Complementary benefits of multivariate and hierarchical models for identifying individual differences in cognitive control

Michael C. Freund*[1,3], Ruiqi Chen[2], Gang Chen[5], and Todd S. Braver[1,3,4]

[1]Department of Cognitive, Linguistic, and Psychological Sciences, Brown University
[2]Division of Biology and Biomedical Sciences, Washington University in St. Louis
[3]Department of Psychological & Brain Sciences, Washington University in St. Louis
[4]Department of Radiology, Washington University in St. Louis
[5]Scientific and Statistical Computing Core, NIMH, NIH, Bethesda, MD, USA

April 26, 2024

Abstract: Understanding individual differences in cognitive control is a central goal in psychology and neuroscience. Reliably measuring these differences, however, has proven extremely challenging, at least when using standard measures in cognitive neuroscience such as response times or task-based fMRI activity. While prior work has pinpointed the source of the issue — the vast amount of cross-trial variability within these measures — no study has rigorously evaluated potential solutions. Here, we do so with one potential way forward: an analytic framework that combines hierarchical Bayesian modeling with multivariate decoding of trial-level fMRI data. Using this framework and longitudinal data from the Dual Mechanisms of Cognitive Control project, we estimated individuals' neural responses associated with cognitive control within a color-word Stroop task, then assessed the reliability of these individuals' responses across a time interval of several months. We show that in many prefrontal and parietal brain regions, test–retest reliability was near maximal, and that only hierarchical models were able to reveal this state of affairs. Further, when compared to traditional univariate contrasts, multivariate decoding enabled individual-level correlations to be estimated with significantly greater precision. We specifically link these improvements in precision to the optimized suppression of cross-trial variability in decoding. Together, these findings not only indicate that cognitive control-related neural responses individuate people in a highly stable manner across time, but also suggest that integrating hierarchical and multivariate models provides a powerful approach for investigating individual differences in cognitive control, one that can effectively address the issue of high-variability measures.

*Address correspondence to michael_freund@brown.edu

1 Introduction

A major goal within psychology and neuroscience is to understand the mechanisms that give rise to psychological diversity. How are two minds alike or distinct, in terms of psychological processes? What neural mechanisms underlie this variability? Such questions of *individual differences* are incredibly important to study: not only do they intrinsically interest many, but they also provide a means to test virtually any cognitive psychological theory (*cf.*, Underwood, 1975), as well as yield direct clinical and educational applications (Diamond, 2013; Engle, 2002; Cole et al., 2014). Spurred by these motivations, cognitive neuroscientists have often sought to identify measures of human brain activity that can be used as markers of individual differences. Yet, this enterprise has proven highly difficult, at least when using common non-invasive measures of brain activity, such as task-based fMRI.

One of the main challenges in studying individual differences can be traced to issues of measurement. In the laboratory, people may score differently on experimental tasks due to different reasons, many of which have nothing to do with cognitive properties of the individual that are stable over time, here termed *cognitive traits*, but instead with properties of the particular measurement occasion, here termed nuisance or *noise* factors. Thus, when reasoning about cognitive traits, researchers run the risk of misattributing a particular finding or pattern of results that in fact arise from more transient factors.

The statistic of *test-retest reliability* is instrumental in mitigating this risk. In particular, test–retest reliability is estimated by repeatedly acquiring the same measures from the same sample of individuals over an extended period of time (typically on the order of days, at minimum). In such repeated-measures designs, differences observed between individuals that are constant over *test repetitions* (also known as “sessions”) are assumed to reflect traits, while changes within individuals over repetitions are assumed to reflect noise (which may include systematic longitudinal effects, such as learning). The relative proportion of trait variance defines the test-retest reliability, or the “traitness” of the measure in question. Yet, despite this grounding theoretical framework, in practice it has not been easy to identify strong individual differences in cognition (conventionally, $r > 0.7$; Matheson, 2019). This difficulty has been particularly salient with measures derived from classical cognitive experimental tasks (Hedge et al., 2018) and adaptations of these paradigms for fMRI. Perhaps most confoundingly, however, poor test-retest reliability has often been reported in psychological domains overwhelmingly assumed to be subject to strong individual differences, for example cognitive control and working memory (Elliott et al., 2020). Such failures have invited pessimism over the usefulness of these popular methods for individual differences research (Elliott et al., 2020).

Much of these pessimistic conclusions, however, overlook a nuanced theoretical issue concerning the structure of noise variability: it is hierarchical. Moreover, ignoring this hierarchical structure will lead to an overly pessimistic estimate of reliability (Chen et al., 2021). To elaborate, within the standard test-retest design, noise variability can be decomposed into (at least) two levels: trial-level variability (i.e., changes over trials, within repetition) and repetition-level variability (i.e., changes over repetitions, aggregated over trials). Overwhelmingly, prior work has estimated reliability through a two-stage summary statistic approach, which involves first aggregating a measure over trials, then assessing the relative proportion of trait variability via linear correlation (see Figure 1 for a walkthrough; see also Haines et al., 2020). But, because the estimation of the linear correlation does not use any information regarding the amount trial-level variability, the summary statistic approach essentially “conceals” the negative impact such variability has on estimated test–retest reliability. In other words, summary statistic estimates of test–retest reliability are underestimated (Figure 1). The amount of underestimation depends in part on how strongly the measure differs across trials, and how effective the first-step aggregation was at reducing the standard errors of the scores. Theoretically, it is possible that the bias is small, but true physiological and behavioral measures vary too strongly over trials for this approach to be accurate in reality (i.e., given attainable amounts of data; Rouder et al., 2023).

This consideration has motivated an alternative, hierarchical Bayesian modeling approach for assessing individual differences (Chen et al., 2021). A hierarchical Bayesian approach does not attempt to squash trial-level variability through aggregation, but instead, it attempts to directly estimate the magnitude of each component of variability from the disaggregated trial-level measures. In this way, the negative impact of trial-level noise on reliability can be explicitly factored out (Figure 1B). In line with this reasoning, results from studies adopting a hierarchical Bayesian approach to individual differences suggest a more optimistic outlook, in which individual differences can in some cases be identified after accounting for the contributions of trial-level noise (Chen et al., 2021; Haines et al., 2020; Snijder et al., 2023).

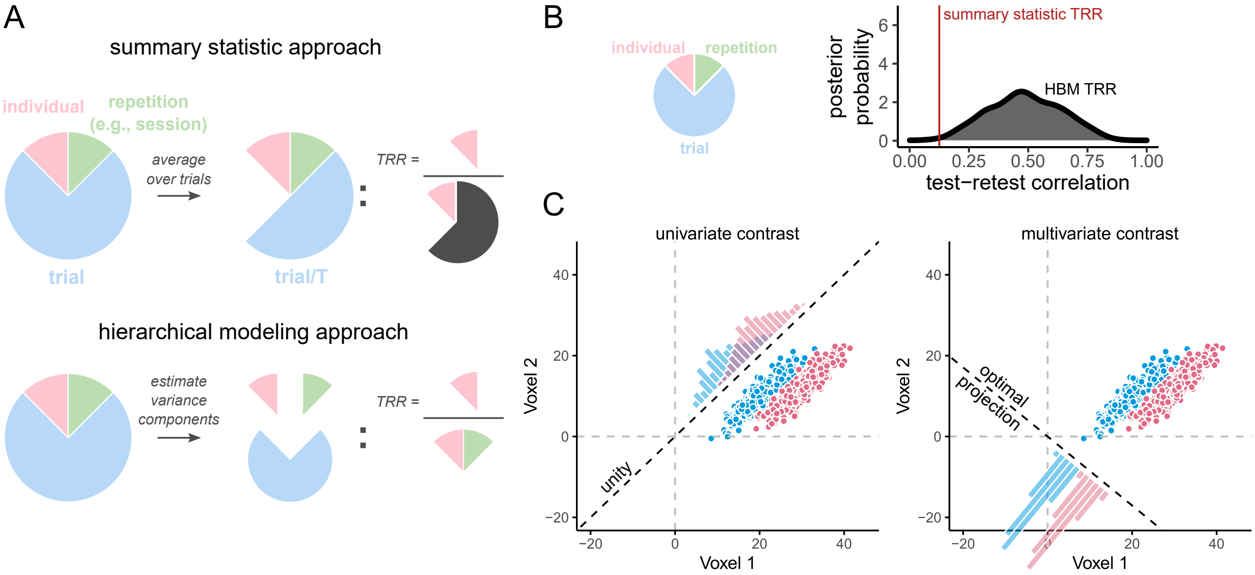

Figure 1: See following page for caption.

Nevertheless, these studies have also illustrated that the hierarchical Bayesian approach is not a panacea for investigating individual differences with high-variability measures (Rouder et al., 2023). This is because the issues caused by high trial-level variability are not fully circumvented in hierarchical Bayesian models. Instead, the issues are redirected, so that they manifest differently within the model estimates. That is, instead of diminishing reliability, increasing trial-level variability leads to *more uncertainty* in the reliability estimate (Figure 1C). Within a recent report, for example, hierarchical Bayesian reliability estimates of fMRI contrasts within select regions of interest were typically higher than their summary statistic counterparts, but were also highly uncertain, with the lower-bound confidence interval generally located well below zero (Chen et al., 2021). Thus, although this "redirection" is useful in that it yields unbiased estimates of test–retest reliability, it does not allow individual difference researchers to avoid grappling with issues of overwhelming trial-level variability.

A plausible remedy for these issues is provided by multivariate pattern analysis (MVPA) decoding. By exploiting the high spatial dimensionality of fMRI, decoding can suppress trial-level variability by identifying dimensions of neural coding that are relatively insensitive to such fluctuations (Figure 1D). Yet, prior work on individual differences has almost entirely eschewed these noise-suppressing decoding techniques, and has instead opted to use more traditional univariate models, which can be much more vulnerable to trial-level variability. This combination of facts suggests a constructive hypothesis: MVPA decoding can improve the power of individual differences analyses by increasing the certainty (i.e., precision) of the estimated trait-level parameters. Some indirect evidence is indeed in line with this hypothesis (Kragel et al., 2021; Xu et al., 2018; Yoo et al., 2019), however, a direct and more extensive investigation of these issues is warranted, in particular within the domain of task-related fMRI (*cf.*, resting state).

Here, using a subset of densely-sampled individuals (N = 27, 6 sessions each) acquired as part of the Dual Mechanisms of Cognitive Control fMRI dataset (Braver et al., 2021), we compared the sensitivity of various approaches to individual differences analyses using a classical fMRI measure of cognitive control function, the Stroop-effect contrast (incongruent versus congruent trials), as derived from the color-word Stroop (1935) task. We compared the reliability of a traditional univariate version of this measure to a multivariate contrast, obtained from an MVPA decoder, that we specifically constructed to suppress trial-level variability. To provide an accurate and complete comparison, we performed these analyses within the context of a hierarchical Bayesian modeling framework for individual differences (Chen et al., 2021) and subsequently compared the results to those derived from a summary statistic approach. In line with theory, we found that relative to univariate models, MVPA decoding improved the ability to identify strong individual differences in fMRI measures of cognitive control from frontoparietal cortex. In a hierarchical Bayesian framework, these improvements manifested as increased estimation certainty (i.e., precision) of individual difference parameters. These results highlight the efficacy of a joint hierarchical Bayesian–multivariate decoding approach, in which the statistical accuracy of hierarchical Bayesian models is boosted by the statistical power of MVPA decoding, and suggest that non-invasive neural measures like task-based fMRI may be an adequate means for addressing questions regarding individual differences in cognition.

Figure 1: **Graphical intuition for test-retest reliability and benefits of multivariate pattern analysis. A, B**, Two approaches to estimating test-retest reliability: summary statistic versus hierarchical modeling. Psychological experimental tasks are typically composed of hierarchically structured "levels": trials are completed within testing repetitions (e.g., sessions) within individuals. Measures derived from such tasks exhibit some amount of variability at each of these levels: for example, a person's performance may fluctuate over trials, across different days (repetitions), while also reliably differing over time from another person's average performance (individual differences). Rigorously studying individual differences requires disentangling individual variability from variability at the other levels. This extent to which a measure supports such disentangling is quantified by test–restest reliability (TRR). **A, top**, The summary statistic approach estimates reliability by first aggregating over trials per individual, attempting to quash trial-level variability prior to computing the proportion of individual-level variance. Despite this aggregation, however, a potentially substantial amount of residual trial-level variability remains "concealed" in the aggregated scores, and thereby in the denominator of the reliability statistic. **A, bottom**, In hierarchical approaches, however, the variance is decomposed into separate components, and trial-level variability is explicitly factored out of the test–retest reliability computation. **B**, When trial-level variability is high (pie chart), the test–retest reliability estimated through summary statistical approaches (red vertical line) will tend to be shrunken relative to estimates from hierarchical Bayesian models (HBM; the central tendency of the grey density indicates the average posterior test–retest correlation). By contrast, in hierarchical models, such heightened trial-level variability would instead manifest as decreased estimation precision (spread of grey density). Unfortunately, both of these scenarios would impair inferential power. **D**, Common forms of multivariate pattern analysis attenuate trial-level variability. Left and right panels show the same (simulated) fMRI data, acquired from a hypothetical region of interest composed of two voxels (x and y axes). Individual trials of a Stroop task evoked particular patterns of activity across these two voxels (points). Each trial belongs to one of two conditions, *incongruent* (red) or *congruent* (blue), whose difference forms the Stroop effect contrast. Both univariate analysis and multivariate pattern analysis can be used to compute the Stroop effect contrast. **Left panel**, In a univariate analysis, measures from different voxels are aggregated together *uniformly*, by taking the spatial mean (i.e., each voxel is weighted equally then summed). Graphically, this corresponds to projecting the condition means onto the unity line, which runs through the origin and $x = 1, y = 1$ (diagonal dotted line). Yet, depending on the shape and relative configuration of these distributions, such a projection may lead to a highly variable measure (overlapping histograms along unity line). **Right panel**, In a multivariate pattern analysis, measures from different voxels can be aggregated together *optimally*. Graphically, this corresponds to projecting the condition means onto the line that yields the most separation between projected classes (i.e., minimizing overlap between the histograms of the projections). As a result, this procedure can lead to projections in which trial-level variance is squashed relative to univariate analysis.

2 Method

2.1 Subjects

Data were obtained as part of the Dual Mechanisms of Cognitive Control project (Braver et al., 2021). All subjects consented to participate in the study under Washington University IRB protocols. At the time of analysis, 32 subjects had completed the test–retest component of this project. We selected 27 of these subjects (number of females = 16, age range at initial session = [19, 42]) to include in the present analyses. This selection ensured that all subjects in our sample had complete data for all sessions that met minimal quality-control criteria (for each scanning run, complete behavioral and fMRI measures, low numbers of missed responses, and motion levels and dropout artifacts that were qualitatively judged to be modest). Of these subjects, 15 had also participated in the Human Connectome Project Young Adult study Van Essen et al. (2013).

2.2 Design

As much of the debate over test-retest reliability has focused on the color-word Stroop (1935) task as a paradigmatic example (e.g., Hedge et al., 2018; Rouder and Haaf, 2019; Haines et al., 2020), we focused exclusively on data from this task within the larger DMCC project dataset Braver et al. (2021). In this task, names of colors were visually displayed in various nameable hues, and subjects were instructed to "say the name of the color, as fast and accurately as possible; do not read the word." The primary manipulation concerns the *congruency* between the meaning of the text and the hue in which it is rendered. A colored word is either "congruent", such that the hue corresponds to the hue the text expresses (e.g., "RED" in red font), or "incongruent" (e.g., "GREEN" in blue font).

