The response time paradox in functional magnetic resonance imaging analyses

Jeanette A. Mumford[1*], Patrick G. Bissett[1], Henry M. Jones[2], Sunjae Shim[1], Jaime Ali H. Rios[1], Russell A. Poldrack[1]

[1]Department of Psychology, Stanford University, Stanford, CA, United States of America.
[2]Department of Psychology, University of Chicago, Chicago, IL, United States of America.

*Corresponding author(s). E-mail(s): jeanette.mumford@gmail.com;

Abstract

Response times (RTs) are often the main signal of interest in cognitive psychology, but are often ignored in functional MRI (fMRI) analyses. In fMRI analysis the intensity of the signal serves as a proxy for the intensity of local neuronal activity, but changes in either intensity or duration of neuronal activity can yield identical fMRI signals. Therefore, if RTs are ignored and pair with neuronal durations, fMRI results claiming intensity differences may be confounded by RTs. We show how ignoring RTs goes beyond this confound, where longer RTs are paired with larger activation estimates, to lesser-known issues where RTs become confounds in group-level analyses and, surprisingly, how the RT confound can induce other artificial group-level associations with variables that are not related to the condition contrast or RTs. We propose a new time-series model to address these issues and encourage increasing focus on what the widespread RT-based signal represents.

Introduction

The goal of task-based functional magnetic resonance imaging (fMRI) studies is to infer involvement of particular brain regions or networks in specific cognitive functions. These studies are often designed using the subtraction logic first developed by Donders (1969) for the analysis of response times (RTs), in which comparisons are made between different task conditions thought to differ with regard to involvement of some specific cognitive function(s). For example, in the well-known Stroop task, stimuli are presented where color and text of words are either congruent (e.g. “blue” presented in blue) or incongruent (e.g. “blue” presented in red) (Stroop, 1935). Individuals are consistently slower at naming stimulus color when the written word is incongruent compared to congruent, and this difference in response times is interpreted as indexing engagement of an additional cognitive process in the incongruent condition, such as conflict detection or resolution (Botvinick et al., 2001). Similarly, greater activation in regions such as the dorsal medial frontal cortex (dMFC) in fMRI studies of the Stroop task have been interpreted as reflecting a specific role in these cognitive processes (Botvinick et al., 1999; MacDonald et al., 2000; Kerns et al., 2004).

The facile nature of the common inference from activation to “involvement” belies the complexity of the link between fMRI signals and underlying neuronal activity (cf. Logothetis (2008)). Here we focus on disambiguating these interpretations of activation: namely, whether a difference in activation reflects differential engagement of a particular computation, or engagement of the same computation for a different amount of time. Because of the slow nature of the blood oxygen-level dependent (BOLD) response, measured in most fMRI studies, it is difficult to distinguish the degree to which a difference in evoked BOLD response reflects an increase in amplitude of neuronal response versus a difference in response duration (Figure 1). This indeterminacy has been known since the early days of fMRI (Savoy et al., 1995; Jezzard et al., 2001) and establishes the importance of considering the potential for differences in duration of neural activity to confound amplitude estimates in some fMRI tasks.

For clarity, we define some terms used throughout the present paper. Time-series level analysis refers to linear models of fMRI time-series. Group-level analyses refer to analyses of estimated fMRI contrasts across a set of subjects including a single group average (1-sample t-test), group average comparisons (2-sample t-test) and linear associations with a covariate (e.g., phenotype). Between-trial RT adjustment is formally carried out in the time-series level. Between-subject RT confounds will only impact group-level models involving group comparisons or associations but not single group averages.

Over a decade ago, Grinband et al. (2008) started the discussion about between-trial modeling of response times in fMRI analysis. They proposed a variable epoch model, where trial-by-trial neuronal durations were assumed to track with RTs as opposed to a common modeling practice that assumed each trial was sufficiently modeled as a brief impulse of constant duration (e.g., .1s). The impulse model was considered to be adequate due to the belief that differences in RTs would not be detectable. The variable epoch model yielded more powerful results than the impulse model or an impulse model with an RT-modulated regressor. Importantly, this model's performance studied within-subject power for a single condition versus baseline and assumed the neuronal activation duration mirrors the RTs. The assumption that neural activity duration scales with RT may be particularly suitable when the neural signature is theoretically linked to the accumulation of information towards a response, as in the accumulation of activity towards a response in the Diffusion Decision Model (Ratcliff, 1978). There are likely tasks and regions of the brain for which this assumption and therefore this modeling strategy is appropriate, but this is a strong theoretical commitment that may not be suitable for all tasks or for the whole brain. How this model performs for condition differences and when model assumptions are violated has not been well studied.

