayed in blue hue) independently for each voxel. We analyzed a mostly-incongruent trial condition to obtain balanced estimates of each trial type. E, Empirical similarity matrices. To address these questions, a neuroanatomically precise technique is needed that does not rely on temporal dissociations, but can instead read out multiple, simultaneously encoded sources of rule and incongruency-related information, in which subjects were instructed before each Stroop trial about which task to perform (color-naming, word-reading; MacDonald et al., 2000; Flodén et al., 2011). 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 Flodén et al., 2011; N = 9 in MacDonald et al., 2000). This fact alone warrants a follow-up study. Second, it is unclear whether the results extend to the more-standard Stroop-task design, in which task rules are not explicitly instructed before each trial, but are instead internally maintained. For example, goal-relevant coding in DLPC may depend on such explicit rule instruction. 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 DLPC (or other frontoparietal regions) in Stroop is primarily preparatory, and is less critical during actual interference resolution.
task information from individual brain regions of interest (ROIs). Multivariate (multivoxel) pattern analysis (MVPA) of fMRI, in popular use for over a decade (Edelman et al., 1998; Haxby et al., 2001; Cox and Savoy, 2003), 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 et al., 2008), a specific MVPA framework, to test for dissociations in frontoparietal coding during Stroop-task performance (see Fig. 1C–E). We conducted a retrospective analysis of data collected as part of the Dual Mechanisms of Cognitive Control project (Braver et al., 2020). Our primary goal was proof-of-principle: to demonstrate the potential of RSA for testing theorized distinctions in neural coding within cognitive control tasks, such as the Stroop (Freund et al., 2021).

Materials and Methods

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

**Code, data, and task accessibility**

Code (R Core Team, 2019) and data to reproduce all analyses, in addition to supplementary analysis reports, are publicly available (https://doi.org/10.5281/zenodo.4784067). As part of the planned data release of the Dual Mechanisms of Cognitive Control project, raw and minimally preprocessed fMRI data have been deposited on OpenNeuro (https://doi.org/10.18112/openneuro.ds003465.v1.0.3). Additionally, the authors will directly share the specific fMRI data used for this study on request.

Task scripts are available at the 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; 38 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). One set from the Model selection analysis (see below) was used in all analyses, and a validation set (N=16; 11 women, 5 men), which was only used in the Model selection analysis (see below). This unbalanced partitioning was done to account for the familial structure present within our sample. Specifically, subjects within each set (primary validation) were all unrelated; however, subjects within the validation set were co-twins of 16 subjects within the primary analysis set. 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 co-twins were randomly selected. Critically, partitioning the sample in this way ensured that the primary analysis set was a random sample of independent subjects.

The partitioning of the data into two subsets also afforded the opportunity to use the validation subset as held-out data for evaluation of the brain–behavior model within the Model selection analysis. 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 statistically "double-dipping" could have occurred. But because the sets are familialy dependent, it is perhaps more accurate to consider the 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. To evaluate this matter, follow-up control analyses were conducted in which the co-twins were removed from the primary analysis set.

Experimental design and statistical analysis

**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 subsets of color-word stimuli (randomly interleaved 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 set, of ‘black’ (0, 0, 0), ‘green’ (0, 128, 0), ‘pink’ (255, 100, 180), and ‘yellow’ (255, 255, 0). These words were centrally presented in uppercase, bold Courier New font on a gray 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 incongruity statistics are standard manipulations to elicit proactive control (Bugg, 2014; e.g., Gonthier et al., 2016) and were performed to investigate questions outside the scope of the current study. Thus, as described further below, the unbiased stimulus set was excluded from all analyses.

Each trial (see, e.g., Fig. 1C) began with a central fixation cross, presented for 300 ms on a gray 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 intertrial interval (triangle of fixation crosses) was 900, 2100, or 3300 ms, selected randomly (with uniform probability). Each of two scanning runs consisted of three blocks of 36 trials, intermixed with four resting fixation blocks, during which a fixation cross appeared for 30 s. This formed a mixed block–event design (Petersen and Dubis, 2012). Each of the 16 mostly incongruent stimuli — that is, each unique colored word (e.g., “BLUE” displayed in red hue) — was 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. Balanced designs ensure that differences in the total number of trials per condition cannot explain any differences observed in pattern correlations among conditions.

**Display and recording systems.** The experiment was programmed in E-Prime 2.0 (2013, Psychology Software Tools), 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 MR-compatible electronic microphone (MicroOptics Technologies); because of mechanical failure, however, we replaced this microphone with the noise-cancelling FOMRI III (OptioAcoustics), which subsequent participants used. A voice-onset processing script (from the MATLAB Audio Analysis Library) was used to derive RT estimates on each trial via spectral decomposition (the accuracy of which was verified by manually coding RTs 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/dualmekanisms/tree/master/preparationsAndConversions/audio).

Importantly, we verified that the change in microphone did not induce confounding between-subject variance in RT measures of interest. While RT estimates recorded via the Micro-Optics microphone tended to be slower ($b = 102.59, p < 0.01$) and more variable ($\chi^2 = 3655, p < 0$), the magnitude of the Stroop effect was not observably impacted by the microphone change ($b = 5.88$, $p = 0.64$).

Image acquisition, preprocessing, and GLM. The fMRI data were acquired with a 3T Siemens Prisma (32 channel head-coil; CMRR multiband sequence, factor = 4; 2.4 mm isotropic voxel, with 1200 ms TR, no GRAPPA, ipat = 0), and subjected to the minimally preprocessed functional pipeline of the Human Connectome Project (version 3.17.0), outlined by Glasser et al. (2013). More detailed information regarding acquisition is found on the Project OSF site (https://osf.io/tbbf9/).

All analyses were conducted in volumetric space; surface maps are displayed in figures only for ease of visualization. Before revision of this manuscript, the data were reprocessed with fMRIPrep, using the standard fMRIPrep pipelines (Esteban et al., 2019, 2020). At this point, the preprocessed results with the HCP pipelines were inadvertently removed. Thus, some follow-up control analyses were conducted with the fMRIPrep-preprocessed data (Extended Data Figs. 2-4, 3-2, 4-4). The fMRIPrep pipeline was implemented in a Singularity container (Kurtzer et al., 2017) with additional custom scripts used to implement file management (more detail on the pipeline is available at https://osf.io/6p3en/).