The Stroop dataset within the DMCC project is highly amenable to assessing test-retest reliability due to the extensive amount of repeated measures acquired. Each subject completed several hundred trials of this task within each of six scanning sessions, administered on different days. By design, these six scanning days were clustered into two "repetitions" of three sessions – an initial "test", then subsequent "retest" repetition. Across subjects, the time between repetitions spanned 36–1558 days (median = 169, IQR = 258). Assessing test–retest reliability across these repetitions thus reflects a relatively challenging benchmark for consistency.

These sessions were largely similar within each repetition. Each consisted of two scanning runs of approximately 12 minutes that contained a minimum of 108 trials, with inter-trial intervals sampled with uniform probability from one, two, or three TRs (1.2, 2.4, or 3.6 s). The same set of eight words, and set of eight corresponding hues, were used in each session as stimuli. Yet by design, the sessions also subtly differed in ways that influenced the size of the behavioral congruency effects that they elicit. Although these manipulations were conducted to investigate questions outside of the scope of the present study, we exploit them in our test–retest reliability analyses. In what we have referred to as the "baseline" type of session (216 trials), the first session within each repetition, behavioral congruency effects were maximized (Braver et al., 2021). This maximization was due to the low frequency of incongruent relative to congruent trials (33%; Logan and Zbrodoff, 1979). The other two "proactive" and "reactive" sessions (216, 240 trials), which were counterbalanced in order across participants, manipulated expectancies regarding incongruent trials. The expectancy manipulations were accomplished by increasing the relative percentage of incongruent versus congruent trials, either in a session-wide manner (to 67%, proactive session), or selectively for specific colors (while keeping the session-wide percentage at 30%, reactive session). The theoretical reasons for the manipulations are described in detail within Braver et al. (2021). We exploit the robust congruency effects elicited by the baseline session in our analyses, by using this session in particular to evaluate test–retest reliability of our univariate fMRI measures (see Section 2.5).

2.3 Image acquisition and preprocessing

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 1.2 s TR, no GRAPPA, ipat = 0), then subjected to standard fMRIPrep pipelines (Esteban et al., 2020, 2018). As part of the preprocessing pipeline, data were projected into surface space (fsLR8K) once, were then smoothed (with Gaussian kernel of FWHM = 4 mm), and were divisively normalized to reflect percent signal change relative to the timeseries mean. These pipelines were implemented in a Singularity container (Kurtzer et al., 2017) with additional custom scripts used to implement file management. More details on the preprocessing and pipeline are available at `https://osf.io/6p3en/` and (Etzel et al., 2022). Container scripts are available at `https://hub.docker.com/u/ccplabwustl`.

2.4 Timeseries models

To generate single-trial estimates of the fMRI activation, we summarized the minimally preprocessed fMRI timeseries using a simple "selective averaging" model. Specifically, we averaged together the second, third, and fourth observations (i.e., time-points in TR) following the onset of a trial (i.e., corresponding to a window of 2.4–4.8 s post-stimulus onset). Prior to selective averaging, we detrended the timeseries for a given participant and repetition with 5-th order polynomials per run (with order selected by $1+\text{floor}(D/150)$, where D is the duration in seconds of a run), as well as with 6 motion parameters concatenated in time over both runs. This detrending was performed via 3dDeconvolve in *AFNI*. Additionally, trials that had a TR within the averaged window with frame-wise displacement > 0.9 mm were censored.

Although there are several other methods for obtaining trial-level fMRI observations (Mumford et al., 2012), we opted for this simple selective averaging model for two reasons. First, the selective averaging model can be extended easily within future work to the other tasks within the DMCC dataset. These tasks have relatively complex, multi-event trial structures that would be difficult to model with fixed-shape regressors. Second, compared to other GLM-based approaches that simultaneously model all trials with individual regressors (i.e., a "least-squares–all" approach; Mumford et al., 2012), the precision of the single-trial estimates in the selective averaging method is not diminished by collinearity within the design matrix. Thus, relative to least-squares–all, the single-trial estimates furnished by selective averaging may be more biased by neighboring trials' activity; however, given that trial-types were well-randomized within each run, we assumed that such bias was likely minimal and less costly in terms of imprecision than using a method such as least-squares–all. (This assumption was also supported by pilot analyses conducted on a subset of the data.)

2.5 Spatial brain activity models

Here, we provide an overview of the univariate and multivariate modeling approaches we used. For a more comprehensive mathematical description, please refer to the Supplemental Method.

Once we divided the single-trial activation estimates into Schaefer atlas regions (Schaefer et al., 2018), we centered these estimates (see Supplemental Method Section 10.2.1), then used two different methods of summarizing the spatial patterns within each region: a univariate model and a multivariate model. Each of these models aggregates information that is spatially distributed across all vertices within a given region. Thus, we refer to them as "spatial models" (equivalently, they could be referred to as "spatial filters"). In particular, these spatial models perform linear dimensionality reduction, in which the spatial pattern on each trial is projected onto a single dimension, yielding a single summary score, per trial and region. In other words, both types of models compute a weighted sum across vertices:

$$\text{score}_t = \sum_{v=1}^{V} \text{activ}_{vt} \cdot \text{weight}_v \tag{1}$$

Here, activ_{vt} is the activation pattern estimates furnished by the timeseries model on trial t, in vertex v out of V total vertices within the region of interest.

The key difference between these models lies in the nature of the weights. In the univariate spatial model, the weights are all positive and equal across vertices:

$$\text{weight}_v = 1/V \tag{2}$$

This projection amounts to estimating the spatial mean.

In the multivariate spatial model, the weights are estimated such that they optimize a criterion. We used linear discriminant analysis (LDA; Fisher, 1936; Hastie et al., 2009, which finds the weights that maximize, within the projection scores, the variability *between the condition means* relative to the variability *within the conditions*. In other words, weights are learned that suppress trial-level variability relative to trial-averaged variability in the scores. Notably, this criterion aligns well with our goal of obtaining a favorable *trial/subject* variability ratio. Such weights are provided by the linear discriminant function within LDA (Equation 6 in Supplemental Method). Here, we simply write this function for a single vertex v:

$$\text{weight}_v = (\text{mean incon activ}_v - \text{mean congr activ}_v) \cdot \text{scaling factor}_v \tag{3}$$

where mean incon activ and mean congr activ are the mean levels of activity for incongruent and congruent conditions, respectively, averaged over trials. (See Equation 6 in the Supplemental Method for the full expression.) The scaling factor is a critical part of this operation, in that it varies across vertices, incorporating information regarding not only the amount of trial-level variability in each vertex, but also how this variability co-varies with each of the other vertices within the region. In this way, these vertex-wise scaling factors enable LDA to squash unreliable spatial dimensions and expand reliable ones.

2.5.1 Implementation of LDA

Despite these similarities between models, additional steps were necessary to curb overfitting within the multivariate models. We fitted the multivariate models in leave-one-session out cross-validation, in which the weights for a given "testing" session were estimated using concatenated data from the other two "training" sessions. To minimize bias in decoding results caused by class imbalance and other confounding factors, we used a stratified random undersampling algorithm to generate the training set for LDA (see Supplemental Method Section 10.2.2). We implemented LDA using the *rda()* function from the R package *klaR*, with parameters fixed at *gamma = 0.25, lambda = 1*, providing a moderate amount of regularization (Weihs et al., 2005). The obtained weights were then multiplied by the trial-level data from the held-out test session to generate the trial-specific scores for test–retest reliability analysis.

2.6 Reliability models

After the fMRI activity patterns within each parcel were summarized into univariate and multivariate scores, or projections, per trial, the test-retest reliability in these projections was estimated. We compared two different methods of estimating test–retest reliability: the Intraclass Correlation, which is a standard "summary statistic" method, and correlations derived from hierarchical Bayesian models. We refer to both of these methods generally as "reliability models".

We fitted both summary statistic and hierarchical Bayesian reliability models similarly and independently for each of the 400 Schaefer-atlas brain regions, and each type of spatial model (univariate, multivariate). This enables us to compare the independent and interactive effects of spatial and reliability modeling methods, similar to decomposing an effect into main effects and interactions.

We scope our test–retest reliability analyses, however, exclusively on projections from the baseline session. This decision was made to streamline reporting of our results. There is strong reason to expect that individual differences in the Stroop effect were maximized within this session (see Section 2.2; Braver et al., 2021). Consequently, this decision likely works *against* our key hypothesis that multivariate models improve properties of test–retest reliability relative to univariate models. This is because the test–retest reliability of multivariate projections is not only dependent on baseline-session data, but also data from the other two sessions (proactive and reactive), as those sessions constituted the training set (in cross-session cross-validation). By contrast, univariate projections within the baseline session do not have this dependency. In other words, focusing on baseline-session data not only allows us to simplify our results, but it gives univariate models their "best-shot" in comparison against multivariate models. Thus, any improvements we see in properties of test–retest reliability due to multivariate models provides strong evidence for their superiority.

2.6.1 Summary-statistic method

The Intraclass Correlation Coefficient (ICC) is a standard way to measure reliability. In particular, the most common form quantifies the consistency instead of absolute agreement between two repeated measures on a same group of participants (ICC(3, 1) in Shrout and Fleiss, 1979).

Averaging the trial-level activation estimates over trials, we formed one summary score per subject, repetition, and condition (incongruent, congruent). The contrast of interest is the Stroop effect, that is, the difference between means of incongruent versus congruent trials. In this setting, ICC(3, 1) of the Stroop effect is defined as the Pearson correlation coefficient in participants' Stroop effects across repetitions (Supplemental Method 10.3.1).

2.6.2 Hierarchical Bayesian Analysis

In the hierarchical Bayesian method of estimating reliability, projections are modeled at the trial level, and the amount of variance at different levels — such as across trials versus across subjects — is separately captured. Unlike intra-class correlation, a hierarchical Bayesian decomposition enables individual-difference correlations to be estimated in a manner that accounts for the impacts of variance at each level.

The structure of our hierarchical models generally followed prior work (Chen et al., 2021). Although we provide a simplified description here, please refer to the Supplemental Method for details. To provide robustness to outliers, we assumed that the output of the spatial models (score in Equation 1) were generated from student-*t* distributions. For readers familiar with Wilkinson and Rogers (1973) notation (e.g., *lme4* syntax), a key part of our model can roughly be expressed as

$$\text{score} \sim \text{congruency} * \text{repetition} + (\text{congruency} * \text{repetition} \mid \text{participant})$$

in which the non-parenthetical terms on the right-hand side denote the population ("fixed") effects, while parenthetical terms denote individual-level ("random") effects. The pivotal quantity of test–retest reliability is contained within the parameters associated with the individual-level effects. The above expression, however, only reflects a small portion of the hierarchical models we fitted. For full descriptions, please refer to Supplemental Method 10.3.2 and 10.4.

The hierarchical models were implemented using the R package *brms* with noninformative or weakly-informative hyperpriors automatically selected by *brms* (Bürkner, 2017). Parameters were estimated through Markov Chain Monte Carlo (MCMC) with four chains with 2000 iterations each (including 1000 warm-up iterations). The training time for each model (i.e., for univariate or multivariate contrasts of a single parcel in the baseline session) was approximately an hour, running on four threads on an Intel Xeon E5-2670 CPU (2.60 GHz).

To ensure that we used models well-fitted to our dataset, we assessed four models of varying complexity and compared them in estimates of their ability to predict new data points. These models varied in terms of the complexity of the covariance structures in the location and scale parameters of the *t* distributions (Supplemental Methods 10.4). Due to computational constraints, this comparison was conducted in a subset of 32 parcels within the 400-parcel Schaefer atlas. These parcel locations were based on prior work that demonstrated a core set of regions associated with demands for cognitive control (Assem et al., 2020, within the Glasser atlas), and were independently selected in prior work (see Braver et al., 2021 Supplemental Materials https://osf.io/pa9hj for identifying corresponding regions within the Schaefer atlas). Expected log-predictive density with Bayesian leave-one-out cross-validation (ELPD_LOO) was computed by *brms* through the *loo* package (Vehtari et al., 2024). A model of intermediate complexity was supported by this criteria in all 32 parcels for both univariate and multivariate approaches (Reduced Model 2; Supplemental Method 10.4.2, Supplemental Results). Another model fit statistic, widely applicable information criterion (WAIC), also supported this model (not reported). Therefore, we then separately fitted this model to data from all 400 parcels, selecting it as the basis of analyses reported in the Results.

3 Results

3.1 Pronounced univariate activation in fronto-parietal networks at the population level

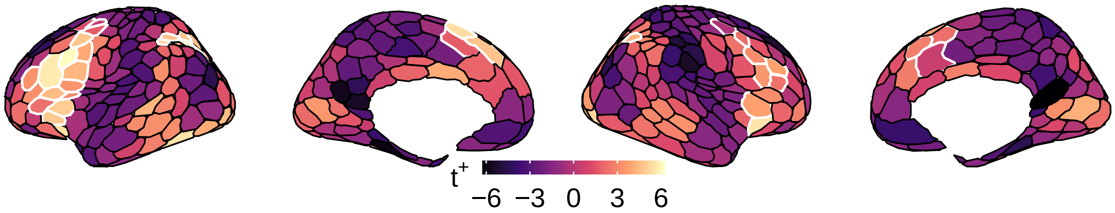

Figure 2: **Population-level univariate activation to Stroop task demands.** Per brain parcel, t^+ statistics show the sign and magnitude of univariate activation. These statistics were computed by estimating the spatially averaged difference of BOLD activity estimates on incongruent versus congruent trials, then dividing this estimate by its standard error, estimated through a hierarchical Bayesian model. Regions of interest, which are outlined in white borders, were defined using a larger portion of the Dual Mechanisms of Cognitive Control dataset.

To provide a basis for assessing reliability of cognitive control-related BOLD responses at the level of individuals, we first characterized BOLD responses that were co-localized across individuals within particular brain regions. Namely, we assessed which cortical regions (parcels within the Schaefer 400 17-Network atlas) increase BOLD activity on more demanding incongruent Stroop trials, relative to less demanding congruent trials, generally across subjects, as such changes in activity are expected to be exhibited by a region involved in cognitive control. This contrast was implemented by first spatially averaging the BOLD signals across vertices within each parcel separately on each trial, then estimating the population-level change in activation within a hierarchical Bayesian model (Reduced model 2 in Supplemental Method 10.4.2). Hereafter, we refer to this contrast as a “univariate activation” contrast, as it relies on a spatially univariate (uniform) averaging of signals from a given region. Figure 2 displays the results of this analysis.

To illustrate the continuous magnitude and sign of the change in activation across parcels, we depict the effect size with a statistic we refer to as t^+, which incorporates both the mean level and amount of uncertainty of activation change (note, however, that although similar in form, the definition of t^+ does not straightforwardly translate to the frequentist’s t-statistic). Using this contrast, prominent increases in activation were observed in frontal and parietal cortex, particularly within the left hemisphere (Figure 2), in a pattern that is strongly consistent with extensive prior findings (e.g., Assem et al., 2020; MacDonald et al., 2000).