Yarkoni et al. (2009) examined the relationship between RT and fMRI signal across a variety of tasks in an effort to better understand how RTs were related to BOLD signal amplitude. RT-driven amplitude differences were found to be due to "time on task" or simply duration differences in the neuronal signal across trials (constant activity that varies in duration), as opposed to differences in the magnitude of neuronal activity. A compelling result of this work was the identification of a widespread network of brain regions that showed significant correlation with RTs across a variety of tasks including the dMFC which was previously described as reflecting conflict in the Stroop task. This calls into question how RT-based activation differences might be interpreted given they are present across many tasks.

A subsequent series of papers focused on the Stroop incongruent versus congruent contrast and whether dMFC-based activation reflected conflict, as proposed by a prominent theory Botvinick et al. (2001). This inspiring discourse across multiple publications illustrates the challenge of interpreting RT-correlated activation and demonstrates the rigorous work required to combine behavioral theory of a task with imaging analysis results when RT-correlated activation is found. Both Grinband et al. (2011) and Carp et al. (2010) showed differences in activation between slow and fast congruent trials were similar to differences between all congruent and incongruent trials in the dMFC, indicating the commonly found effect was driven by RTs. Follow-up work by Yeung et al. (2011) suggested that in their theory of conflict monitoring, RT and conflict could not be so cleanly dissociated. Ultimately, a consensus was not reached (Brown, 2011; Grinband et al., 2011; Nachev, 2011) and it did not have widespread impact on task modeling or interpretation. Of the 22 publications identified from a PubMed search for "stroop task fmri 2021", only 4 addressed RT in their analyses and interpretation of their results. It is beyond the scope of this present work to come to an agreement on how to interpret Stroop-based fMRI activation maps. Our goal is to revive the discussion of how RTs can impact fMRI results and offer a new modeling framework that flexibly unconfounds RT from condition differences.

Limited focus has been given to links connecting RT adjustment in time-series analyses to between-subject RT confounds in group-level analyses. For example, if between-trial RT adjustment is ignored in the Stoop task, incongruent versus congruent fMRI differences are likely to correlate with incongruent versus congruent RT differences. In earlier work (Carp et al., 2012), this relationship was alluded to in a result where correlation between age and the Stoop incongruent versus congruent contrast changes when between-trial RT adjustment is performed. This work highlights the possibility of between-trial RT adjustment impacting between-subject analyses, but the link has not been formally defined. We formalize the relationship and reveal a surprising issue where ignored between-trial RT variability can introduce other confounds *not* related to RTs in group-level models. For example, if "time spent outdoors" (TO) is equally related to two conditions, but not the condition difference, failure to address between-trial RT variability can introduce a TO effect to group-level analyses of the condition difference contrast *even when RT is not related to TO.* This problem