Preprocessing, to estimate activation patterns, we fit a whole-brain voxelwise GLM to BOLD time-series in AFNI, version 17.0.0 (Cox, 1996). To build regressors of primary interest, we convolved with an HRF [via AFNI’s BLOCK(1,1)] 16 boxcar time courses, each coding for the initial second of presentation of a mostly incongruent stimulus that resulted in a correct response. We also included two regressors [similarly created via BLOCK(1,1)] to capture signal associated with congruent and incongruent trials of noninterest (unbiased stimuli) that prompted correct responses, an error regressor coding for any trial in which a response was incorrect or omitted (via BLOCK), a sustained regressor coding for task versus rest (via BLOCK), a transient regressor coding for task-block onset [as a set of piecewise linear spline functions via TENTzero(0,16,8,8)], six orthogonal motion regressors, five polynomial drift regressors (order set automatically) for each run, and 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 × 3 blocks × 36 trials (144 from the mostly-incongruent stimulus group, 72 from unbiased). Frames with FD > 0.9 were censored.

The distractor model assumed that conditions with the same word will evoke identical patterns. This model predicts $R_{ij} = 1$, otherwise 0. In the “target” matrix in Fig. 1E, 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 $i = j$, this model predicts $R_{ij} = 1$, otherwise 0. If the “target” matrix in Fig. 1E, the only cells equal to 1 are those in which the stimulus hues match.) The incongruency model assumed that only conditions that evoked incongruent patterns were considered, regardless of the hue or word. If $i$ and $j$ are both incongruent, $R_{ij} = 1$, otherwise 0. In the “incongruency” matrix in Fig. 1E, the only cells equal to 1 are those in which both stimuli are incongruent.) As a covariate of noninterest, we also included a model capturing similarity between congruent conditions (if $i$ and $j$ are both congruent, $R_{ij} = 1$, otherwise 0). We additionally examined alternative approaches to RSA modeling of incongruity, to see whether our results were robust to this parameterization of incongruency (see Alternative RSA incongruity models).

Alternative RSA incongruity models.

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 $y$ and design matrix of this regression were assembled in a series of steps. (1) The 120 unique off-diagonal elements 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 $r$. (4) The vector $r$ was pretrained to remove a specific nuisance component. This component stemmed from the task design, and throughput each model, each 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 3 times in the first run versus 6 times in the second (vice versa for the other half). As each scanning run contains a large amount of run-specific noise (Mumford et al., 2014; Alink et al., 2015), this imbalance across runs could lead to a bias in the resulting \( \beta \) coefficients, in which pattern similarity of stimuli that mostly occurred within the same run would be inflated. We formalized this component of bias as another model similarity vector, \( v \), with elements equal to 1 if the run in which condition \( j \) most frequently occurred, otherwise 0. The magnitude of this bias was estimated as the slope term \( b_{\text{v}} \) in a linear regression \( r = v_{\text{b}} + b_{\text{e}} + e \), where \( b_{\text{e}} \) is the intercept coefficient and \( e \) is the residual vector. The model \( v \) was scaled by its magnitude and then subtracted from \( r \), forming the RSA response vector \( r' = r - v_{\text{b}} \). We additionally used an alternative, down-sampling technique to verify that our primary findings were robust to this issue (see Downsampling analysis).

Thus, the RSA regression yielded three \( \beta \) coefficients of interest: \( \beta_{\text{target}}, \beta_{\text{distr.}}, \beta_{\text{incon.}} \). 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 \( \beta_{\text{target}} \), 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, nonparametric dimensionality reduction technique, to visualize the structures of activity patterns within selected regions (see Fig. 2): ventral somatomotor cortex (corresponding to “mouth” homunculi), primary visual cortex (V1), and (left) DMFC. These parcels were selected to highlight coding of each task dimension. For each selected region, we averaged observed correlation matrices across subjects and then subtracted these values from 1 to obtain a dissimilarity matrix. Before averaging, we z-transformed (inverse hyperbolic tangent, arctanh) correlations, and inverted this transform after averaging. (In contrast to the RSA regression above, we did not rank-transform correlation matrices, as non-metric multidimensional scaling incorporates a monotonic regression). Similar to our RSA, we prewhitened each similarity matrix before conducting this procedure (see Step 4 in Estimation of coding strength \( \beta \)). Each mean dissimilarity matrix was submitted to an implementation of Kruskal’s nonmetric multidimensional scaling, `vegan::metaMDS()` in R, to generate a two-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 \( \beta \)) obtained from our three ROIs within each hemisphere, and for our three RSA models. Fixed effects were estimated for the interaction of RSA model, ROI, and hemisphere. Random effects by subject were estimated for the interaction of RSA model and ROI, with a full covariance structure (9 \( \times \) 9; Barr et al., 2013). This model was fit with `lmets::lmert()` in R (Bates et al., 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 et al., 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-model (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.

**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 nontarget color word spoken by a subject before 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., because of high scanner noise or poor enunciation) were coded as such.

We fit two hierarchical models on these data: one on errors and one on RTs. Several observations were excluded from these models. From the error model, only trials with responses coded as “unintelligible” were excluded (54). From the RT model, several types of trials were excluded: trials with RTs >300 ms (1) or <250 ms (53; 52 of which were equal to zero). A cluster of fast and unrealistically irrevocable RTs from 2 subjects (23/216, 26/216) that were likely because of an artifact of insufficient voice-onset signal within the recording. Trials with a residual RT that was more extreme than three interquartile ranges from an initial multilevel model fitted to all subjects data (of the structure of the model equation below; as in Baiyen and Milin, 2010). All were trials with incorrect (137), unintelligible (54), or no response (52). In total, 232 trials were excluded (0-62 per subject), leaving 10,352 trials for analysis (154-216 per subject).
RTs for subject s were modeled (following Laird and Ware, 1982 notation) as follows:

\[
\begin{align*}
\text{RT}_s & = (1, \text{stroop})_s \beta + (1, \text{stroop})_u \theta + \epsilon_s \\
\text{u}_s & \sim N(J_0, (T_2, 2)) \\
\text{e}_s & \sim N(0, c, \sigma^2)
\end{align*}
\]

where RT, is a column vector of the RTs from subject s, 1 is an all-ones vector (intercepts), stroop, is a vector indicating incongruent trials, and \( e_s \) is a subject-specific parameter by which their residual variance was scaled. Critically, \( b \) and \( u \), contained coefficients corresponding to the classical Stroop interference effect contrast (incongruent – congruent), for the group \( (b) \) and subject \( (u) \) levels. The scaling parameters \( c \) relaxed the assumption that each subject had a common residual variance, a well-warranted complexity, given the vastly improved fit of the heterogeneous-variance model \( (\chi^2_p = 4299, p = 0.00; \Delta BIC = -3856, \Delta AIC = -4203) \). To accomplish this, the RT model was fitted with \( nlme\)(lm4) in R (Pinheiro et al., 2019). The behavioral error model had similar predictors, but assumed binomially distributed error, with logit link function (fitted in lm4).