To complement this continuous measure, we also employed a dichotomous measure, by using “regions of interest” (ROIs) based on the larger Dual Mechanisms of Cognitive Control dataset from which this subset of data was taken. Specifically, in Braver et al. (2021) a set of 35 parcels were identified from a larger sample of participants (N = 80) that showed consistent activation according to cognitive control demands, and which included the Stroop effect contrast, but also parallel contrasts across three additional tasks studied within that report. Compared to using the present sample and task alone, defining ROIs on the basis of the larger Dual Mechanisms dataset likely yields more accurate definitions of core brain regions associated with cognitive control across multiple tasks. These ROIs included brain areas classically associated with cognitive control, such as mid-lateral prefrontal and posterior parietal cortices (Figure 2, white borders), and which most prominently belonged to the fronto-parietal control network.

In subsequent analyses, we will use both of these complementary statistics (t^+, ROI definitions) to illustrate relations between univariate activation at the population level, and properties of other neural measures at the individual level, such as test–retest reliability.

3.2 Hierarchical modeling reveals highly reliable estimates at the individual level

Having established that univariate contrasts reveal highly robust activity changes on average *across* individuals, we next asked whether these same univariate contrasts are also highly consistent *within* each individual, across repeated testing sessions, which in our sample were separated by several months to years of intervening time. In other words, what is the test–retest reliability of fMRI activation to Stroop-task demands?

To provide a comprehensive assessment of test–retest reliability, we used two different approaches for estimating reliability and compared their results. The first approach was via a widely used method, the intra-class correlation, which relies on a two-stage procedure in which "summary statistics" are computed, then correlations in these statistics are subsequently estimated. The second was via a hierarchical Bayesian modeling approach. Using the same models we fitted to estimate the population-level effects separately in each brain parcel (i.e., in Figure 2), we also estimated the posterior distribution of test–retest correlations within individuals. Following prior work Chen et al. (2021), to summarize these posterior correlations with a single value, we used the maximum *a posteriori* probability (MAP) estimate (i.e., the single correlation value at which the density of the posterior probability is maximal).

Figure 3, top, depicts the results of both of these approaches, as applied to individuals' univariate activation contrasts. When estimated through a summary statistic approach, test–retest reliability is quite modest, reaching maximum values of only $r \sim 0.5$, primarily within our ROIs (white borders), which tended to contain strong population-level effects. The results from the hierarchical Bayesian model were quite different. As revealed by these models, test–retest reliability was often close to maximum, again prominently within ROIs, but also within a much broader set of regions. In fact, the number of regions with conventionally "high" reliability ($r > 0.7$) dramatically increased from 3 to 186, indicating that nearly 50% of the brain parcels were highly reliable.

By directly contrasting the summary-statistic and hierarchical correlations within each parcel, the differences between the estimates can be explicitly illustrated. Two findings here are notable. First, the change in reliability was most prominent for regions in which the intra-class correlations was already distant from zero. For regions with positive intra-class correlations, hierarchical modeling dramatically improved reliability, while for regions with negative intra-class correlations, hierarchical modeling reduced reliability. For regions with near-zero intra-class correlations, little change was revealed. This finding is consistent with prior theoretical and empirical demonstrations that, compared to hierarchical Bayesian models, intra-class correlation multiplicatively underestimates test–retest reliability (Chen et al., 2021). Second, improvements in reliability revealed by hierarchical modeling also tended to be particularly large in regions with strong

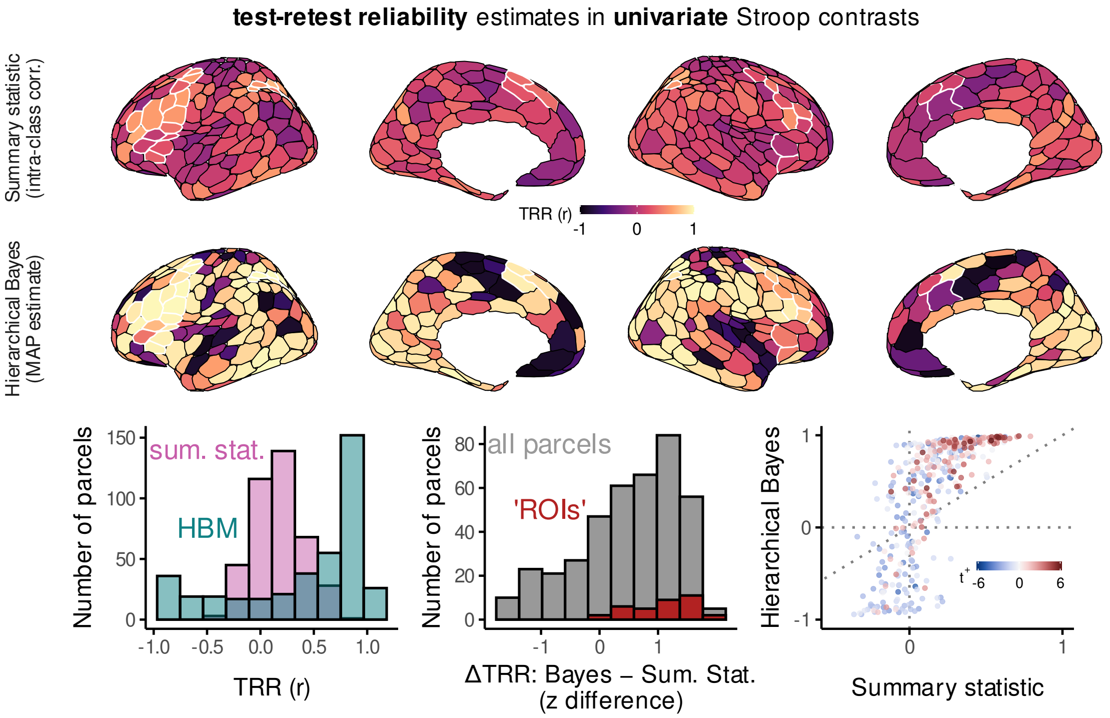

Figure 3: **Test–retest reliability estimates in univariate Stroop contrasts. Surface plots at top**, Test–retest reliability (TRR) correlation coefficients (r) estimated through “summary statistic” (upper row) or hierarchical Bayesian models (bottom row). To summarize the posterior correlation distributions in the hierarchical models, we used the maximum *a posteriori* (MAP) estimate. White borders illustrate regions of interest (ROIs) identified in a prior report (Braver et al., 2021). **Bottom left**, Histogram of the distribution of these test–retest correlations across all cortical parcels (hierarchical Bayesian models or HBM, pink; summary statistic, sum. stat.). **Bottom middle**, Histogram of the difference in test–retest correlations between hierarchical Bayes MAP estimates minus summary-statistic estimates, over all cortical parcels and also in ROIs. Prior to subtraction, correlation coefficients were z-transformed (i.e., inverse hyperbolic tangent). **Bottom right**, Scatterplot of the relation between test–retest correlation estimates from summary statistic models (x axis) and hierarchical Bayesian models (y axis), over all cortical parcels (points). The color scale illustrates the sign and magnitude of the population-level univariate Stroop effect contrast (t^{+}, defined in Figure 2). Dotted lines illustrate x and y intercepts, as well as the unity line.

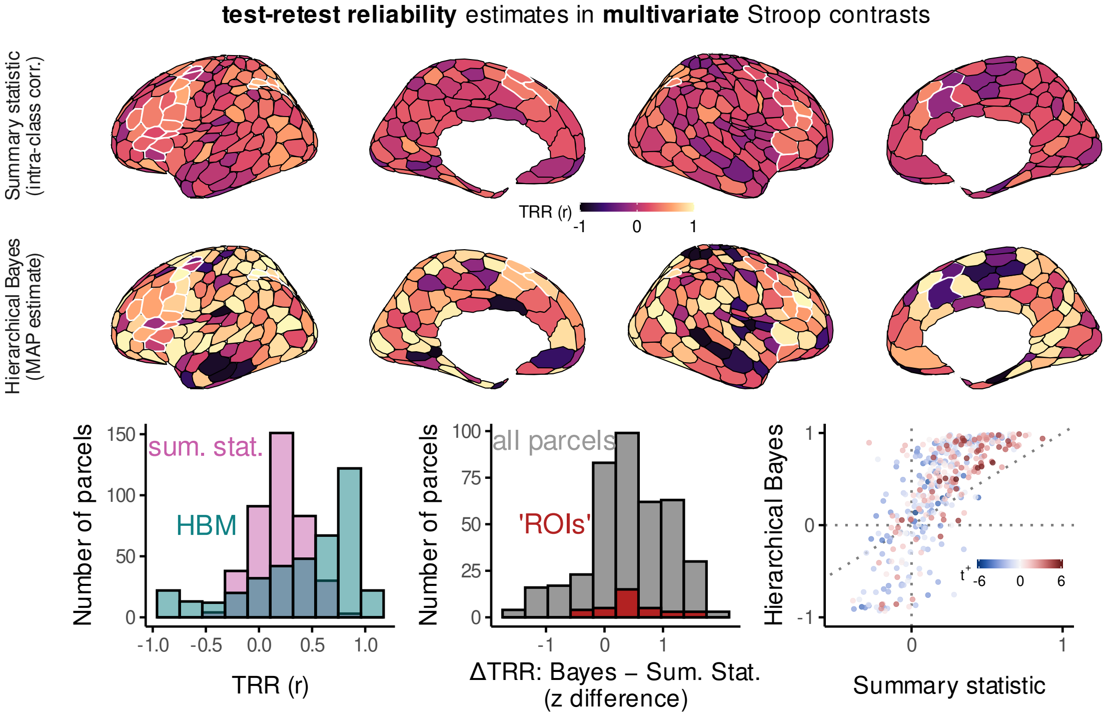

Figure 4: **Test–retest reliability estimates in multivariate Stroop contrasts.** All plots are analogous to those in Figure 3.

population-level effects, that is, fronto-parietal regions which were consistently activated to Stroop-task demands across subjects. Consequently, the same experimental contrasts used to identify regions that are involved in cognitive control, generally within the population, can also be used to individuate people, in terms of the extent to which their neural responses within these regions were modulated by the task (e.g., Braver et al., 2010).

We then sought to characterize reliability of an alternative contrast, of multivariate activation. Similar to the univariate activation contrast, the multivariate contrast we used implemented a linear transform of spatial patterns of brain activity, in which activity evoked by easier congruent trials was subtracted from activity evoked by more difficult incongruent trials. But instead of a uniform weighting of signals within a region of interest, the multivariate contrast relied on an optimal estimation of the weighting vector (see Method). The reliability properties of such a contrast have not yet been established. In general, findings were quite similar to univariate contrasts: reliability was dramatically improved by hierarchical modeling, and the most reliable estimates were revealed in many of the same regions (Figure 4). In several regions, however, the gains revealed by hierarchical modeling were not quite as extreme as observed univariate contrasts. We will return to this observation in the Discussion.

Collectively, these findings demonstrate that, both univariate and multivariate fMRI contrasts can be highly consistent within individual, at least when basing the estimation on the most likely (MAP) estimates from hierarchical models.

3.3 Imprecision limits interpretation of individuals' univariate activation estimates

The results summarized in Figure 3 reflect the single most likely test–retest correlation value per brain region, as provided by the MAP estimator. While a convenient summary, this estimator conveys no information about *how much more likely* the MAP estimates are, relative to other outcomes in which reliability is lower. For instance, if the model assigns almost as much probability to a pattern of results in which reliability is considerably lower than in Figure 3, the prior results should not be well trusted. Evaluating these possibilities requires assessing the *precision* or *uncertainty* in the test–retest reliability estimates. Such an assessment is naturally supported within a Bayesian modeling framework, through analysis of the tails of the posterior distribution of test–retest correlations.

We found that, even as the single most likely test–retest reliability estimates of many regions approached a maximal value of one, in these same regions, the hierarchical model also assigned relatively high probabilities to correlation values that were quite low. In many cases, correlation values below zero were assigned a non-negligible probability. We depict this uncertainty in Figure 5 (upper), in which the mean posterior test–retest correlation estimates (dark to hot hues) are thresholded based on their (im)precision (fully saturated to unsaturated grey). That is, the only regions displayed in fully saturated colors are those with 95% of their posterior correlation density or more located above zero. This graphic therefore combines information about the central tendency (dark to hot hues) and lower tail (saturation) of the posterior (Taylor et al., 2023). From this view, only a handful of parcels now have a combination of both high reliability (> 0.7), and high certainty in the reliability estimate, a finding which converges with prior work (Chen et al., 2021).

In analyses focused on individual difference questions, observing such a high degree of uncertainty in individual-level variables would severely complicate inferences, weakening conclusions drawn about any relations observed.

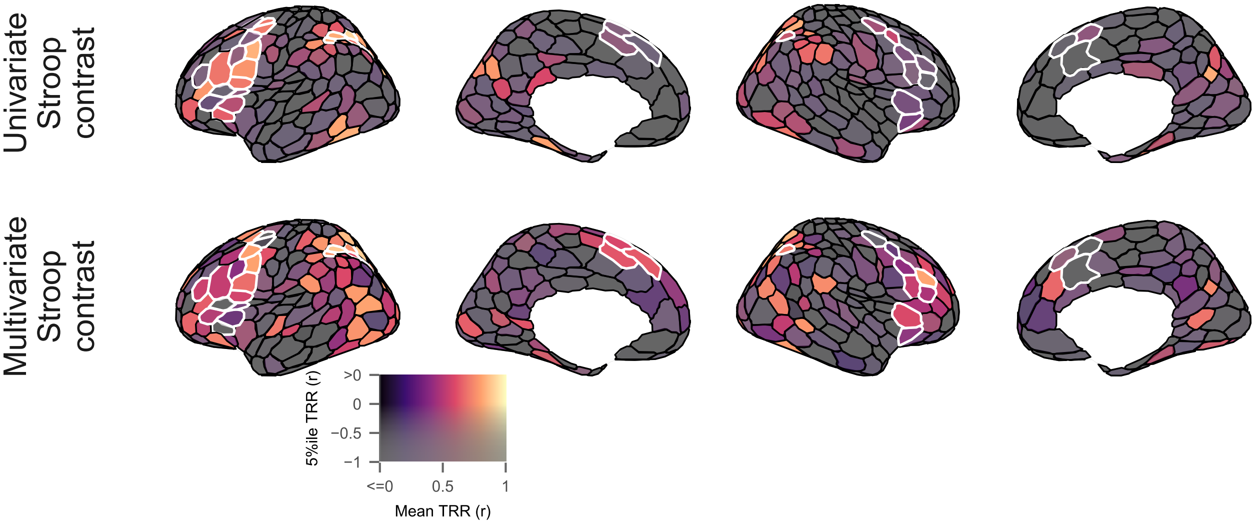

Figure 5: **Test–retest reliability correlations thresholded by certainty.** All correlations were estimated within hierarchical Bayesian models, as the *mean* of the posterior distribution over test–retest correlation values. The black-to-hot dimension in the colormap indicates the mean posterior test–retest correlation for univariate (top) and multivariate (bottom) contrasts. In contrast to Figure 3, we use the mean to summarize the posterior here, as the mean is more sensitive to the full distribution of the posterior, and so more strongly reflects behavior of the tails, than does the MAP estimate. The saturated-to-unsaturated dimension in the color-map reflects the lower bound of the test–retest correlation estimate (95th percentile). A soft threshold is applied to the color-map, so that only parcels that were estimated to have positive reliability with high certainty are displayed with full saturation (more colorful or deeper black), while those with increasingly negative lower bounds are displayed with quadratically decreasing saturation (more grey; Taylor et al., 2023).