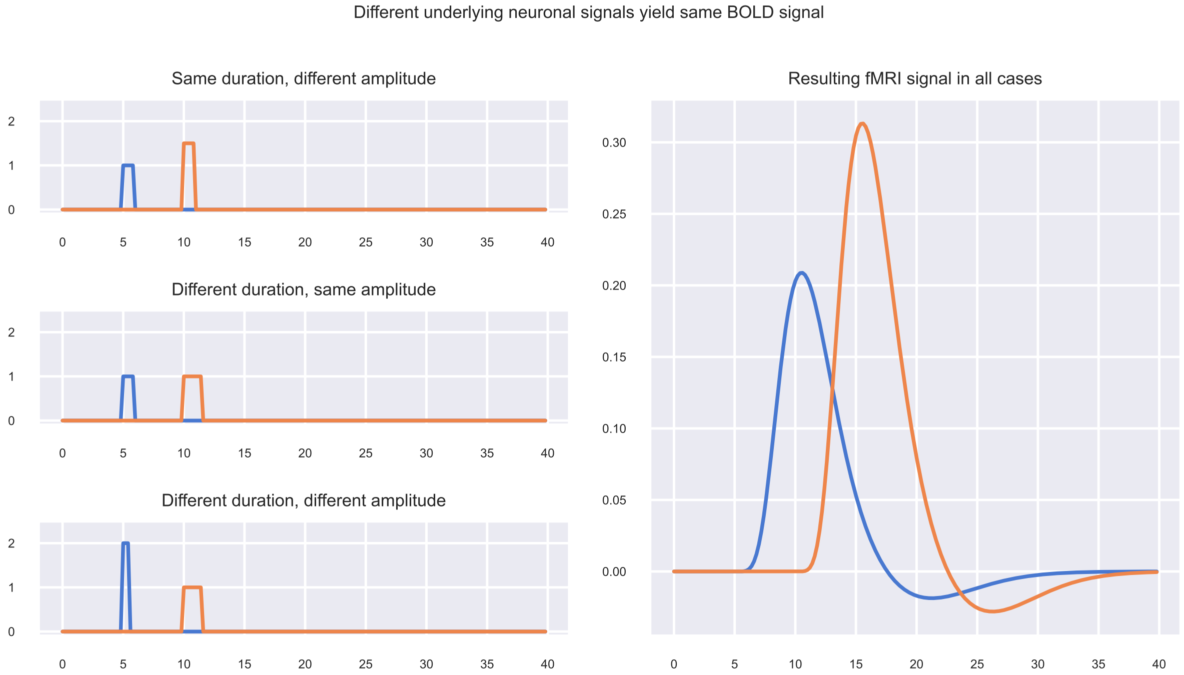


Fig. 1: How amplitude and duration of neuronal signal interact to yield similar BOLD responses. The left hand column shows 3 different examples of neuronal signals that evoke the same BOLD response shown in the right hand panel.

cannot be fixed within group-level analyses, but we show here that it can be eliminated using an improved time-series analysis.

The overarching goal of the present work is to revive interest in understanding and addressing RT-correlated activation to improve our interpretation of the fMRI signal. The standard modeling approach ignores RTs and is at risk of inflated Type I errors whenever RT effects are present, which includes a large set of voxels in the brain. BOLD signals can also relate to RTs at different levels for each condition in an interaction effect and the model proposed by Grinband et al. (2008) is more appropriate to use in regions where an interaction is hypothesized, but has limitations for whole brain analysis. We extend and improve upon previous ideas to incorporate RT into fMRI data analyses, proposing a new modeling strategy that flexibly accommodates both areas with signal durations that scale with RTs and those that do not, with controlled type I error rates, making it particularly suitable for the common procedure of whole brain analysis. We show how conflict between model assumptions and data generating process can inflate type I error rates, reduce power and introduce both RT-related and RT-unrelated confounds in group-level analyses. The presence of group level confounds is an especially important consideration when using contrast estimates supplied by databases for large neuroimaging studies, since the solution requires repeating the time-series analysis, which may not be possible for many users of these databases. Last, we replicate the findings of Yarkoni et al. (2009) by demonstrating a large RT-based activation pattern across a set of 7 fMRI tasks, each with approximately 91 subjects.

Results

Simulations

Statistical models used for fMRI data generally involve convolution of a vector representing trial or stimulus onsets with a canonical hemodynamic response function (Figure 1) to create regressors for use in linear modeling (Poldrack et al., 2009). Trials can be represented either as delta functions or as boxcar functions with some duration; when a boxcar is used, the boxcar duration is set to a constant value such as stimulus duration or a brief default value (e.g., 0.1 second). In Figure 2 this model is referred to as “Constant duration, no RT” (hereafter as ConstDurNoRT). Because of the indeterminacy described above (Figure 1), the specific constant value used for stimulus duration will not generally impact the statistical inferences derived from the model, as it will simply scale the values of the parameter estimates along with their variances (assuming the trial durations are relatively short, <2s). This standard approach does not include any information about response times; thus, if two conditions differ in RTs when true activation magnitudes do not differ, the condition with longer RTs may have higher estimated activation if the signal durations track RTs.