This hierarchical modeling framework enabled us to estimate the amount and internal consistency of individual variability in the Stroop interference effect within both RTs and errors while accounting for trial-level error (e.g., Haines et al., 2020). We used these subject-level estimates to validate that our behavioral measures met prerequisite properties for individual differences analyses. To assess the amount of between-subject variability in the Stroop interference effect, a nested model comparison was conducted, in which the models fitted above were compared with a random-intercept model. Stroop effects differed significantly across subjects in RT \( (\chi^2_p = 87.53, p = 0.00; \Delta BIC = -69.08; \Delta AIC = -83.53) \), but not in accuracy \( (\chi^2_p = 0.69, p = 0.71; \Delta BIC = 18; \Delta AIC = 3) \). We therefore did not further analyze accuracy. To assess internal consistency (defined here as cross-run correlation), we fit a model with separate congruency factors (fixed and random) per run, and a full \( 4 \times 4 \) covariance structure, which was used to obtain the cross-run correlation in Stroop effect. Individual Stroop interference effects in RT were estimated to be 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 that error measures were inadequate for individual differences analyses (see above), so we focused solely on RT measures.

We began with the RT model described in the preceding section. However now, for a given RSA model and ROI, we incorporated into the fixed effects each subject’s estimated coding strength, \( \beta \), by interacting this coding-strength term with the congruency factor. This formed a cross-level, continuous-by-categorical interaction, \( \beta \cdot \text{stroop} \). The coefficient on this interaction term described how the Stroop interference effect (within-subject) varied across subjects as a function of their coding strength (between-subject). To test our hypotheses, we performed contrasts on these interaction coefficients, which are outlined within Results describing Figure 3.

Model selection. To complement our hypothesis-driven brain–behavior analyses, we used a more data-driven model-selection approach. An expanded set of 24 superparcels (in addition to our six ROIs) was defined (see Fig. 3B for list, see Extended Data Fig. 3-3). Some superparcels were included as ROIs, others were included as negative controls (i.e., regions that were not predicted to be important for explaining behavioral performance). Subject-level coefficients of the Stroop interference effect contrast were extracted from the behavior-only RT model and used as the response vector (in the model equation, the slope elements of \( u \); also known as BLUPs or conditional modes). RSA was conducted on each superparcel (see Estimation of coding strength \( \beta \)), furnishing three \( \beta \) coefficients per superparcel. These 72 measures were fitted to the response vector via elastic net regression, implemented via glmnet/glmmnet() in R (Friedman et al., 2010). Parameter \( \alpha \) was set to 0.5 (balancing lasso and ridge penalties). The parameter \( \lambda \) was tuned via 10-fold cross-validation (default) using the “1-SE rule”: the smallest \( \lambda \) > 1 SE of the minimum \( \lambda \) across folds was saved (Hastie et al., 2009); this routine was repeated 1000 times, and the minimum mode (there were ties) of the saved \( \lambda \)s 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, refitting 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 \( \beta \), separately conducted on each MMP cortical parcel. Inferential statistics followed those suggested by Nili et al. (2014). One-sided signed-rank tests were conducted for significance testing (>0). \( p \) values were FDR-corrected over all 360 parcels, separately for each task dimension (target, distractor, incongruency).

Univariate activation analyses. We additionally conducted a standard “univariate activation” analysis on these data. This was not meant to evaluate whether univariate activity was a plausible confounding variable in our analysis, but rather to provide some basis for comparing our data to most extant neuroimaging studies of Stroop. For a given ROI (or MMP parcel), \( \beta \) coefficients from the first-level fMRI GLM were averaged over voxels by stimuli, then over stimuli by congruency. These mean values were then contrasted, analogous to the behavioral Stroop interference effect (incongruent – congruent). This statistic gives an estimate of the overall (across-voxel) difference in fMRI activity within a given brain region on incongruent versus congruent trials.

Follow-up control analyses

To establish the robustness of our results, we conducted several control analyses that examined a number of confounds and concerns: potential differences in signal-to-noise ratio (SNR) across prefrontal ROIs, the effects of head motion, different RSA models for incongruency coding, the presence of bias imposed by the experimental design, within-run versus between-run RSA estimation, and the effects of downsampling to account for different trial numbers across runs. These are each described next.

Comparison of SNR ratios. To test for differences in SNR between DMFC and DLPCF, 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 cross-validated “lower-bound” noise ceiling estimator of Nili et al. (2014), as this yields a lower-variance estimate, and therefore more powerful contrast across regions, than the non-cross-validated “upper-bound” (Hastie et al., 2009). For a given region, the lower-bound noise ceiling is defined for each subject \( s \) in \( 1, \ldots, N \) as \( \text{Cor}(y, y_{-s}) \); the linear correlation between a subject \( s \) observed similarity vector, \( y_s \), and the group-mean vector excluding subject \( s \), \( y_{-s} \) (for definition of \( y \), see Estimation of coding strength \( \beta \)). As each of the \( N \) estimates is interdependent, we used a percentile bootstrap to contrast noise 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 (Schuirmann, 1987; Lakens, 2017) 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
Figure 3. Individual-level results. A, The coding strength of task variables (β) within right DLPFC and left DMFC accounted for individual differences in the size of the Stroop interference effect (RT) in hypothesized ways. Inset text represents linear correlation coefficients (r) and 95% CIs (estimated via bias-corrected and accelerated bootstrap). Lines indicate a regression line of RT onto Model Fit (β), and confidence bands indicate 95% CIs of predicted values from this regression obtained via percentile bootstrap. The direction of these observed brain–behavior relationships was robust to alternative, more conservative RSA estimation techniques (cross-run RSA and cross-validated RSA; Extended Data Figure 3-2). For display of relationships between target and incongruency coding in all ROIs (DLPFC, DMFC, LPPC) and Stroop effect, see Extended Data B. A larger set of 24 cortical regions were defined for use in a data-driven model selection analysis. Surface maps display in various colors all left-hemisphere regions (excluding left lateral occipital-temporal cortex; for full list and definitions, see Extended Data Figure 2-1). Three RSA models (target, distractor, incongruency) were fitted per region, yielding a set of 72 coding-strength estimates. These 72 estimates were used to predict Stroop interference effects within a cross-validated model selection procedure. B, Additional brain–behavior associations identified via cross-validated model selection. 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. For description of inset text, lines, and bands, see A. D, Estimate of prediction accuracy of the selected model. Within a validation (held-out) set of 17 subjects (monovygotic co-twins of subjects within the analysis sample), observed Stroop effects were moderately correlated (red line) to predictions from the selected model. A permutation test (gray distribution) indicated the selected model significantly explained validation-set variance, suggesting that the selected model results are relatively stable.