3.4 A combined hierarchical and multivariate approach yields highly reliable, and highly precise, measures of individuals

Are multivariate contrasts subject to similar amount of imprecision as univariate contrasts? Given that multivariate contrasts optimize different criteria than univariate contrasts (see Method), we suspected they may yield more precise individual-level measures. A hint to this question is already provided by Figure 5, bottom, in which the reliability of multivariate contrasts are thresholded by their lower-bound tail, in an identical manner to univariate contrasts (top). Compared to univariate contrasts, although the central tendencies of multivariate contrast reliability in several regions is somewhat numerically lower, over twice as many regions (80 versus 39) have a high certainty (> 95% probability) of positive reliability. Of note, this category also included over half (19/35) of the ROIs. For instance, several regions in lateral PFC (particularly right hemisphere), posterior parietal cortex, and superior frontal cortex are only identified as highly certain to have positive reliability when the multivariate contrast is used. In Table 1, we list the 40 parcels with the highest (most positive) lower-bound estimates of test–retest reliability. Most of these parcels (26/40) are located within two fronto-parietal networks, Dorsal Attention A and Control A.

To provide a more comprehensive comparison, however, we calculated the precision of test–retest reliability separately in univariate and multivariate contrasts, then contrasted their magnitudes within each brain parcel. We used precision, that is, the reciprocal of the standard deviation of the test–retest correlation posterior, so that higher values indicate more precise estimates. (To ensure a sensitive comparison across a large range of values, we logarithmically transformed SDs prior to taking the reciprocal.) As illustrated in Figure 6, multivariate contrasts generally led to more precise individual-level contrast estimates.

We can now consider the joint impact of hierarchical and multivariate approaches on reliability of individual-level variables. For both univariate and multivariate contrasts, hierarchical modeling revealed that the central tendency of the posterior test–retest reliability distribution are often much greater than expected based on less-accurate summary statistic estimates (shown for multivariate contrasts in Figure 7 left panel, y axis). Yet relative to univariate contrasts, multivariate contrasts more strongly pulled the tails of the posterior distributions closer to their central tendencies (Figure 7 left panel, x axis). This not only increased the certainty in the plausible range of the parameter estimates, but also made these model results easier to summarize and interpret with point estimates. Such changes were relatively widespread, in that over 50% of parcels (233/400) show both improved reliability compared to summary statistic modeling (Figure 7, and improved certainty in individual-level estimates relative to univariate (Figure 7, left panel, upper right quadrant). Thus, this combination of methods can lead to a complement of benefits that has eluded prior work in this area: both *highly reliable* individual-level estimates, as well as *highly certain identification* of individual-difference associations.

3.5 Multivariate contrasts achieve high individual-level precision by squashing trial-level variability

How do multivariate contrasts improve the certainty of individual-level estimates? We illustrate a means by which this improvement is achieved. Theoretical work has demonstrated that a pivotal quantity for individual difference analyses is the ratio of trial-level variability to individual-level variability (Chen et al., 2021; Figure 1), which we refer to here as the "variability ratio". When trial-level variability greatly outweighs individual-level variability (a high variability ratio), uncertainty in correlations among individual-level variables is maximized. Conversely, if trial-level variability is reduced, for example, by obtaining measurements less susceptible to such noise, then certainty in correlations will increase. Thus, relative to univariate contrasts,

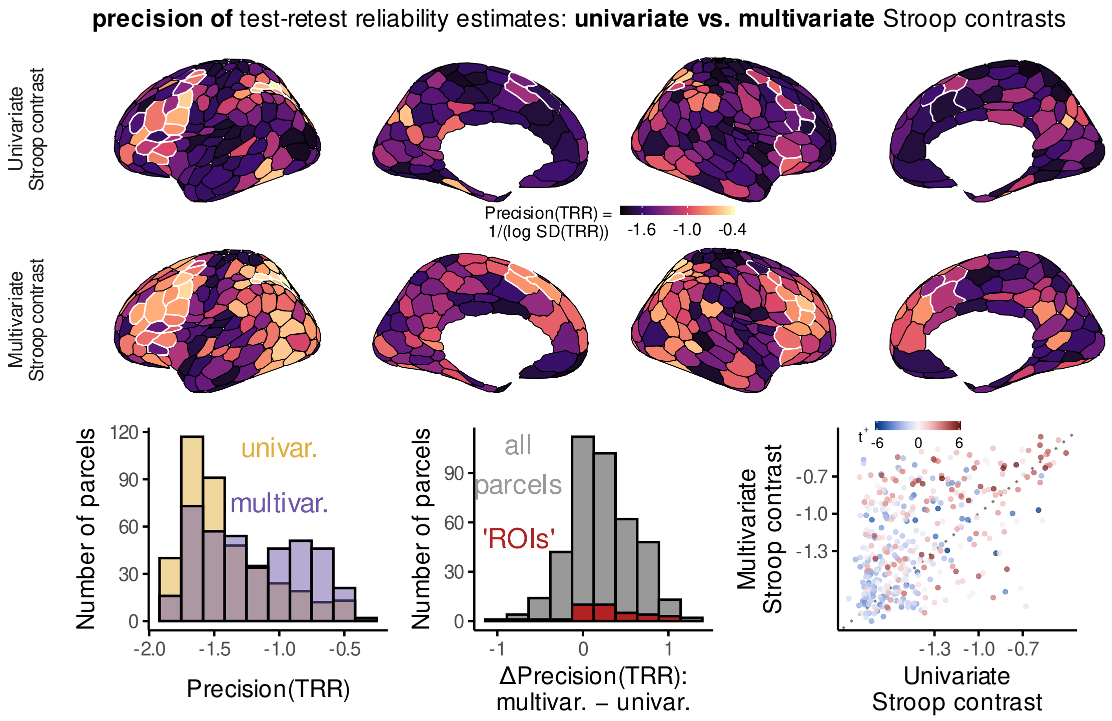

Figure 6: **Precision of test–retest reliability estimates in univariate versus multivariate contrasts.** We defined the precision of test–retest reliability correlations as the reciprocal of the log-transformed standard deviation across samples from the posterior distribution. Higher values indicate less variable (more certain) estimates of test–retest correlations. Aside from the difference in the statistic of interest, all plots are analogous to those in Figure 3.

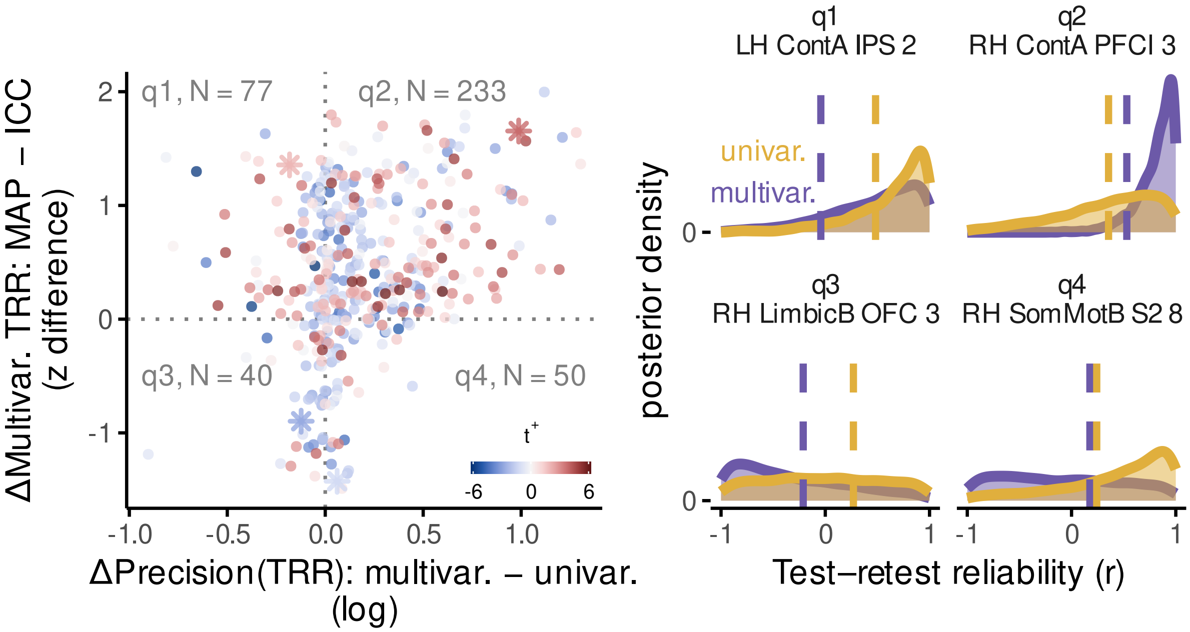

Figure 7: **Joint impact of multivariate contrasts and hierarchical Bayesian modeling on individual-level correlation estimates. Left**, Scatterplot of the joint benefits of multivariate and hierarchical modeling. The x axis depicts the increase in test–retest correlation precision associated with using a multivariate (multivar.) versus univariate (univar.) contrast (x axis). The y axis depicts results from multivariate contrasts only: namely, the increase in test–retest correlation strength revealed by hierarchical (maximum *a posteriori* estimate, or MAP) versus summary-statistic models (intra-class correlation coefficients, or ICC). "N=(#)" indicates the number of parcels in each quadrant. The color-map illustrates the population-level univariate activation estimate associated with each brain parcel (see Figure 3). Asterisks indicate locations of example parcels whose posterior densities are displayed in the right panels. **Right**, Example posterior densities of test–retest correlation values. Dotted vertical lines indicate values of summary statistic reliability estimates. These four regions were chosen as representative examples of the posterior densities within each of the four quadrants in the left panel of this figure (marked by asterisks). *LH ContA IPS 2* is located within rostral aspect of left intraparietal lobule, *RH ContA PFCl 3* within rostral right mid-lateral PFC, *RH LimbicB OFC 3* within right orbitofrontal cortex, and *RH SomMotB S2 8* near right somatomotor and supramarginal gyri.

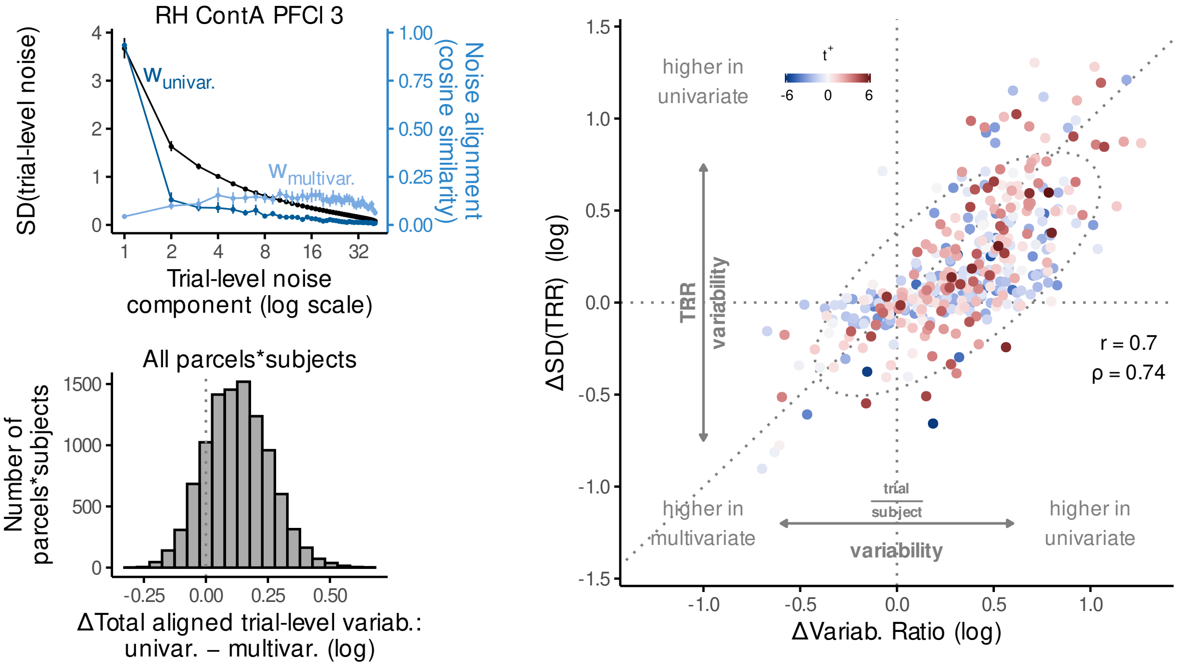

Figure 8: See following page for caption.

Figure 8: **Increased precision through reduced susceptibility to trial-level variability. Left, top**, Analysis of principal components of trial-level variability and their relation to univariate and multivariate contrasts in an example region. Principal component analysis was used to identify the spectrum of components of trial-level ”noise”, that is, the set of dimensions along which this PFC region fluctuates over trials, within each task condition. Each dimension (x axis) corresponds to a particular spatial pattern over vertices within the region, while the black points (left y axis) display how much the region fluctuated along this dimension over trials (in units of standard deviation, or square rooted eigenvalues), on average over subjects. Most variability is concentrated in the first dimension, indicating that trial-level variability was overwhelmingly low dimensional. The blue dots and lines (right y axis) illustrate how these dimensions of trial-level fluctuations are aligned to the univariate (dark blue, $\mathbf{w}_{\text{univar.}}$) and multivariate (light blue, $\mathbf{w}_{\text{multiv.}}$) weight vectors. The univariate weight vector is primarily aligned to the principal dimension of trial-level variability, whereas the multivariate weight vectors are least aligned to this dimension, and instead more heavily weight intermediate dimensions. For all points, error bars illustrate boostrapped 95% confidence intervals of between-subject variability in means. **Left, bottom**, In most brain parcels and subjects we measured, multivariate weight vectors were susceptible to less total trial-level variability than univariate weight vectors. This susceptibility was estimated separately in each parcel and subject by computing the log ratio of the total amount of trial-level SD in the direction of univariate versus multivariate weight vectors. Positive values indicate univariate weight vectors were aligned to a greater total amount of trial-level variability than multivariate weight vectors. **Right**, Improvements in *trial/subject* variability ratio translated into improvements in certainty of individual-level correlations. The x axis represents log(univariate variability ratio) – log(multivariate variability ratio), while the y axis represents log(univariate SD(TRR)) – log(multivariate SD(TRR)). In other words, positive values indicate that, relative to univariate contrasts, multivariate contrasts either led to a *smaller and thus more favorable* trial/subject variability ratio (x axis), or to a *less uncertain and thus more interpretable* test–retest correlation (y axis). Note that the values and interpretation of the y axis here are equivalent to that of the x axis in Figure 7. For clarity in this figure, however, we relabel this figure’s y axis to match the interpretation of its x axis, wherein higher values of the underlying statistic (SD, trial/subject variability ratio) are less favorable for individual differences analyses.

multivariate contrasts may improve certainty by finding a dimension that is less susceptible to trial-level noise than the uniform dimension, to which univariate contrasts are bound (Figure 1).