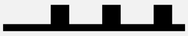
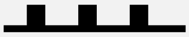
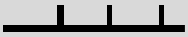
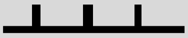
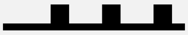
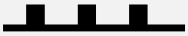
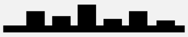
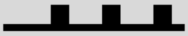
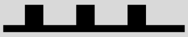
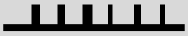
Model name		Unconvolved regressor	Duration	Modulation
1	Constant Duration, no RT (ConsDurNoRT)		.1s	None
			.1s	None
2	RT Duration (RTDur)		RT	None
			RT	None
3	Constant Duration, RTMod (ConsDurRTMod)		.1s	None
			.1s	None
			.1s	RT*
4	Constant Duration, RTDur (ConsDurRTDur)		.1s	None
			.1s	None
			RT*	None

Fig. 2: Models assessed in the simulation study. Models are described by name, unconvolved regressor visualization, duration used for boxcars of unconvolved regressors and definition of the modulation used, when present. Convolved regressors were used in data generation and modeling. The first model does not include any response time information, the second model addresses RT through the duration of the regressors and the third and fourth models add an RT regressor to the first model either using an RT modulated regressor (Model 3) or an RT duration regressor (Model 4). *See Discussion Section for details on why RT is not centered for the RT modulated regressor of Model 3 and why the RT Duration regressor of Model 4 is not orthogonalized.

Grinband et al. (2008) developed a modeling approach to address response times in fMRI data, where boxcar durations for each trial varied by trial response times (labeled as "RT Duration/RTDur" in Figure 2). This approach will appropriately scale the parameter estimates for regions in which neural activity durations match RT durations, which we will refer to as "duration scales with RT". This models an interaction and since constant duration regressors are not included, the model implies each condition has a different linear relationship between BOLD activation and RT, and that BOLD activation is 0 when RT is 0. One shortcoming of this model is it will not correctly model activation in regions where neural activity duration does *not* scale linearly with RTs.

To address these issues, we created a generalized model of RT that can identify RT effects separately from the task effect (corrected for RT); the two implementations of this model are shown as "Constant Duration, RTMod" (ConsDurRTMod) and "Constant Duration, RTDur" (ConsDurRTDur) in Figure 2. Each of these models starts with the ConsDurNoRT regressors and adds a single RT regressor.ConsDurRTDur models RTs through duration and ConsDurRTMod models RTs through parametric modulation. In both cases any differences in RTs between conditions will be removed by the RT regressor, leaving the condition difference effects to be interpreted as unconfounded estimates of activation in relation to the experimental manipulation. This model can be extended to a full interaction model by splitting the single RT regressor into two RT regressors, one for each condition. This will be further described in the Power Analysis results. Notably we have not orthogonalized the RT regressor in ConsDurRTDur or mean centered the RTs in the RT regressor of ConsDurRTMod, which will not have any impact on the estimate of the contrast of interest (condition difference) but may negatively impact the interpretation of other contrasts or introduce RT confounds in group level analyses. This will be further discussed in the Discussion section.

Response time data were simulated based on RTs from two different tasks: the Stroop task (based on our data) and reported RT distribution parameters from a two-alternative, forced-choice categorization task from Grinband et al. (2008). In each case RTs were generated by sampling from an ex-Gaussian distribution (Ratcliff and Murdock, 1976); the specified ex-Gaussian parameters led to RTs that were generally longer for the two-alternative, forced-choice categorization task (mean = 1337, sd = 706.5) compared to Stroop (mean = 690, sd = 177.5). Another difference is the variance relative to the mean is smaller for the Stroop task (coefficient of variation of .528 and .257 for the two-alternative, forced-choice categorization and Stroop tasks, respectively). Interstimulus intervals (ISIs) were sampled from a Uniform distribution and condition order was randomly presented. Time-series data where duration scales with RT were generated using the RTDur model and data where duration did not scale with RT were generated using the ConstDurNoRT model.

For all models the contrast of interest was condition 2 - condition 1 and all simulation-based results correspond to group-level analyses with 100 subjects. See the Methods for further details.