A previous 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 on the design matrix containing predicted BOLD timescourses of our 16 conditions of interest. The RSA incongruency model parameterized the congruent–congruent correlations (i.e., $R_i$ where $i$ and $j$ are both patterns from congruent trials, denoted here simply as $CC$) with a separate nuisance regressor. That is, these cells were effectively excluded, and the model instead computed the contrast $II - IC$. This exclusion was done because we have no specific hypotheses regarding how congruent trials should be encoded relative to one another. Other parameterizations are possible, however, including models that (a) incorporate CC correlations within the "baseline" or intercept term, by omitting the congruent nuisance regressor [i.e., $II - ave(IC+CC)$], or (b) that omit IC correlations from the contrast (i.e., $II - CC$). We note, however, that (b) is a suboptimal parameterization as it effectively excludes 40% (48/120) of observations per subject (i.e., all IC cells). To verify that our results were not dependent on the particular modeling approach we chose, we compared the observed RSA model fits to these alternative parameterizations. In brief, subject's model fits were highly similar between the parameterizations (see Group and Exploratory whole-cortex RSA; Extended Data Fig. 2-5). Thus, to help streamline the Results, we only conducted and reported analyses using the original RSA incongruency parameterization (in which the congruency model was a nuisance covariate).

Design bias. For the current study, a typical within-run form of RSA estimation was implemented, in which correlations were computed among activation patterns estimated within the same scanning run and first-level GLM. Within-run RSA has been criticized because it is susceptible to design biases that occur when trial orders are insufficiently randomized within the experiment (Cai et al., 2019). A priori, this was not a strong concern in the current design, as trial orders were fully randomized both within and between participants. Nevertheless, we conducted several diagnostic and robustness analyses to validate that our...
results, and conclusions were not impacted by this potential bias. First, we estimated the extent of the potential bias by running through our RSA pipeline data simulated under a “worst-case” scenario, that is, when SNR = 0 (details described within Extended Data Fig. 1-2) and across a wide range of autocorrelation strengths. At worst, the three models were weakly biased (within 0.02-0.05 of α = 0.05; Extended Data Fig. 1-2). This result indicated that while the bias was present, it was relatively minimal (compare Cai et al., 2019). Second, we validated that these simulated estimates were realistic, by using our actual fMRI data to estimate the false positive rate empirically. To do this, we conducted RSA on the first-level GLM coefficients from each subject’s timecourse. As these voxels should contain no brain activity signal but similar noise characteristics as those of interest (a ventricle mask of 2431 voxels was obtained from AFNI servers: https://afni.nimh.nih.gov/pub/dist/tgz/suma_MNI_N272.tgz). By treating the group-level mean and SD of these RSA model fits as the parameters of a non-central null distribution, $X - N(\beta, SD(\beta))$, we computed the empirical false positive rate as $P(X > z_{0.05})$, the proportion of this distribution greater than the customary one-tailed z criterion. All rates were within 0.03 of $\alpha = 0.05$ (target = 0.07, distractor = 0.04, incongruency = 0.08), confirming the bias was quite minimal.

Between-run RSA. As an alternative to within-run RSA, various between-run estimation approaches have been proposed which have been shown to be less sensitive to potential design biases. We opted not to use between-run RSA for our primary analyses, both because of the reduced effects of design bias established above, but also because between-run RSA need not be considered to be conservative than within-run RSA (Cai et al., 2019). Moreover, several particulars of the present design are known to further hamper its power. Namely, between-run RSA makes incomplete use of the data, an issue that is exacerbated to the maximum extent possible in the present case, as our design has only the minimum number (two) of across-validations folds (runs; Diedrichsen et al., 2020). Additionally, because the image acquisition sequence involved a reversal of phase-encoding direction across the two runs, this effectively adds a strong nonlinear component of noise if between-run RSA is used.

Nevertheless, to examine further the extent of design bias in our data, as well as the robustness of our results to the drop in power imposed by between-run RSA, we conducted a follow-up analysis of our primary results using between-run RSA approaches. We conducted two forms of between-run RSA: the first used “cross-run RSA”, which operates on the cross-correlation of patterns between scanning runs (see Alink et al., 2015), and the second used “cross-validated RSA”, which operates on the inner product of pattern contrasts between runs (see Walther et al., 2016). We selected these two forms of RSA as they have complementary benefits. Cross-run correlation is most comparable to our within-run correlation, as they are both linear correlations. However, using this method within our data set also necessitated using downsampling (as the numbers of trials per condition were not perfectly balanced at the run-level; see Downsampling analysis), which increases the variance of resulting estimates because of discarding data. In contrast, cross-validated RSA is insensitive (in terms of expected value) to the issue of trial numbers per condition (Diedrichsen et al., 2020). Using this method, therefore allowed us to conduct the RSA using all the data at once, without downsampling. But cross-validated RSA tests a more constrained hypothesis than cross-run RSA. Whereas cross-run RSA can be sensitive to nonlinear differences between conditions, cross-validated RSA tests linear discriminability between conditions. Thus, when a nonlinear boundary separates conditions, cross-validated RSA will fail to detect an effect, whereas cross-run RSA could succeed. A nonlinear boundary would occur, for example, when one condition (e.g., incongruent stimuli) drives a reliable, common, response while the other conditions (e.g., congruent stimuli) either drive unreliable, or stable but heterogeneous, responses. Nevertheless, to make cross-validated RSA as comparable as possible to our primary RSA, we z-score standardized patterns before computation and omitted spatial prewhitening (Walther et al., 2016).

Downsampling analyses. Last, we checked whether the primary results were robust to the prewhitening method of data preprocessing, which was introduced to handle the imbalance of trials across runs (see Estimation of coding strength $\beta$, step 4). In this analysis, we instead handled this issue by performing RSA after equating the number of trials per run*condition, by iteratively downsampling conditions with random subsets of trials. Specifically, we first fitted GLMs on fMRI time-series, separately for each scanning run, that contained a single regressor per trial (LS-A method of Mumford et al., 2012). The minimum number of times we presented each unique Stroop stimulus in a single run was 3; this was the number to which we downsampled all conditions with >3 occurrences. For these conditions, we randomly sampled three trials, and averaged GLM coefficients voxelwise over these trials. (For 3-trial conditions, we simply averaged all trials that were present.) This formed 16 separate condition-level coefficient vectors (activity patterns) per run. We then averaged these coefficient vectors across run, then estimated condition × condition correlation matrices from these patterns (averaging across runs was omitted from the downsampled cross-run RSA; see preceding section). We repeated this resampling, averaging, and correlation process 1000 times, and averaged the resulting correlation matrices across iterations (after an arctanh transform). These correlation matrices were then submitted to the same RSA as outlined in Estimation of coding strength $\beta$, with the omission of the prewhitening step (4).