This hypothesis makes two assumptions. In particular, the uniform dimension should indeed be highly susceptible to trial-level variability. Additionally, the dimension used by the multivariate contrast should be less susceptible to trial-level variability. Are these borne out in our data? Consider an example region in left dorsal prefrontal cortex, which consists of 58 vertices. In this region, BOLD activity can fluctuate across trials in 58 different ways (i.e., along 58 different dimensions). Identifying and ranking these dimensions by the amount of across-trial variability that occur along them, we can see that a majority of the trial-level variability is concentrated in only one or two dimensions (Figure 8, left top, black points and lines). This identification and ranking also allows us to assess how the spectrum of trial-level variability relates to the univariate and multivariate contrasts. We assessed how the univariate and multivariate weight vectors, which are spatial patterns that define how information is pooled across vertices within a region, are aligned to each of these ranked dimensions of trial-level variability. Indeed, the univariate weight vector is strongly aligned to the largest dimension of trial-level noise (Figure 8, left top, dark blue). The multivariate weight vector, however, is least aligned to this high-variability dimension, and instead weights other intermediate dimensions more strongly (light blue). These patterns can be summarized within each parcel by computing the total amount of trial-level variability in the direction of each contrast's weight vector (i.e., by weighting the SD of each dimension by its alignment to univariate or multivariate contrasts, then summing across dimensions). Performing this summarization in every parcel of every subject, we can see that in nearly all of them, univariate contrasts are susceptible to a greater amount of trial-level variability than multivariate contrasts. (Figure 8, left bottom). These findings indicate that the trial-level variability is of low spatial frequency within our brain parcels (i.e., predominantly uniform across vertices), and that only multivariate contrasts can exploit this uniformity to find alternative, higher spatial frequency dimensions that are more robust to trial-level variability.

Having demonstrated the difference in susceptibility to trial-level variability, we then sought to link changes in the trial/subject variability ratio to changes in the certainty of test–retest correlations. Using hierarchical models fitted in each brain parcel, we first computed the trial versus subject variability ratio separately for univariate and multivariate contrasts. Next, to show how multivariate contrasts impact this ratio, we subtracted these ratios between contrast types (after a log transform) to form a change statistic, which is positive when univariate contrasts yielded a higher trial/subject variability ratio, and negative when multivariate contrasts yielded a higher ratio (Figure 8, bottom, x-axis). Following similar logic, we computed a change statistic for the standard deviation of test–retest correlation, such that positive values indicate greater *imprecision* with univariate contrasts, while lower values indicate greater *imprecision* with multivariate contrasts (Figure 8, bottom, y-axis). Most brain regions fall in the all-positive quadrant, indicating that multivariate contrasts made both ratios more favorable for individual difference-focused analyses. Importantly, though, the change in variability ratio was also strongly positively correlated with the change in precision. This relation is consistent with improvements of correlation precision being driven by suppression of trial-level variability in the measures.

		Univariate TRR			Multivariate TRR		
Parcel (Schaefer 400-17)	t^+	MAP	5%ile	ICC	MAP	5%ile	ICC
LH DorsAttnA SPL 3	4.58	0.97	0.46	0.61	0.94	0.75	0.87
RH DorsAttnA SPL 4	2.23	0.99	0.67	0.79	0.97	0.74	0.75
LH ContA IPS 5*	4.65	0.98	0.46	0.54	0.99	0.67	0.71
LH VisCent ExStr 11	1.82	0.68	-0.16	0.35	0.95	0.63	0.67
LH DorsAttnA SPL 6	0.60	0.95	-0.51	0.36	0.98	0.62	0.78
RH DorsAttnA SPL 7	0.43	0.95	0.23	0.47	0.96	0.59	0.66
LH DorsAttnA SPL 7	1.20	0.68	-0.66	0.14	0.99	0.59	0.67
LH DorsAttnB PrCv 1*	4.02	0.98	0.55	0.59	0.87	0.55	0.65
LH DorsAttnA SPL 1	2.87	0.97	0.33	0.37	0.86	0.52	0.70
LH DorsAttnA SPL 2	2.02	0.78	-0.41	0.34	0.87	0.51	0.71
RH DorsAttnA SPL 5	-0.31	0.77	-0.18	0.48	0.99	0.51	0.65
RH DorsAttnA SPL 2	2.91	0.97	0.52	0.58	0.88	0.50	0.67
LH DorsAttnA SPL 4*	4.76	0.99	0.76	0.71	0.85	0.49	0.58
RH ContA IPS 1	2.81	0.91	-0.26	0.42	0.87	0.49	0.64
LH DorsAttnA ParOcc 1	-2.69	-0.86	-0.93	-0.24	0.98	0.48	0.58
LH DefaultB IPL 1	-2.22	0.49	-0.74	-0.11	0.93	0.47	0.65
LH DefaultB PFCl 2*	4.76	0.88	-0.55	0.29	0.98	0.47	0.58
LH ContA IPS 4*	4.94	0.98	0.49	0.62	0.91	0.46	0.64
RH ContA PFCl 3*	3.80	0.69	-0.57	0.36	0.98	0.46	0.53
LH ContA Temp 1	-0.69	0.94	-0.41	0.21	0.87	0.46	0.60
LH DorsAttnA TempOcc 4	4.06	0.92	-0.76	0.26	0.82	0.45	0.62
LH DorsAttnA ParOcc 2	1.12	0.99	0.61	0.59	0.96	0.44	0.48
RH DefaultB Temp 2	3.04	0.07	-0.76	0.07	0.93	0.43	0.50
RH DorsAttnA TempOcc 2	2.11	0.98	0.08	0.49	0.98	0.42	0.48
LH VisCent ExStr 10	-1.61	0.74	-0.88	0.15	0.98	0.41	0.26
LH DorsAttnB PostC 5	-0.03	0.95	-0.08	0.51	0.95	0.41	0.47
LH SalVentAttnA ParOper 1	-3.18	0.85	-0.67	0.16	0.95	0.40	0.60
RH ContA IPS 4	2.67	0.94	-0.17	0.62	0.84	0.40	0.50
LH DorsAttnB PostC 3	0.75	0.97	0.28	0.57	0.84	0.39	0.61
LH ContA PFCd 1	-2.26	0.88	-0.74	0.08	0.93	0.34	0.53
RH SalVentAttnA ParOper 1	-3.35	0.56	-0.78	0.03	0.94	0.34	0.55
RH DefaultC IPL 1	-3.38	0.83	-0.31	0.33	0.96	0.33	0.45
LH DorsAttnB FEF 2*	1.22	0.92	-0.71	0.32	0.93	0.33	0.46
RH ContB IPL 4*	4.34	0.91	-0.18	0.55	0.86	0.31	0.46
RH DorsAttnA SPL 8	-2.83	-0.63	-0.89	-0.07	0.97	0.29	0.50
LH DorsAttnA SPL 5	1.77	0.95	0.06	0.37	0.95	0.28	0.48
RH ContC pCun 1	-2.64	0.98	0.48	0.63	0.98	0.27	0.51
LH ContA IPS 3*	5.95	0.93	0.18	0.40	0.68	0.26	0.54
LH ContA PFCl 1*	4.33	0.98	0.37	0.56	0.80	0.25	0.43
LH ContA IPS 1*	3.41	0.99	0.57	0.59	0.69	0.25	0.53

Table 1: **Regions of interest and associated test–retest reliability statistics.** Displayed are key statistics from Schaefer-atlas parcels that exhibited the strongest test–retest reliability in the multivariate Stroop activation contrast. Asterisks by parcel names indicate regions of interest associated with cognitive control demands, defined in a prior report (Braver et al., 2021; 11 out of 35 also exhibit strong test–retest reliability). The strength of population-level univariate activation in each region summarized in the statistic t^+ (descending order). Test–retest reliability in univariate activation contrasts (Univariate TRR) and in multivariate contrasts (Multivariate TRR) are also displayed. The maximum *a posteriori* estimate (MAP) indicates the central tendency of posterior TRR correlations (i.e., the most likely point estimate), while the 5th percentile of the posterior (5%ile) indicates a lower-bound on the estimate (i.e., the uncertainty in the estimate). Rows are sorted by this statistic. Intra-class correlation coefficient estimates (ICC) are also displayed.

4 Discussion

We investigated whether a current, widely-used psychological test of cognitive control can yield *reliable* fMRI measures of individual differences over repeated testing sessions. In particular, we focused on characterizing the way in which two contemporary modeling frameworks, Hierarchical Bayesian modeling and MVPA, jointly impact the derived fMRI measures and their associated estimates of test-retest reliability. We found that their combination clearly afforded complementary benefits for estimating individual differences. Hierarchical Bayesian modeling generally led to higher estimates of test-retest reliability than the more traditional "summary-statistic" framework, but in most cases, these estimates were highly uncertain, reflecting strong trial-to-trial variability in the derived measure. This variability, however, was squashed by the application of MVPA, which in turn substantially increased the certainty of the Bayesian estimate of reliability. Therefore, by combining these contemporary modeling frameworks, widely-used task fMRI designs can be used to produce not only highly reliable, but also highly precise measures of individual differences.

Our findings caution against sweeping claims of the inadequacy of using task-based fMRI to study individual differences. For example, in one recent study, authors analyzed several extant datasets of common task-based fMRI designs, and reported that the test-retest reliability of univariate activation contrasts in key regions of interest were consistently low (Elliott et al., 2020). These results were interpreted as evidence that "the task-fMRI literature generally has low reliability" and conclude that such task-fMRI measures are "not currently suitable for ... individual-differences research". Our findings undermine this conclusion, as we present evidence that one of the most commonly used task designs in fMRI, the color-word Stroop task, can indeed elicit highly reliable fMRI measures. This discrepancy between findings can be accounted for, perhaps entirely, by our use of hierarchical modeling methods to estimate test-retest reliability, as opposed to their use of summary-statistic methods. When trial-level noise is high — which is invariably the case for fMRI and many behavioral measures of cognition — summary-statistic estimates of reliability are most inaccurate (Figure 1; Chen et al., 2021). In such scenarios, when investigators fail to grapple with these complexities within their approach to data analysis, the risk of faulty and overly pessimistic inferences increases. Our findings join a growing body of work in illustrating the usefulness of hierarchical Bayesian analysis as a principled way of approaching this issue, as it enables complex multi-level variance structures to be estimated and decomposed (Haines et al., 2020; Chen et al., 2021; Rouder et al., 2023; Rouder and Haaf, 2019; Snijder et al., 2023). Nevertheless, to our knowledge, only one other study has extended this approach to fMRI data (Chen et al., 2021), thus the benefits of this approach for fMRI data analysis are only beginning to be explored.

In addition, our findings mark an advance from prior modeling work on individual differences in that we provide clear evidence for a partial remedy to the "reliability crisis". Prior work has used hierarchical Bayesian frameworks to pinpoint the primary limiting factor for individual difference studies: the overwhelming influence of trial-level noise (Chen et al., 2021; Rouder et al., 2023). While this insight is invaluable, these studies also demonstrate that hierarchical Bayesian modeling, on its own, does not provide a solution to the psychometric issues gripping this field. In contrast, our findings demonstrate a partial solution to this issue, at least in the case of fMRI, is likely provided by MVPA (Kragel et al., 2021). We found that application of MVPA substantially suppressed trial-level noise relative to individual differences, and this led to more certain and interpretable estimates — without compromising the high degree of test-retest reliability. Thus, the use of multivariate rather than univariate contrasts can substantially boost power for detecting individual differences in task-fMRI responses to psychological manipulations.

A viable research strategy, then, may be to use task-fMRI and psychological manipulations to collect modest sample sizes of relatively densely-sampled individuals, and use MVPA in conjunction with hierarchical modeling to study their individual differences. Notably, such a goal is attainable through the efforts of a single laboratory. At first impression, these conclusions may seem to contrast with those of a recent study concluding that thousands to millions of individuals are necessary to detect individual difference relations between fMRI and behavioral measures (Marek et al., 2022). Yet, our conclusions do not actually conflict with theirs. First, this conclusion was based primarily on analyses of resting-state and structural MRI, as opposed to task-based fMRI, which was our focus here. Second, while a small number of task-based fMRI analyses were reported in Marek et al. (2022), the results of those analyses exclusively pertain to what the authors referred to as "brain-wide association studies", a niche category of individual difference questions, in which all of the covariance between an individual's fMRI activation contrast from a given task and their behavioral performance on the same task is considered to be noise, and therefore completely statistically discarded (see Extended Data Figure 3 of Marek et al., 2022; Spisak et al., 2023). Such a narrow and unusual definition of individual-difference correlations conflicts with a foundational assumption in psychology: that measures in different tasks are partially generated from a smaller set of latent psychological dimensions, such that dependence on common dimensions drives similarity between measures (Spearman, 1904; Bollen, 2002; DeYoung et al., 2022). As a result, the results from that study have limited to no bearing on the interpretation of results reported here. In fact, as their results show, when this common variance is not statistically discarded, individual-level correlations between task-fMRI and behavioral measures reach moderate effect sizes (*r*s between 0.3 and 0.6) — even despite the use of summary-statistic correlations, which are known to be downwardly biased (Chen et al., 2021), and univariate contrasts, which we have shown here to yield suboptimal precision for individual-level correlations.

Our findings also demonstrate how it is the case that MVPA leads to substantially lower trial-level variability than univariate contrasts. Because we used a highly interpretable multivariate framework, linear discriminant analysis, we were able to straightforwardly formulate univariate contrasts as a special case of a multivariate decoder, in which the decoder weights, which define how to summarize the spatial activity pattern, are uniform across vertices (Equation 2; Supplemental Method 10). Critically, the only feature that differed between our multivariate and univariate models was the nature of these weights: both models were applied to the same input data, and both implemented a linear scalar projection. The fact that trial-level variability was squashed in our multivariate contrast therefore implies that the MVPA decoder was able to find a dimension (i.e., weights) along which the signal-to-noise ratio was more favorable than the uniform dimension, to which univariate contrasts are bound (see Figure 1 D for a cartoon depiction). Analyses of group-level trends bolster this interpretation, in which we find that the principal component of trial-level noise in many brain regions is strongly aligned to the uniform dimension, whereas dimensions that encode the task conditions are considerably less aligned with the uniform dimension (Figure 8). Whether these improvements result from suppressing noisy vertices (Walther et al., 2016), exploiting noise correlations (Walther et al., 2016; Bejjanki et al., 2017), or capturing signal heterogeneity (Davis et al., 2014; Roth et al., 2018; Harrison and Tong, 2009), and the extent to which these features depend on particulars of preprocessing (e.g., smoothing, confound regression, HRF modeling) remain questions for future research. Such questions would be highly tractable to address under the framework we have presented here.

In a relevant prior study, univariate and multivariate decoding models were compared through simulations from a generative hierarchical model (Davis et al., 2014). Interestingly, these simulations included both trial and individual-level variability, and one conclusion reached was that MVPA may be *less* sensitive to *individual-level* variability than univariate models. This conclusion may initially seem to conflict with ours, in that we found MVPA models can yield more favorable measures for targeting individual differences.

Note, however, that this prior study considered a highly restricted case, in which individuals only varied in their univariate (uniform) activation to experimental conditions. This assumes that other factors that can strongly influence multivariate decode-ability — for example, the amount of voxel-level variability induced by experimental conditions, or the structure of their trial-level variability (e.g., noise correlations) — are not variable across subjects. In real data, however, it seems unlikely that these factors do not differ across subjects. Indeed, even though we did not explicitly optimize the MVPA decoders to enhance individual-level variability (but rather, to enhance condition-level versus trial-level variability within each individual), the resulting output patterns were nevertheless able to strongly individuate participants within our sample.