Error rates and power

We first assessed the false positive rate for each model on each of the simulated data sets for the condition comparison contrast (Figures 3, Extended Data Fig. 1). In all cases the ConsDurRTDur model appropriately controlled Type I error but the ConsDurRTMod model failed to control error rates for large RT differences (>1s) when the signal duration scales with RT because longer RTs (>2s) are not fit as well by an RT modulated regressor. The modulated regressor assumes the BOLD activation increases linearly with duration, which is an assumption that only holds well for RTs under 2s, which will be explored in detail in the "RT-based and RT-driven group-level confounds" section. Due to model assumption violations, ConstDurNoRT had inflated error rates when activation duration did scale with RT, and RTDur had inflated error rates when the signal duration did not scale with RT. Thus, the most commonly used model for task fMRI analysis, ConstDurNoRT, suffers from substantial inflation of false positives in the face of RT differences between conditions, because it inaccurately attributes the confounding RT signal to differences in the intensity of the underlying neuronal signal. Larger Type I error rates observed with the Stroop-based RT reflects a reduced RT standard deviation compared to the forced-choice categorization task, making RT-based differences easier to detect. Results with a longer ISI (3-6s), are similar (Extended Data Fig. 1).

Only models with controlled Type I errors are considered in the power analyses since the inflated error rates indicate bias in the contrast estimates. When signal duration does not scale with RTs, power for ConsDurNoRT (true model) and ConsDurRTDur are compared (right panels, Extended Data Fig. 2) and ConsDurRTMod is only considered if RTs are less than 2s (right panels, Figure 4). In this data setting, the true model, ConsDurNoRT, has maximum power and ConsDurRTDur/ConsDurRTMod have negligable power loss when the RT difference is .1s while ConsDurRTDur has a slight power reduction when the RT differences is .8s (Extended Data Fig. 2). For large RT differences there can be mild collinearity between the RT regressor and the conditions, which explains the power loss.

When signal duration does scale with RTs, power is considered for RTDur (true model) and ConsDurRTDur. The data in this setting follow an interaction effect, so we added a 4 regressor interaction model that combines the regressors from RTDur and ConsDurNoRT, where the contrast of interest is the comparison of the RTDur regressors. When RTs are <2s (Figure 4), we add a second 4 regressor interaction model using RT modulated regressors, which has been proposed by Carp et al. (2010, 2012); Weissman and Carp (2013). In this setting the RTDur model will have maximal power (orange line, left panels Figures 4 and Extended Data Fig. 2). Interestingly, the ConsDurRTDur model has only a slight power loss when the RT difference is large (red lines, Extended Data Fig. 2) and is negligible for the RT difference of .1s (Figure 4). See Extended Data Fig. 3 for an illustration of how condition differences from ConsDurRTDur can approximate RTDur differences in this data setting. The 4 regressor interaction models have large losses in power for both RT difference settings (Extended Data Fig. 2 and Figure 4). This large power loss is due to variable contrast estimates caused by high collinearity between RT regressors and condition regressors, where the variance inflation factors for the contrast of interest were almost always above the commonly accepted threshold of 5.

RT-based and RT-driven group-level confounds

The foregoing analyses, along with the previous work by Grinband, focused on confounding of RT between-trials, which impacts average condition effects. Here we introduce a new problem of between-*subject* RT confounds. Within-subject differences in average RT, corresponding to the contrasted conditions, can confound group-level analyses involving group comparisons or associations. For example, the incongruent versus