Results

Influential theories of cognitive control have proposed specific dissociations in the type of task information encoded by human medial and lateral frontoparietal cortex (Fig. 1A,B). But previous studies have largely approached this question indirectly, by using 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 using the similarity structure of neural activity patterns evoked within these regions to estimate their informational content. In particular, through RSA, we compared neural coding of three distinct types of Stroop-task information — target, distractor, and incongruency (Fig. 1D,E) — within each 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 representation of task information (Fig. 1B, middle). Second, we examined individual-level effects (i.e., individual differences), to test for dissociations in brain–behavior relationships (Fig. 1B, right). These two analyses were ROI-based and primarily focused on dorsomedial frontal and lateral frontoparietal regions (Fig. 1D), 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 anatomic profile of each task dimension.

Group

DMFC and DLPFC 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 Table 1.

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: $\hat{\beta} = 0.11, p = 0.00$, right: $\hat{\beta} = 0.08, p = 0.00$). Additionally, within left DMFC, a preference for this dimension was observed: incongruency information was encoded more strongly than either target or distractor information (vs target: $\Delta\hat{\beta} = 0.07, p = 0.05$; vs distractor: $\Delta\hat{\beta} = 0.10, p = 0.00$; $p$ values adjusted across hemisphere; Fig. 2A). This DMFC
Table 1. Group-level results from RSA in frontoparietal ROIs.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>β</th>
<th>σ</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMFC (L) incon.</td>
<td>0.11</td>
<td>0.03</td>
<td>3.65</td>
<td>0.00026</td>
</tr>
<tr>
<td>DMFC (R) incon.</td>
<td>0.08</td>
<td>0.03</td>
<td>2.82</td>
<td>0.00485</td>
</tr>
<tr>
<td>DMFC (L) target</td>
<td>0.04</td>
<td>0.01</td>
<td>2.61</td>
<td>0.00918</td>
</tr>
<tr>
<td>DMFC (R) target</td>
<td>0.05</td>
<td>0.01</td>
<td>3.94</td>
<td>0.00008</td>
</tr>
<tr>
<td>DMFC (L) distr.</td>
<td>0.01</td>
<td>0.01</td>
<td>0.78</td>
<td>0.43667</td>
</tr>
<tr>
<td>DMFC (R) distr.</td>
<td>0.00</td>
<td>0.01</td>
<td>0.19</td>
<td>0.84907</td>
</tr>
<tr>
<td>DLPC (L) incon.</td>
<td>0.04</td>
<td>0.02</td>
<td>2.04</td>
<td>0.04155</td>
</tr>
<tr>
<td>DLPC (R) incon.</td>
<td>0.05</td>
<td>0.02</td>
<td>2.39</td>
<td>0.01682</td>
</tr>
<tr>
<td>DLPC (L) target</td>
<td>0.03</td>
<td>0.01</td>
<td>2.56</td>
<td>0.01062</td>
</tr>
<tr>
<td>DLPC (R) target</td>
<td>0.04</td>
<td>0.01</td>
<td>3.18</td>
<td>0.00147</td>
</tr>
<tr>
<td>DLPC (L) distr.</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.98</td>
<td>0.32561</td>
</tr>
<tr>
<td>DLPC (R) distr.</td>
<td>0.00</td>
<td>0.01</td>
<td>0.31</td>
<td>0.75797</td>
</tr>
<tr>
<td>LPPC (L) incon.</td>
<td>0.07</td>
<td>0.02</td>
<td>3.07</td>
<td>0.00211</td>
</tr>
<tr>
<td>LPPC (R) incon.</td>
<td>0.06</td>
<td>0.02</td>
<td>2.64</td>
<td>0.00818</td>
</tr>
<tr>
<td>LPPC (L) target</td>
<td>0.03</td>
<td>0.01</td>
<td>2.74</td>
<td>0.00617</td>
</tr>
<tr>
<td>LPPC (R) target</td>
<td>0.05</td>
<td>0.01</td>
<td>4.38</td>
<td>0.00001</td>
</tr>
<tr>
<td>LPPC (L) distr.</td>
<td>0.00</td>
<td>0.01</td>
<td>0.28</td>
<td>0.77722</td>
</tr>
<tr>
<td>LPPC (R) distr</td>
<td>0.02</td>
<td>0.01</td>
<td>0.86</td>
<td>0.39359</td>
</tr>
<tr>
<td>incon. — target</td>
<td>DMFC (L)</td>
<td>0.07</td>
<td>0.03</td>
<td>2.26</td>
</tr>
<tr>
<td>incon. — target</td>
<td>DMFC (R)</td>
<td>0.03</td>
<td>0.03</td>
<td>0.91</td>
</tr>
<tr>
<td>incon. — distr.</td>
<td>DMFC (L)</td>
<td>0.10</td>
<td>0.03</td>
<td>3.22</td>
</tr>
<tr>
<td>incon. — distr.</td>
<td>DMFC (R)</td>
<td>0.08</td>
<td>0.03</td>
<td>2.64</td>
</tr>
<tr>
<td>target — distr.</td>
<td>DMFC (L)</td>
<td>0.03</td>
<td>0.02</td>
<td>1.63</td>
</tr>
<tr>
<td>target — distr.</td>
<td>DMFC (R)</td>
<td>0.05</td>
<td>0.02</td>
<td>3.18</td>
</tr>
<tr>
<td>incon. — target</td>
<td>DLPC (L)</td>
<td>0.01</td>
<td>0.02</td>
<td>0.46</td>
</tr>
<tr>
<td>incon. — target</td>
<td>DLPC (R)</td>
<td>0.01</td>
<td>0.02</td>
<td>0.44</td>
</tr>
<tr>
<td>incon. — distr.</td>
<td>DLPC (L)</td>
<td>0.06</td>
<td>0.02</td>
<td>2.41</td>
</tr>
<tr>
<td>incon. — distr.</td>
<td>DLPC (R)</td>
<td>0.05</td>
<td>0.02</td>
<td>2.09</td>
</tr>
<tr>
<td>target — distr.</td>
<td>DLPC (L)</td>
<td>0.04</td>
<td>0.02</td>
<td>2.76</td>
</tr>
<tr>
<td>target — distr.</td>
<td>DLPC (R)</td>
<td>0.04</td>
<td>0.02</td>
<td>2.33</td>
</tr>
<tr>
<td>incon. — target</td>
<td>LPPC (L)</td>
<td>0.04</td>
<td>0.03</td>
<td>1.63</td>
</tr>
<tr>
<td>incon. — target</td>
<td>LPPC (R)</td>
<td>0.01</td>
<td>0.03</td>
<td>0.43</td>
</tr>
<tr>
<td>incon. — distr.</td>
<td>LPPC (L)</td>
<td>0.07</td>
<td>0.03</td>
<td>2.75</td>
</tr>
<tr>
<td>incon. — distr.</td>
<td>LPPC (R)</td>
<td>0.04</td>
<td>0.03</td>
<td>1.63</td>
</tr>
<tr>
<td>target — distr.</td>
<td>LPPC (L)</td>
<td>0.03</td>
<td>0.02</td>
<td>1.87</td>
</tr>
<tr>
<td>target — distr.</td>
<td>LPPC (R)</td>
<td>0.03</td>
<td>0.02</td>
<td>1.96</td>
</tr>
<tr>
<td>DMFC (L) — DLPC (L) incon.</td>
<td>0.06</td>
<td>0.02</td>
<td>2.38</td>
<td>0.01737</td>
</tr>
<tr>
<td>DMFC (L) — DLPC (L) target</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.09</td>
<td>0.92467</td>
</tr>
<tr>
<td>DMFC (L) — LPPC incon.</td>
<td>0.04</td>
<td>0.03</td>
<td>1.47</td>
<td>0.14112</td>
</tr>
<tr>
<td>DMFC (L) — LPPC target</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.56</td>
<td>0.57868</td>
</tr>
<tr>
<td>incon. — target</td>
<td>DMFC (L) — DLPC (R)</td>
<td>0.06</td>
<td>0.03</td>
<td>2.20</td>
</tr>
<tr>
<td>incon. — target</td>
<td>LPPC (L) — DLPC</td>
<td>0.03</td>
<td>0.03</td>
<td>1.06</td>
</tr>
</tbody>
</table>