While our expectations that hierarchical modeling and MVPA leads to benefits in test-retest reliability were borne out in general, there were a minority of brain parcels that did not follow such a pattern. In some cases, there were a group of parcels for which the hierarchically estimated reliability was lower than that estimated through summary statistics. Examining this unorthodox set more closely, we found that they all had near zero or even negative summary statistic reliability (Figure 4). This pattern is consistent with the fact that hierarchical reliability estimates are multiplicatively scaled relative to summary-statistic estimates (Chen et al., 2021). In other cases, MVPA decreased the precision of test-retest reliability compared to univariate analysis (Figure 6). We suspect that this pattern reflects a source of noise to which MVPA is susceptible while univariate analysis is not: changes in the signal or noise topographies across cross-validation splits. Here, for our multivariate models, we cross-validated over data acquisition sessions, which were not only administered on different days, but also involved minor differences in experimental manipulations (see Method 2.6). These factors likely hampered the ability of our decoders to generalize across sessions. As a result, we view our results as providing a relatively pessimistic example of the benefits of multivariate modeling. Future experiments specifically tailored to support this type of analysis, we predict, will yield even more reliable and precise individual difference measures.

Future work would make valuable contributions by exploring several directions that build on these findings. We have demonstrated a framework that can yield highly reliable task-fMRI measures at the individual level. It remains to be seen, however, whether this improved reliability actually translates into improved predictive power for individual differences. Thus, the next clear step will be to use this framework to predict other cognitive or behavioral measures of interest. Finding that MVPA methods are more predictive than univariate methods would provide strong validation of this framework for studying *cognitively relevant* individual differences.

Another direction worthy of exploration is whether trial-level variability can be usefully decomposed in the service of studying individual differences. Here, we have considered such variability to be "noise". This consideration was reflected in our use of linear discriminant analysis, which explicitly suppresses trial-level variability, and by our focus on characterizing the reliability of subject-level means (or "locations"), as opposed to standard deviations (or "scales"), in fMRI measures of interest. These choices may have obscured cognitively relevant individual differences. For example, trial-level BOLD variability may reflect the operation of non-stationary control or attention processes that fluctuate over trials. As such, one may wish to adopt a decoding method that does not suppress such variability (see Kobak et al., 2016 for similar logic). In addition, the amount of trial-level variability may itself be useful as an individual difference variable, for example of the capacity for sustained attention (Williams et al., 2019; Saville et al., 2011). It is also tempting to consider that a fair amount of this trial-level variability may be explainable in terms of as-yet-unmeasured cognitive processes. Indeed, large cross-trial variability is an intriguing phenomenon that warrants further explanation.

Finally, although our results suggest some promise for the use of classical experimental designs to study individual differences, we do not wish to discourage the exploration of novel experiments or task designs. Most likely, there will be distinct advantages to "return to the drawing board" of task design. Classical designs have been optimized for group-level power, which perhaps occurred at the expense of eliciting subject-level variability (Hedge et al., 2018). As a result, individual variability may be more strongly elicited by more naturalistic, less constrained, or higher-dimensional designs (e.g., Shallice and Burgess, 1991; Rosenberg and Finn, 2022; Sonkusare et al., 2019; Nastase et al., 2020), or by bespoke tasks developed in conjunction with computational cognitive models (Zorowitz and Niv, 2023). But note, however, that neither of these cases will allow one to escape the need to develop sophisticated, theory-driven modeling and analysis approaches to make sense of the data. In addition to classical experimental designs, we predict that novel approaches would also be well served by the framework we have illustrated here, which exploits the complementary benefits of MVPA and hierarchical Bayesian modeling.

5 Data and Code Availability

The fMRI data used in the current study comes from the Dual Mechanisms of Cognitive Control Project Braver et al. (2021) and will be made available upon publication. Scripts for primary analyses will be made available at `https://github.com/mcfreund/trr`. For detailed descriptions of the task design, see Braver et al. (2021), and of image pre-processing and quality control, see Etzel et al. (2022).

6 Author Contributions

Michael C. Freund: *Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing–original draft, Writing–review & editing*. Ruiqi Chen: *Formal Analysis, Methodology, Software, Visualization, Writing–review & editing*. Gang Chen: *Methodology, Writing–review & editing*. Todd S. Braver: *Conceptualization, Funding Acquisition, Resources, Supervision, Writing–review & editing*.

7 Funding

This research was supported by National Institutes of Health grants R37 MH066078 and R21 AT009483 awarded to T.S.B.

8 Declaration of Competing Interests

None declared.

9 Acknowledgements

We thank all current and former Dual Mechanisms of Cognitive Control team members for their efforts in collecting and curating the dataset on which this study is based. In particular, we thank Kevin Okansen, Leah Newcomer, Erin Gourley, Maria Gehred, Alex Kizhner, Rachel Brough, Al Tay, Catherine Tang, and Anxu Wang, for their major efforts in data collection. We thank Carolina Ramirez and Mitch Jeffers for developing the data preprocessing codebase and Nicholas Bloom for developing the experiment software. We thank Joset Etzel for project management and for developing quality control criteria we used to select our sample of participants. In addition, we thank the members of the Cognitive Control and Psychopathology Lab for many fruitful discussions and feedback. M.C.F. also thanks Aaron Cochrane for helpful comments on a draft of this manuscript.

factor	index	levels
trial	t	$1, ..., T_{c,p,s}$
condition	c	incon, congr
(spatial) model	m	univ, multiv
participant	p	$1, ..., P$
(brain) region	n	$1, ..., N = 400$
repetition	r	test, retest
session	s	baseline, proactive, reactive
vertex	v	$1, ..., V_n$

Table 2: Experimental factors and indices.

10 Supplemental Method

Compared to the main text, this Supplement contains a more comprehensive and unified description of the spatial brain activity and reliability models we used.

10.1 Mathematical Notation

The transpose operation is written $^\top$. Non-bold font is used for scalars while lowercase bold is used for column vectors and uppercase bold for matrices. Italicized subscripts are used to index sets of scalars, vectors, or matrices: for example, $\mathbf{X}_{i,j}$ for $i \in 1, ..., I$ and $j \in 1, ..., J$ refers to the i,j-th matrix in a set of IJ matrices. Importantly, this subscript notation should not be confused with indices for elements of individual vectors or matrices. To avoid such confusion, we write element-wise indices with parenthetical superscripts: for example, $X_{i,j}^{(12)}$ refers to the element in the first row and second column of matrix $\mathbf{X}_{i,j}$. For clarity within inline text, all index notation will be omitted.

The specific notation used for factor indices are listed in Table 10.1.

10.2 Spatial brain activity models

The selective averaging procedure described within Section 2.4 furnishes a set of activation estimates per trial $t \in T$ and vertex $v \in V_n$, arranged within a data matrix $\mathbf{X}$ of size $T \times V_n$. These estimates were segmented into Schaefer-atlas regions $n \in N = 400$, centered (see Section 10.2.1), then submitted to univariate or multivariate spatial models $m \in \{\text{univ}, \text{multiv}\}$.

Thus, both univariate and multivariate spatial models can be written as a scalar projection of each trials' pattern:

$$\mathbf{y}_{p,s,r,n,c,m} = \mathbf{X}_{p,s,r,n,c}\mathbf{w}_{p,s,r,n,m} \tag{4}$$

where $\mathbf{y}$ is a vector of length T that holds the projections for each trial.

Following this notation, the univariate weights are those that estimate the spatial mean:

$$\mathbf{w}_{n,m=\text{univ}} = \mathbf{1}/V_n \tag{5}$$

where $\mathbf{1}$ is a vector of length V_n whose entries are all ones. Likewise, the multivariate weights were provided by the linear discriminant function:

$$a_{p,s,r,n}\mathbf{w}_{p,s,r,n,m=\text{multiv}} = \mathbf{S}^{-1}_{p,\neg s,r,n}(\bar{\mathbf{x}}_{p,\neg s,r,n,c=\text{incon}} - \bar{\mathbf{x}}_{p,\neg s,r,n,c=\text{congr}}) \tag{6}$$

where $\bar{\mathbf{x}}$ is the mean vector of activation estimates, averaged over trials and of length V_n; $\mathbf{S}$ is the mean within-condition covariance matrix, averaged over conditions and of size $V_n \times V_n$; and a is the Euclidean norm of the expression on the right-hand side, so that $\mathbf{w}$ is unit length. The multivariate models were fitted in leave-one-session out cross-validation, in which the weights for a given session s were estimated using concatenated data from the other two sessions, $\neg s$. To provide a moderate amount of robustness to the model, we regularized $\mathbf{S}$ by shrinking the off-diagonal values towards zero by a fixed amount of 25% (*cf.* Diedrichsen et al., 2016).

10.2.1 Centering

Prior to fitting the spatial models, the trial-level activation estimates from the timeseries model were centered. In particular, the values in each vertex were centered at their mean of condition means within each run. This centering is also known as “mean pattern” or “cocktail” centering (Misaki et al., 2010; Walther et al., 2016). The goal of implementing this centering was to reduce nuisance variance that resulted from global changes across different scanning runs and sessions (i.e., “global”, in the sense of changes that impact all conditions equally within each run). While this form of centering has been criticized for rendering linear correlations difficult to interpret (eGarrido et al., 2013), this concern does not apply in the present case with an LDA-based decoder (which implements a centering operation by default; see Misaki et al., 2010; Walther et al., 2016). To avoid imposing systematic differences between runs via centering, we estimated the mean using only the specific subset of trials that were present in each scanning run with matched stimulus features across runs and with balanced proportions of congruent and incongruent conditions within each run (see King et al., 2019 for a similar approach). Finally, we implemented these steps prior to fitting both multivariate and univariate spatial models, to ensure that the resulting model outputs $\mathbf{y}$, differ only due to differing weights $\mathbf{w}$, and not differing input data $\mathbf{X}$.

10.2.2 Stratified random undersampling

To minimize bias in decoding results caused by class imbalance and other confounding factors, we used a stratified random undersampling algorithm to generate the training set for LDA. By randomly sampling subsets of trials within our training set without replacement, we built a collection of 100 smaller training sets that contained an approximately equal number of trials k of each word and color within both congruent and incongruent conditions of each scanning run. We selected k as the minimum number of occurrences in any given combination of these factors within a given scanning run. Due to slight differences in trial balancing, the exact number k differed across across runs and sessions: $k \in \{4, 5\}$ in the baseline session, $k \in \{1, 2\}$ in the proactive session, and $k = 6$ in the reactive session. Note that these differences should not lead to systematic biases in the results, as we pool over both runs of two sessions when training decoders, so that the imbalance is systematically washed out.

10.3 Reliability models

The output of the spatial models, $\mathbf{y}$, forms the “outcome” variable within reliability models. Recall that we fit identical reliability models to outputs of both spatial model types and to each brain region, and that we also focused exclusively on modeling reliability within baseline-session projections (Section 2.6). Therefore, for clarity in the remaining notation, we omit indices for for model m, brain region n, and session s. For example, we now denote the output of spatial models as $y_{c,r,p,t}$.

10.3.1 Summary-statistic method

First, we averaged the trial-level activation estimates over trials, forming two summary scores per subject and repetition: $\bar{y}_{c,r,p} = \frac{1}{T_{c,p}} \sum_{t=1}^{T_{c,p}} y_{(c,r,p,t)}$. Thus, we refer to this method as the "summary statistic" method. The contrast of interest is the Stroop effect, that is, the difference between means of incongruent versus congruent trials: $\hat{y}_{r,p} = \bar{y}_{c=\text{incon},r,p} - \bar{y}_{c=\text{congr},r,p}$.

Then, at the population level, Stroop effects can be modeled by a set of Gaussian distributions:

$$\begin{aligned} \hat{y}_{r,p} &\sim \mathcal{N}(b_r + \beta_p, \sigma^2) \\ \beta_p &\sim \mathcal{N}(0, \xi^2) \end{aligned} \tag{7}$$

where b_r is the "fixed" (population-level) Stroop effect associated with repetition r, and β_p is the "random" (individual) Stroop effect associated with participant p, in general across sessions. Under this formulation 7, the intra-class correlation coefficient is defined as the proportion of variance:

$$\text{ICC}(3,1) = \frac{\xi^2}{\xi^2 + \sigma^2}. \tag{8}$$

Alternatively, it can be shown that ICC(3, 1) is the same as the Pearson correlation coefficient between the Stroop effects of both repetitions over the participants p:

$$\text{ICC}(3,1) = \frac{\xi^2}{\xi^2 + \sigma^2} = \frac{\text{Cov}(y_{r=\text{test},p}, y_{r=\text{retest},p})}{\sqrt{\text{Var}(y_{r=\text{test},p})\text{Var}(y_{r=\text{retest},p})}} = \text{Corr}(y_{r=\text{test},p}, y_{r=\text{retest},p}). \tag{9}$$

10.3.2 Hierarchical Bayesian modeling method

The full or "maximal" model structure is denoted in Equation 10. The rest of this section explains this formulation in detail.

$$\begin{aligned} y_{c,r,p,t} &\sim \mathcal{T}(\alpha_{c,r,p}, \nu_{c,r,p}) \\ \alpha_{c,r,p} &= m_{c,r} + \mu_{c,r,p} \\ \log \nu_{c,r,p} &= \gamma_{c,r} + \tau_{c,r,p} \\ (\boldsymbol{\mu}, \boldsymbol{\tau})_p &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} &= \begin{bmatrix} \Sigma^{(11)} & \dots & \Sigma^{(18)} \\ \vdots & \ddots & \vdots \\ \Sigma^{(81)} & \dots & \Sigma^{(88)} \end{bmatrix} \end{aligned} \tag{10}$$

We begin by assuming the observed output of the spatial models, y (*cf*, $\mathbf{y}$ in Equation 1), is generated from a set of Student's t-distributions, each of which are fully described by a location parameter α and scale parameter ν. Conceptually, these location and scale parameters of the t-distribution are analogous to the mean and standard deviation parameters of the Gaussian distribution. The t-distribution is used here instead of Gaussian, however, as its fatter tails provide greater robustness to outlying values (e.g., Chen et al., 2021). Importantly, the subscripts on α and ν indicate that individual parameters are estimated for each participant p, condition $c \in \{\text{incon}, \text{congr}\}$, and repetition $r \in \{\text{test}, \text{retest}\}$ combination.

Note that the the condition and repetition factors are parameterized with a "flat" coding scheme (also sometimes referred to as a "no-intercept" dummy coding scheme). Under this scheme, a given participant's α parameter coefficients specify their mean projections for each condition and repetition combination (e.g., incon within the test repetition).

Next, we assume that these location parameters $\alpha_{c,r,p}$ are a sum of two components: $m_{c,r}$, a population-level fixed effect, which all participants share; and $\mu_{c,r,p}$, a participant-level random effect, which indicates the deviation of the p-th participant's score from the group score $m_{c,r}$, within repetition r and condition c.

Similar to the location parameter, we decompose the (log of the) scale parameter into population and participant-level effects. We denote these effects with $\gamma_{c,r}$ and $\tau_{c,r,p}$, which are analogous to their location-parameter counterparts. In effect, these parameters enable the model to account for differing amounts of residual variability for each participant, repetition, and condition combination.

Finally, we assume that each participant's eight location and scale parameters, now written as a single 8-element vector $(\boldsymbol{\mu}, \boldsymbol{\tau})$, were sampled from a single multivariate Gaussian distribution with zero mean and covariance matrix $\mathbf{\Sigma}$. The covariance matrix $\mathbf{\Sigma}$ describes the linear relationships among these parameters over subjects.