*Group-mean RSA model fits and contrasts between RSA models (target, distractor, incongruency) and between ROIs (DMFC, DLPC, LPPC in each hemisphere). Each row displays the statistical estimates from a given RSA model and ROI, or from a contrast between RSA models or ROIs. Contrasts were conducted either between RSA models (within ROI) or between ROIs (within RSA model). For example, incon. — distr. [DMFC (R)] indicates the contrast between target and distractor RSA model fits (‘coding strengths’) within right DMFC. β indicates the mean RSA model fit over subjects. Unless noted, contrasts were averaged across hemisphere. For contrasts performed separately in each hemisphere, p-values were FDR-adjusted for two comparisons. Products of bracketed terms indicate interactions between RSA model and ROI. incon., Incongruency; distr., Distractor.

# Group incongruency

We next focused on lateral frontoparietal regions and the coding of target information. DLPCF indeed encoded target information (left: β = 0.03, p = 0.01 right: β = 0.04, p = 0.00), and more strongly than distractor information (Δβ = 0.04, p = 0.00). We did not find a preference, however, for target over incongruency information in DLPCF (Δβ = -0.01, p = 0.62), nor did we observe significantly stronger target coding than in DMFC (β = 0.00, p = 0.92). A qualitatively similar pattern of results was observed in LPPC: 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 (Table 1).

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

We next tested a series of hypotheses to scrutinize and extend our results.

First, we conducted a positive control analysis to bolster confidence in the statistical power of RSA methods within the present design. In particular, we sought to determine whether our methods could detect dissociations in task coding that are strongly expected to exist. For this, 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, evidence of preferential distractor coding was observed (Fig. 2C); distractor: β = 0.04, p<0.001, target: β = 0.02, p = 0.17, distractor vs target: Δβ = 0.03, p = 0.05), whereas, within primary ventral somatomotor cortices (encompassing the “mouth” homunculi), a relatively selective pattern of target coding was found (Fig. 2C); distractor: β = 0.00, p = 0.9, target: β = 0.13, p<0.001, distractor vs target: Δβ = -0.13, p<0.001). Further, representation of these dimensions was strong enough to predominate the overall structures of patterns within low-dimensional embeddings (Fig. 2B). 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 DLPCF was not because of a general deficiency in our distractor coding model.

Second, we conducted sensitivity analyses to assess the robustness of our results to the particular ROI definitions used. In one analysis, we tested more expansive ROI definitions, by using alternatively-defined superparcels (Extended Data Fig. 1-1). These definitions included additional, more rostral PFC parcels (1 in DMFC, 3 in DLPCF), 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 (Extended Data Fig. 2-1). In the other analysis, we examined whether the overall superparcel coding profiles were representative of individual parcels. While DMFC and DLPCF results generally reflected that of constituent parcels (Extended Data Fig. 2-2A,B), interestingly, there was substantial heterogeneity within left LPPC (Extended Data Fig. 2-2C). Similar to left preference, incongruency coding was neuroanatomically dissociated, as this coding scheme was reflected more strongly in left DMFC than in DLPCF (Δβ = 0.06, p = 0.02; Table 1).
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).

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 et al., 2018). It is possible that we might have observed a dissociation in left DMFC but not DLPFC because of differential levels of statistical power across the two regions. We therefore derived an SNR analysis to determine whether this was a viable explanation (see Materials and Methods). A paired-sample bootstrap test did not indicate a systematic difference between DLPFC versus left DMFC group-level SNR ($\Delta z = -0.01. p = 0.60$). However, the SNRs in these regions were also not confirmarily 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.

Fourth, to provide a basis for comparison to most extant neuroimaging research of the Stroop task, we conducted a “univariate activation” analysis, examining whether these brain regions were generally more active during incongruent versus congruent conditions. No regions were found to respond more strongly overall to incongruent versus congruent conditions (Extended Data Fig. 2-3A), although the mean contrast in DMFC (L) was positive (i.e., incongruent > congruent). This null result was not surprising, however, because of the high frequency of incongruent trials within the experiment — which is known to reduce both the behavioral and neural univariate Stroop effect (Logan and Zbrodoff, 1979; Carter et al., 2000; De Pisapia and Braver, 2006). While this null result demonstrates the utility of using RSA in this case, it should not, however, be seen as direct evidence for the increased sensitivity of RSA versus univariate methods, as the univariate and RSA-based tests as implemented here are subject to different constraints and are thus incomparable (Allefeld et al., 2016). Finally, the magnitude of the univariate Stroop effects was only weakly correlated to incongruency coding model fits (Extended Data Fig. 2-3B), suggesting that these measures were nonredundant.

Finally, we tested whether these patterns of results were robust to alternative RSA techniques, including a downsampling technique to equate trial counts across runs, two “between-run” RSA methods (cross-run RSA and cross-validated RSA), and alternative parameterizations of the incongruency coding model (see Materials and Methods). Findings were robust to downsampling and to between-run RSA (Extended Data Fig. 2-4), and were highly similar across different parameterizations (Extended Data Fig. 2-5). Interestingly, however, the detection of incongruency coding in DMFC depended on whether a linear or nonlinear RSA method was used. When using an RSA method that tests linear discriminability between conditions (cross-validated RSA), the incongruency coding effect was abolished in DMFC; whereas when using a comparable method that is sensitive to nonlinear pattern differences (cross-run RSA), the effect remained quite strong. This pattern of results suggests that incongruency information was encoded nonlinearly within DMFC activation patterns. Indeed, this can be seen within the two-dimensional embedding (Fig. 2), as a radial, rather than linear, separation of incongruent (central) and congruent (peripheral) stimuli.

### Table 2. Parameter estimates from hierarchical brain–behavior models that explained individual differences in behavioral performance (RT) with variability in the strength of neural coding ($\beta$, see Method)$^d$

<table>
<thead>
<tr>
<th>Contrast</th>
<th>b</th>
<th>$\sigma$</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMFC (L) incon.$^c$</td>
<td>47.66</td>
<td>25.97</td>
<td>1.84</td>
<td>0.06647</td>
</tr>
<tr>
<td>DMFC (R) incon.</td>
<td>$-0.10$</td>
<td>33.15</td>
<td>0.00</td>
<td>0.95762</td>
</tr>
<tr>
<td>DMFC (L) target</td>
<td>48.40</td>
<td>90.63</td>
<td>0.53</td>
<td>0.59330</td>
</tr>
<tr>
<td>DMFC (R) target</td>
<td>$-24.04$</td>
<td>63.75</td>
<td>-0.38</td>
<td>0.70613</td>
</tr>
<tr>
<td>DLPFC (L) incon.</td>
<td>14.08</td>
<td>41.17</td>
<td>0.34</td>
<td>0.73227</td>
</tr>
<tr>
<td>DLPFC (R) incon.</td>
<td>38.42</td>
<td>33.66</td>
<td>1.14</td>
<td>0.25367</td>
</tr>
<tr>
<td>DLPFC (L) target</td>
<td>$-105.93$</td>
<td>78.34</td>
<td>-1.35</td>
<td>0.17637</td>
</tr>
<tr>
<td>DLPFC (R) target$^c$</td>
<td>$-239.13$</td>
<td>71.68</td>
<td>-3.30</td>
<td>0.00085</td>
</tr>
<tr>
<td>LPPC (L) incon.</td>
<td>$-48.95$</td>
<td>38.13</td>
<td>-1.28</td>
<td>0.19921</td>
</tr>
<tr>
<td>LPPC (R) incon.</td>
<td>$-32.40$</td>
<td>32.22</td>
<td>-1.01</td>
<td>0.31467</td>
</tr>
<tr>
<td>LPPC (L) target</td>
<td>1.90</td>
<td>93.65</td>
<td>-0.02</td>
<td>0.98385</td>
</tr>
<tr>
<td>LPPC (R) target$^c$</td>
<td>$-235.34$</td>
<td>71.23</td>
<td>-3.30</td>
<td>0.00096</td>
</tr>
</tbody>
</table>

Separate models were fit per region and RSA model (target, incongruency) combination. The displayed estimates correspond to the interaction stroop · $\beta$, which indicates how subjects’ coding strengths relate to the magnitude of their Stroop interference effects. Also displayed are contrasts on stroop · $\beta$ estimates, between RSA models (within ROIs) or between ROIs (within RSA model). For example, target - incon. | DLPFC (R) indicates, for right DLPFC, whether target and incongruency coding were differentially related to the size of the Stroop interference effect.$^b$

$^a$ Contrast with $p < 0.05$.

$^b$ Contrast with $0.05 < p < 0.1$.

**Individual**

Primary individual-level results are summarized in Figure 3. Statistical estimates corresponding to results outlined within this section are contained in Table 2 (for scatter plots of all associations, see Extended Data Fig. 3-1).

**Better-performing subjects have stronger lateral frontal-parietal target coding**

The fidelity of target-related information in lateral frontal-parietal 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 ($\beta_{target}$) 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 resolved Stroop interference effects more quickly (DLPFC: $b = -239.13. p = 0.00. r = -0.44 [-0.66, -0.15]; \rho = -0.39 [-0.63, -0.08];$ LPPC: $b = -235.34. p = 0.00. r = -0.44 [-0.63, -0.17]; \rho = -0.35 [-0.59, -0.06];$ values corrected across hemisphere; linear correlation, $r$; rank correlation, $\rho$; bootstrapped 95% CI. [lower, upper]; Fig. 3A; Table 2). Subjects’ target coding estimates were moderately correlated ($r=0.53$) between these two lateral frontal-parietal regions, as expected based on their strong neuroanatomical and functional connectivity (e.g., Petrides and Pandya, 1999; Buschman and Miller, 2007). In neither of these lateral regions was incongruency coding significantly related to the Stroop interference effect (DLPFC: $b = 38.42. p = 0.25. r = -0.17 [-0.14, 0.45]; \rho = -0.10 [-0.19, 0.37];$ LPPC $b = -32.40. p = 0.31. r = -0.15$...
Table 2

Fig. 3-2

A data-driven model selection analysis was conducted to
address this question (see Materials and Methods). We defined
an expanded set of 24 cortical regions (superparcels), including
the six defined and used earlier, which covered various areas that
may be important for performing the Stroop task (Fig. 2B;
Extended Data Fig. 3-3). Conducting RSA on each superparcel
furnished three coding estimates (one per coding model) per
superparcel. These 72 estimates were then used as features in a
cross-validated model selection procedure.

Strikingly, the selected model contained all three hypothe-
sized sizes: (right) DLPFC and LPPC target coding, and
(left) DMFC incongruency coding. In addition, two unexpected
measures were identified, both with negative slopes: left DLPFC
distractor coding ($r = -0.43 [-0.66, -0.14]$, $\rho = -0.33 [-0.60,
-0.04]$) and left ventral visual incongruency coding ($r =
-0.35 [-0.52, -0.13]$, $\rho = -0.43 [-0.62, -0.20]$); bivariate cor-
relations shown in Fig. 3C). We tested the predictive accuracy of
the 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 (Fig. 3D). While not fully independent, these val-
idation set data were from a true hold-out and obtained from dist-
inct individuals. This result therefore bolsters claims regarding
the stability of the selected model.

Nevertheless, to provide a cursory test of a truly independent
validation set (i.e., with no familial dependency to the training
set), we excluded all subjects in the training set who were co-
twins of those in the validation set, and re-conducted this
model selection procedure. This amounted to discarding 16/49
(33%) of training-set observations. The selected model contained
only one variable, which was not in our ROIs (but in early visual
cortex), and was unable to predict held-out Stroop effects ($r =
-0.12$). This is perhaps unsurprising, however, given the sub-
stantial reduction in the size of the training dataset for an already
high-dimensional model. To reduce the dimensionality, we re-
conducted this analysis, focusing now instead on ROIs and
coding schemes of interest — target coding in DLPFC and
LPPC, and incongruency coding in DMFC (within each hemi-
sphere) — and additionally ensured that all variables were used
in prediction (via ridge regression). This model was better able
to predict the held-out Stroop effect ($r = 0.20$), in particular,
relative to a comparable model that contained theoretically “mismatched” ROI×coding scheme combinations (incongruency coding in DLPFC and LPPC, target coding in DMFC; $r = -0.17$; bootstrapped $p = 0.088$).