Contained within $\mathbf{\Sigma}$ is the pivotal quantity of test-retest reliability: the correlation between test and retest repetitions in the Stroop contrast, denoted here as ρ. Due to the flat parameterization of the condition and repetition factors, however, ρ is not explicitly expressed as a single term in $\mathbf{\Sigma}$, but instead implicitly, as a linear combination of its rows and columns. To obtain ρ, we transform $\mathbf{\Sigma}$ by a 2-by-8 contrast matrix $\mathbf{W}$ that encodes this linear combination. Specifically, row r of $\mathbf{W}$ corresponds to the test or retest repetition, while column j corresponds to the j-th element of the participant-level random effect vector $(\boldsymbol{\mu}, \boldsymbol{\tau})$. Where j corresponds to $\mu^{(c=\text{incon},r)}$, $W^{(rj)} = 1$; where j corresponds to $\mu^{(c=\text{congr},r)}$, $W^{(rj)} = -1$; elsewhere, $W^{(rj)} = 0$. Applying this contrast to the random-effect covariance matrix, $\mathbf{W\Sigma W}^\intercal$, and dividing the resulting covariance element by the product of the standard deviations, yields the test-retest correlation in the Stroop effect, ρ.

10.4 Alternative Hierarchical Bayesian Models

10.4.1 Reduced model 1: independent location and scale (ILS)

The first simplification to the "full model" that we considered was omitting the random-effect covariances between the location $\boldsymbol{\mu}$ and scale $\boldsymbol{\tau}$ parameters. This simplification led to a model with the same general form as in Equation 10, except location $\boldsymbol{\mu}$ and scale $\boldsymbol{\tau}$ parameters were assumed to be generated by independent distributions, with independent covariance matrices $\mathbf{\Sigma}_l$, for $l \in \{\text{locat}, \text{scale}\}$.

From Equation 10, the reduced terms are as follows:

$$\begin{aligned}\boldsymbol{\mu}_p &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{l=\text{locat}})\\ \boldsymbol{\tau}_p &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{l=\text{scale}})\\ \boldsymbol{\Sigma}_l &= \begin{bmatrix}\Sigma^{(11)} & \dots & \Sigma^{(14)}\\ \vdots & \ddots & \vdots\\ \Sigma^{(41)} & \dots & \Sigma^{(44)}\end{bmatrix}_l\end{aligned} \tag{11}$$

10.4.2 Reduced model 2: independent location and scale, symmetric covariance structure (ILS Sym)

We further simplified the model with independent location and scale (Equation 11) by additionally assuming that the covariance structure was symmetric between repetitions (Chen et al., 2021). Under a symmetry assumption, the covariance between different conditions within different repetitions is constrained to be equal across permutations of repetitions. For example, a symmetric structure would entail that Cov(incon test, congr retest) equals Cov(congr test, incon retest). To parameterize this covariance structure, we use a different coding scheme to parameterize the condition factor than used in Equation 10. Now, we use a contrast coding scheme, $c' \in \{\text{mean, stroop}\}$, such that, for a given participant p and repetition r, $\alpha_{c'=\text{mean},r,p}$ represents the mean of incongruent and congruent condition means, and $\alpha_{c'=\text{stroop},r,p}$ represents the difference between incongruent and congruent means (i.e., the mean Stroop contrast). Under the symmetry assumption, the mean of incongruent and congruent conditions is independent from the mean Stroop contrast over subjects (Chen et al., 2021).

From Equation 10, the reduced terms are as follows:

$$\begin{aligned}\boldsymbol{\mu}_{p,c'} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{l=\text{locat},c'})\\ \boldsymbol{\tau}_{p,c'} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{l=\text{scale},c'})\\ \boldsymbol{\Sigma}_{l,c'} &= \begin{bmatrix}\Sigma^{(11)} & \Sigma^{(12)}\\ \Sigma^{(21)} & \Sigma^{(22)}\end{bmatrix}_{l,c'}\end{aligned} \tag{12}$$

Note that $\boldsymbol{\mu}$ is now a vector with 2 elements, corresponding to each repetition r. Under this contrast coding scheme, it is now no longer necessary to apply a contrast matrix to $\boldsymbol{\Sigma}$ to obtain the test-retest correlation in Stroop effects ρ.

10.4.3 Reduced model 3: independent location and scale, symmetric covariance structure, homogeneous scale (Homog.)

Finally, the simplest model we considered was a reduction of the model with symmetric covariance structure (Equation 12), in which we additionally assumed that the scale of the residuals, ν, was constant across all conditions, repetitions, and participants.

From Equation 10, the reduced terms are as follows:

$$\begin{aligned} y_{c',r,p,t} &\sim \mathcal{T}(\alpha_{c,r,p}, \nu) \\ \boldsymbol{\mu}_{p,c'} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{c'}) \\ \boldsymbol{\Sigma}_{c'} &= \begin{bmatrix} \Sigma^{(11)} & \Sigma^{(12)} \\ \Sigma^{(21)} & \Sigma^{(22)} \end{bmatrix}_{c'} \end{aligned} \tag{13}$$

Note that, although this homogeneous-variance assumption is likely unrealistic, it is typically made by default: for example, it is required by popular hierarchical-modeling libraries such as *lme4*.

10.5 Analysis of Hierarchical Model Parameters

10.5.1 Population-level Stroop effect

The posterior distribution (MCMC samples) of the ”fixed” (population-level) effect of congruency for each wave ($m_{c'=\text{stroop},r}$ in Equation 10 with re-coded contrast c') was extracted from the models through the *fixef()* function of *brms*. Samples from this distribution were averaged over repetitions r. Then, a t^+ statistic was computed as the ratio between the mean and standard deviation of the posterior distribution.

10.5.2 Point estimate and precision of test–retest reliability

The posterior distribution of TRR ($\Sigma^{(12)}_{l=\text{locat},c'=\text{stroop}}$ in Equation 12) was extracted through the *VarCorr()* function of *brms*. The dispersion of the posterior distribution was summarized by the precision, which is the inverse of the standard deviation.

10.5.3 Variability ratio

To quantify the magnitude of trial-level noise, we computed a ratio: the variability of the Stroop effect within-individuals relative to the variability between individuals. The particular calculation is as follows: for every MCMC sample, the trial-level variability ν is given by the population-level mean of the scale parameter averaged over repetitions ($\nu = \exp \frac{\gamma_{(c'=\text{mean},r=\text{test})} + \gamma_{c'=\text{mean},r=\text{retest}}}{2}$ in Equation 10), while the individual-level variability σ is given by the “random” effect of Stroop averaged over repetitions ($\sigma = \frac{\sqrt{\Sigma^{(11)}_{l=\text{locat},c'=\text{stroop}}} + \sqrt{\Sigma^{(22)}_{l=\text{locat},c'=\text{stroop}}}}{2}$ in Equation 12). The variability ratio $\frac{\nu}{\sigma}$ was computed for each sample and then summarized by the mode of the distribution.

11 Supplemental Results

11.1 Model Comparisons

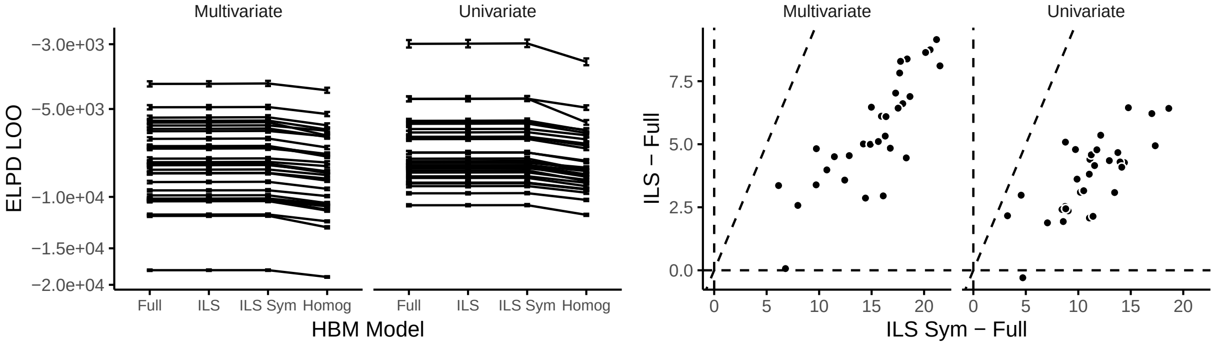

Figure 9: **Results of a model comparison.** This model comparison was conducted on 32 brain parcels (see Method). Expected-log pointwise predictable density (ELPD) was measured in a leave-one-out manner (LOO). **Left**, Lines connect ELPD LOO estimates (y axis) across different models (x axis) from the same parcel. More positive values indicate a better fit, in terms of better estimated ability of the model to account for out-of-sample datapoints. The pattern of ELPD LOO is highly consistent across parcels. Note that because statistics from the Homog. model were considerably lower (worse) than others, the y-axis spacing is non-linear (inverse hyperbolic sine function). **Right**, Within-parcel contrasts of ELPD LOO. Each point is a parcel. X and y axes illustrate the difference between ELPD LOO for the respective models. Dashed lines illustrate unity line and x and y intercepts. The pattern of ELPD LOO is highly consistent across parcels. On the x-axis, most parcels lie above 0, indicating ILS Sym was preferred over the Full model. On the y-axis, most parcels lie above 0, indicating ILS was preferred over the Full model. Additionally, all parcels lie to the left (underneath) the unity line, indicating ILS Sym was preferred over ILS.

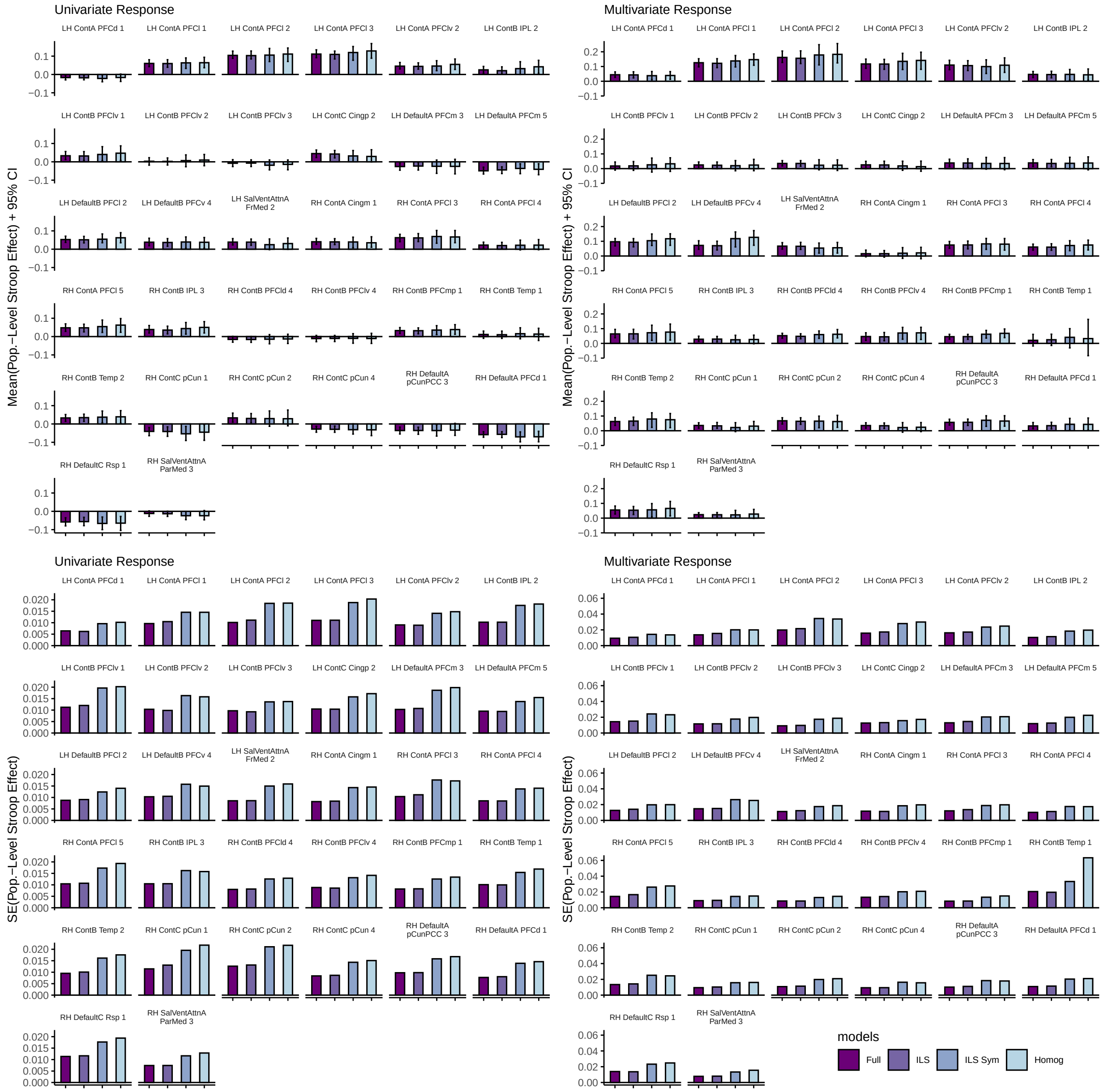

Figure 10: **Estimates of population-level Stroop contrasts for all reliability models fitted in 32 brain parcels used for model comparison. Top**, Bar heights illustrate the mean of the posterior, with errorbars illustrating 95% CI. **Bottom**, Bar heights illustrate the standard error in the mean, measured as the SD of the posterior distribution.

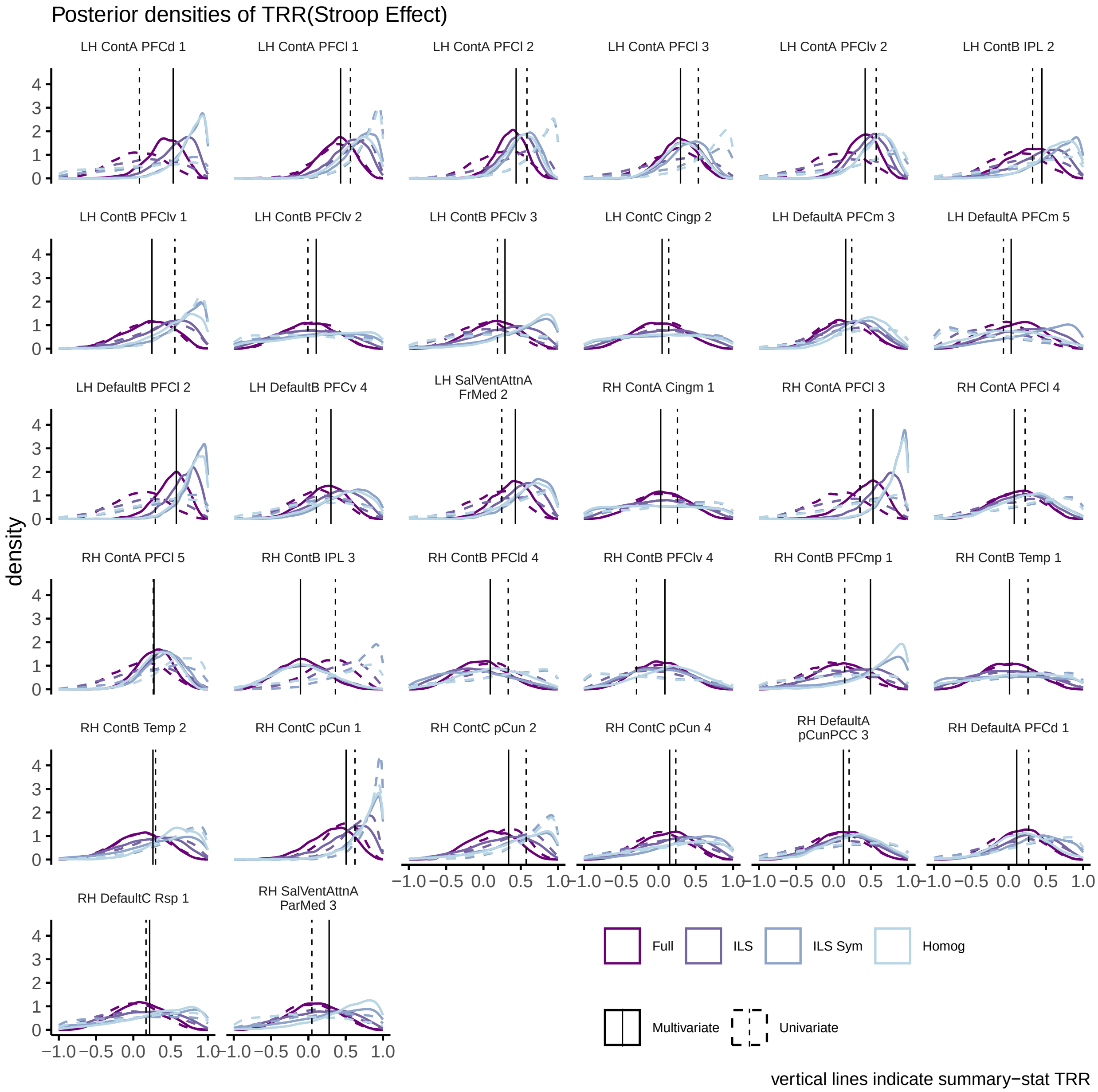

Figure 11: **Posterior densities for individual-level test–retest correlations for all reliability models fitted in 32 brain parcels used for model comparison.**