Results from these model selection analyses 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 whole-cortex RSA

In a final, exploratory analysis, we estimated RSA models 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 andtemporal lobes, including many perisylvian regions (Fig. 4, top; Extended Data Fig. 4-1). Notably, the strongest target coding was observed within regions that receive strong sensory and (or) motor-related input (Extended Data Fig. 4-1). 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 (Fig. 4, middle; Extended Data Fig. 4-3). Aside from this occipital area, these incongruency-coding parcels notably belonged to control networks (frontoparietal, cinguloopercular, and dorsal attention; Extended Data Fig. 4-3). In a third, distinct profile, distractor coding was only observed within early visual cortex (left V1 and V2; Fig. 4, bottom; Extended Data Fig. 4-2).

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. 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.

Finally, we repeated this exploratory analysis using alternative RSA methods involving downsampling and between-run RSA (see Downsampling analysis, and Between-run RSA; Extended Data Fig. 4-4). Across these analyses, the core results were quite robust: we found incongruency coding in parcels within DMFC (though this depended on the use of nonlinear RSA methods, as in Extended Data Figs. 2-4, 3-2), target coding in mid-DLPC, and distractor coding in visual cortex. Our findings were therefore not specific to a particular estimation method.

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 anatomic profiles were identified across our three coding models, further suggesting that these models were successful in measuring coding of distinct task dimensions (Fig. 1A).

Discussion

We analyzed the similarity structure, or representational geometry (Kriegeskorte and Kievit, 2013), of frontoparietal activity patterns associated with cognitive control, during performance of the classic color-word Stroop task. In left DMFC, incongruency coding predominated. While DLPFC and LPPC encoded both target and incongruency-related information, distractor coding was not detected in these regions but was instead 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. Further, in a more comprehensive predictive model that included coding measures from a wide set of 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 (MacDonald et al., 2000; Floden et al., 2011). Most prominently, MacDonald et al. (2000) used a modified Stroop-task design, in which task rules (delivered via precues) 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 DLPFC encodes task-set and rule-related information in a preparatory manner, whereas DMFC encodes incongruency and conflict-related information in a stimulus-evoked manner. In the current study, we leveraged the high spatial dimensionality of fMRI to test 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 Floden et al., 2011; $N = 9$ in MacDonald et al., 2000) and indicate that the dissociations were not dependent on the use of cued task-switching designs. Synthesizing these prior findings with those of the present study hints at a continuity in the putative role of DLPFC during target selection. Rather than exclusively contributing to preparation, DLPFC coding may evolve from proactively representing abstract rule or set-related information, towards more concrete targets and behavioral choices as relevant stimulus information becomes available in the environment. This view accords with work in monkey neurophysiology (Mante et al., 2013; Rigotti et al., 2013; Stokes et al., 2013), yet further work is needed to determine whether similar dynamics occur within human DLPFC and how such dynamics may reflect or interact with specific processes of cognitive control in Stroop-like tasks.

More broadly, the results of this study highlight the utility of RSA and the general representational geometric framework for investigating cognitive control. Previous work has used MVPA decoding in the Stroop task to study the impact of control demand on posterior representations (e.g., Banich et al., 2019). Here, we used the RSA framework to explicitly model and decompose control-related frontoparietal representations. Indeed, a major motivation of the current study was to assess how well RSA measures of frontoparietal coding map to theorized mechanisms of control. 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 and Cohen, 2001; Ridderinkhof et al., 2004; Shenhav et al., 2013). Our results were
generally in line with these accounts, joining with a growing body of research in suggesting that RSA provides a convenient yet powerful framework from which neural measures can be used to test cognitive control theory (for review, see Freund et al., 2021).

Nevertheless, the current work represents only an initial step in using the RSA 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. For instance, a key limitation of the current study was the finding of widespread coding of the target dimension, suggesting a lack of specificity in the target 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 DLPFC and LPPC. Yet, this limitation could be addressed more powerfully by experimental design. Adding specific factorial manipulations, such as a task rule manipulation (see MacDonald et al., 2000; Hall-McMaster et al., 2019) or a response modality manipulation (see Minxha et al., 2020; see also Barch et al., 2001), would enable a richer, more precise set of cognitive control-relevant coding variables to be estimated (Fig. 5).

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 predicted coding profiles emerged in some frontoparietal ROIs, all regions encoded incongruency and target information. This incongruency-coding finding is consistent with prior univariate fMRI research (e.g., Nee et al., 2007; Niendam et al., 2012) and a more recent finding that the responses of single neurons in human dACC and DLPFC are robustly modulated by conflict (Smith et al., 2019). With respect to target coding, however, 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, for example: in macaque LIP (Roitman and Shadlen, 2002), in macaque caudal DLPFC (Mante et al., 2013), and in human dACC and pre-SMA (Minxha et al., 2020; see also Okazawa et al., 2021). 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.

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 (Fig. 3B, left). 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 that are strongly tuned to target features, activate as a function of Stroop incongruency (Egner and Hirsch, 2005). In our task, mid-ventral stream areas may have received biasing input, selectively on incongruent trials, which 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.

Perhaps the most surprising result was the robust negative correlation observed between left DLPFC distractor coding and the behavioral Stroop effect (Fig. 3B, right). 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 were 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 (Melara and Algom,

---

**Figure 5.** Expanded experimental designs. The current target coding model (top, green text; see also Fig. 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 (i.e., participants respond to the present set of stimuli, but under both color-naming and word-reading conditions; MacDonald et al., 2000; light purple arrows and text), 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 (green 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 on 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 (bottom right quadrant resembles our “distractor coding” model in Fig. 1, as word features become target in the word-reading condition). Importantly, too, abstract coding of task rule would also now be distinguishable (“rule” matrix on left). 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 abstract choice options, independent of effecter, could be separated from effecter-dependent motoric coding. For an elegant example of a similar manipulation, see Minxha et al. (2020).
2003; Bugg and Crump, 2012). Indeed, subjects clearly do exploit these associations, as indicated by reduced Stroop effects for stimuli that are “mostly incongruent,” also known as “item-specific proportion congruency” effects (ISPC; Bugg et al., 2011; Jiang et al., 2015; see also Crump and 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 DLPCF 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 goal-relevant coding schemes that originate in LPFC before target-stimulus onset as abstract rule or context coding, but which may morph into target coding after 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 DMPFC or subcortical structures (Jiang et al., 2015; Chiu et al., 2017). As suggested here, it may be possible to measure correlates of these neural coding schemes via RSA.

Experimental factors that encourage 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.

References