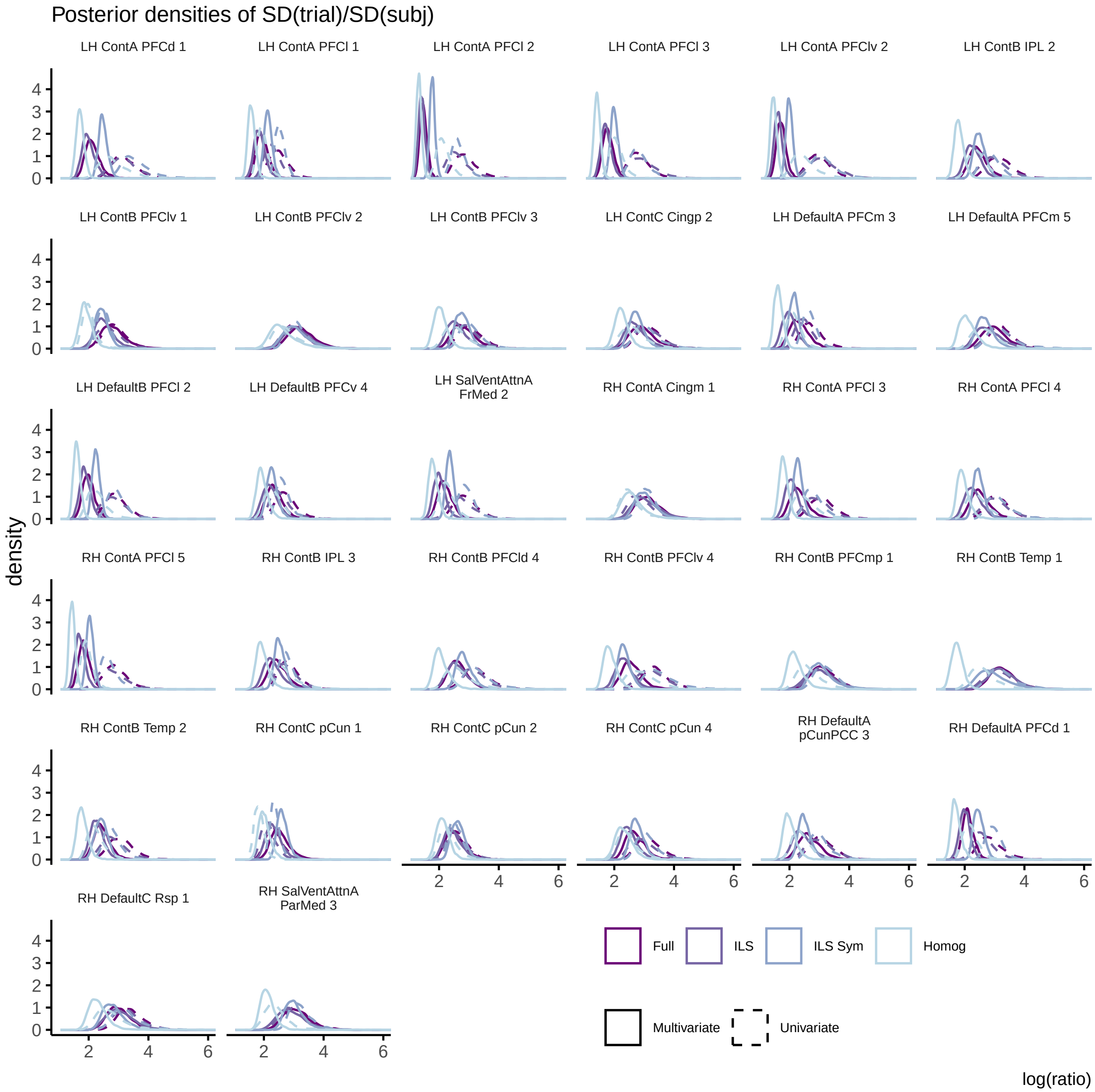

Figure 12: **Posterior densities for trial/subject variability ratios for all reliability models fitted in 32 brain parcels used for model comparison.**

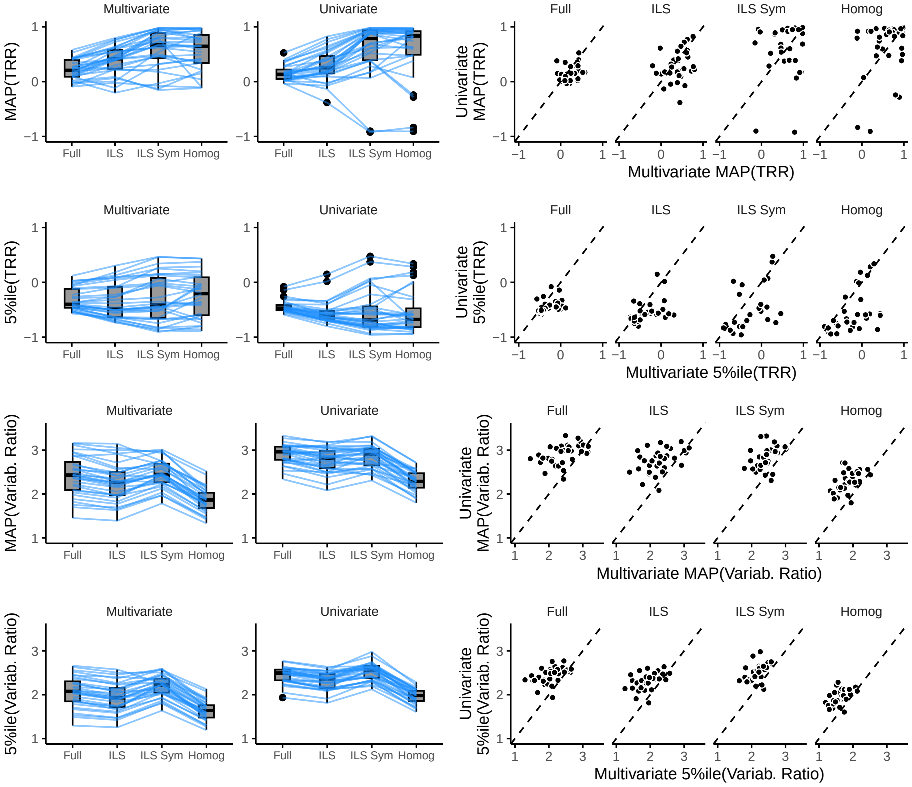

Figure 13: **Statistics of posteriors for all reliability models fitted in 32 brain parcels used for model comparison.**

References

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

Bejjanki, V. R., da Silveira, R. A., Cohen, J. D., and Turk-Browne, N. B. (2017). Noise correlations in the human brain and their impact on pattern classification. *PLoS Computational Biology*, 13(8).

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology*, 53(1):605–634.

Braver, T. S., Cole, M. W., and Yarkoni, T. (2010). Vive les differences! Individual variation in neural mechanisms of executive control. *Current Opinion in Neurobiology*, 20(2):242–250.

Braver, T. S., Kizhner, A., Tang, R., Freund, M. C., and Etzel, J. A. (2021). The Dual Mechanisms of Cognitive Control Project. *Journal of Cognitive Neuroscience*, 33(9):1990–2015.

Bürkner, P.-C. (2017). Advanced Bayesian Multilevel Modeling with the R Package brms.

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., and Haller, S. P. (2021). Trial and error: A hierarchical modeling approach to test-retest reliability. *NeuroImage*, 245:118647.

Cole, M. W., Repovš, G., and Anticevic, A. (2014). The Frontoparietal Control System: A Central Role in Mental Health. *The Neuroscientist*, 20(6):652–664.

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., and Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage*, 97:271–283.

DeYoung, C. G., Sassenberg, T. A., Abend, R., Allen, T., Beaty, R., Bellgrove, M., Blain, S. D., Bzdok, D., Chavez, R. S., Engel, S. A., Ma, F., Fornito, A., Genç, E., Goghari, V., Grazioplene, R., Hanson, J. L., Haxby, J., Hilger, K., Homan, P., Joyner, K., Kaczkurkin, A., Latzman, R. D., Martin, E. A., Passamonti, L., Pickering, A., Safron, A., Servaas, M., Smillie, L., Spreng, R. N., Tiego, J., Viding, E., and Wacker, J. (2022). Reproducible between-person brain-behavior associations do not always require thousands of individuals.

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, 64(1):135–168.

Diedrichsen, J., Provost, S., and Zareamoghaddam, H. (2016). On the distribution of cross-validated Mahalanobis distances. *arXiv:1607.01371 [stat]*.

eGarrido, L., eVaziri-Pashkam, M., eNakayama, K., and eWilmer, J. (2013). The consequences of subtracting the mean pattern in fMRI multivariate correlation analyses. *Frontiers in Neuroscience*, 7.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., and Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, page 0956797620916786.

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, 11(1):19–23.

Esteban, O., Ciric, R., Finc, K., Blair, R. W., Markiewicz, C. J., Moodie, C. A., Kent, J. D., Goncalves, M., DuPre, E., Gomez, D. E. P., Ye, Z., Salo, T., Valabregue, R., Amlien, I. K., Liem, F., Jacoby, N., Stojić, H., Cieslak, M., Urchs, S., Halchenko, Y. O., Ghosh, S. S., De La Vega, A., Yarkoni, T., Wright, J., Thompson, W. H., Poldrack, R. A., and Gorgolewski, K. J. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nature Protocols*, 15(7):2186–2202.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. (2018). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, page 1.

Etzel, J. A., Brough, R. E., Freund, M. C., Kizhner, A., Lin, Y., Singh, M. F., Tang, R., Tay, A., Wang, A., and Braver, T. S. (2022). The Dual Mechanisms of Cognitive Control dataset, a theoretically-guided within-subject task fMRI battery. *Scientific Data*, 9(1):114.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188.

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., and Turner, B. M. (2020). Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox.

Harrison, S. A. and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238):632–635.

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

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50(3):1166–1186.

King, M., Hernandez-Castillo, C. R., Poldrack, R. A., Ivry, R. B., and Diedrichsen, J. (2019). Functional boundaries in the human cerebellum revealed by a multi-domain task battery. *Nature Neuroscience*, 22(8):1371–1378.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C. E., Kepecs, A., Mainen, Z. F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C. K. (2016). Demixed principal component analysis of neural population data. *eLife*, 5:e10989.

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., and Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*, 32(4):622–626.

Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLOS ONE*, 12(5):e0177459.

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

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

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G. M., Uriarte, J., Snider, K., Lynch, B. J., Wilgenbusch, J. C., Pengo, T., Tam, A., Chen, J., Newbold, D. J., Zheng, A., Seider, N. A., Van, A. N., Metoki, A., Chauvin, R. J., Laumann, T. O., Greene, D. J., Petersen, S. E., Garavan, H., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., and Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660.

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*, 7:e6918.

Misaki, M., Kim, Y., Bandettini, P. A., and Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, 53(1):103–118.

Mumford, J. A., Turner, B. O., Ashby, F. G., and Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, 59(3):2636–2643.

Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage*, 222:117254.

Rosenberg, M. D. and Finn, E. S. (2022). How to establish robust brain–behavior relationships without thousands of individuals. *Nature Neuroscience*, 25(7):835–837.

Roth, Z. N., Heeger, D. J., and Merriam, E. P. (2018). Stimulus vignetting and orientation selectivity in human visual cortex. *eLife*, 7:e37241.

Rouder, J. N. and Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, 26(2):452–467.

Rouder, J. N., Kumar, A., and Haaf, J. M. (2023). Why many studies of individual differences with inhibition tasks may not localize correlations. *Psychonomic Bulletin & Review*.

Saville, C. W. N., Pawling, R., Trullinger, M., Daley, D., Intriligator, J., and Klein, C. (2011). On the stability of instability: Optimising the reliability of intra-subject variability of reaction times. *Personality and Individual Differences*, 51(2):148–153.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114.

Shallice, T. and Burgess, P. W. (1991). DEFICITS IN STRATEGY APPLICATION FOLLOWING FRONTAL LOBE DAMAGE IN MAN. *Brain*, 114(2):727–741.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Snijder, J.-P., Tang, R., Bugg, J. M., Conway, A. R. A., and Braver, T. S. (2023). On the psychometric evaluation of cognitive control tasks: An Investigation with the Dual Mechanisms of Cognitive Control (DMCC) battery. *Behavior Research Methods*.

Sonkusare, S., Breakspear, M., and Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically Acclaimed. *Trends in Cognitive Sciences*, 23(8):699–714.

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology*, 15(2):201–293.

Spisak, T., Bingel, U., and Wager, T. D. (2023). Multivariate BWAS can be replicable with moderate sample sizes. *Nature*, 615(7951):E4–E7.

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

Taylor, P. A., Reynolds, R. C., Calhoun, V., Gonzalez-Castillo, J., Handwerker, D. A., Bandettini, P. A., Mejia, A. F., and Chen, G. (2023). Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. *NeuroImage*, 274:120138.

Underwood, B. J. (1975). Individual differences as a crucible in theory construction. *American Psychologist*, 30(2):128–134.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2024). Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klaR analyzing german business cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343, Berlin. Springer-Verlag.

Wilkinson, G. N. and Rogers, C. E. (1973). Symbolic Description of Factorial Models for Analysis of Variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3):392–399.

Williams, D. R., Martin, S. R., and Rast, P. (2019). Putting the Individual into Reliability: Bayesian Testing of Homogeneous Within-Person Variance in Hierarchical Models. Preprint, PsyArXiv.

Xu, Z., Adam, K. C. S., Fang, X., and Vogel, E. K. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods*, 50(2):576–588.

Yoo, K., Rosenberg, M. D., Noble, S., Scheinost, D., Constable, R. T., and Chun, M. M. (2019). Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *NeuroImage*, 197:212–223.

Zorowitz, S. and Niv, Y. (2023). Improving the Reliability of Cognitive Task Measures: A Narrative Review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(8):789–797.