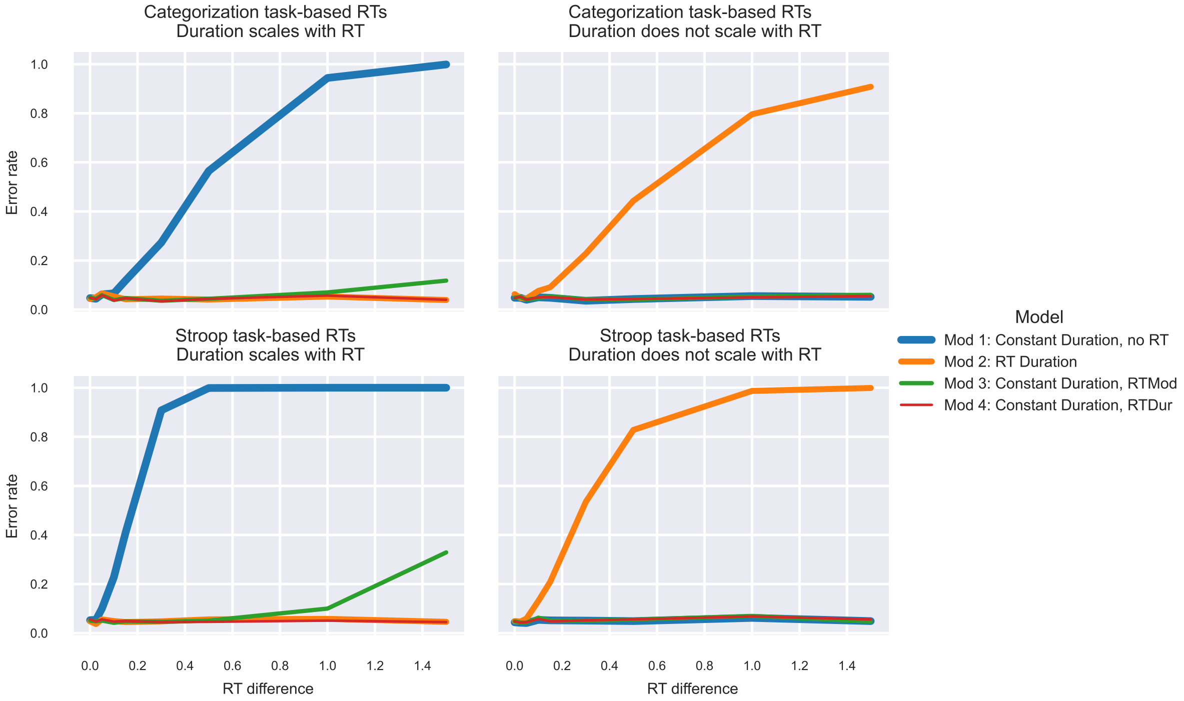


Fig. 3: Type I error as RT difference between conditions increases. The two-alternative, forced-choice categorization task RT distribution was used in the top panels, while Stroop RT distribution was used in the bottom panels. An ISI between 2-4s was used and the inference of interest was for the 1-sample t-test of the condition difference effect with 100 subjects. 2500 simulations were used to calculate the error rate. Models with error rates larger than .05 are not valid.

congruent BOLD contrast estimate may correlate with the differences in average RTs for incongruent and congruent conditions. This is of particular interest given the increasing focus on analyses of brain-behavior correlations in fMRI literature (e.g., Dubois and Adolphs (2016)).

The driving factor of correlations between condition differences in brain activation and corresponding differences in RTs is simply due to the relationship between the activation estimate and RT when the data and model assumptions are in conflict. For example, if signal durations scale with RTs and the ConstDurNoRT model is used (duration = 1s), the relationship between the estimated activation, $\hat{\beta}$, and the true activation, B, is approximately $\hat{\beta} = B \times RT$, for a single trial (left panel, Figure 5), where the trend becomes nonlinear after approximately 2s. As Figure 5 lays out, this implies the relationship between condition difference estimates and RT differences is given by

$$\hat{\beta}_2 - \hat{\beta}_1 = B \times \left(\overline{RT}_2 - \overline{RT}_1\right). \tag{1}$$

In the example shown in Figure 5, even though the true condition difference is 0 for all subjects, the average estimated condition difference is nonzero across subjects and also has a linear relationship with the RT difference (right panel, Figure 5). As is the case with all linear trends, this relationship does not require a non-zero RT difference on average, but is driven by between-subject RT variability making it a potential confound whenever using ConsDurNoRT.

Simulation results in Figure 6 show group RT difference correlations across all models and data types. ConsDurNoRT and ConsDurRTMod produce correlations between contrast differences and RT differences when signal durations scale with RTs, as does RTDur when signal durations do not scale with RTs. ConsDurRTDur does not induce correlations for either signal type. Although no true relationship between average subject RT differences and fMRI condition differences exist, the mismatch between model assumptions and data potentially introduce a group-level model RT confound. We did not introduce a specific between-subject variance in the RT difference, implying the between-subject variance may be underestimated, but we did

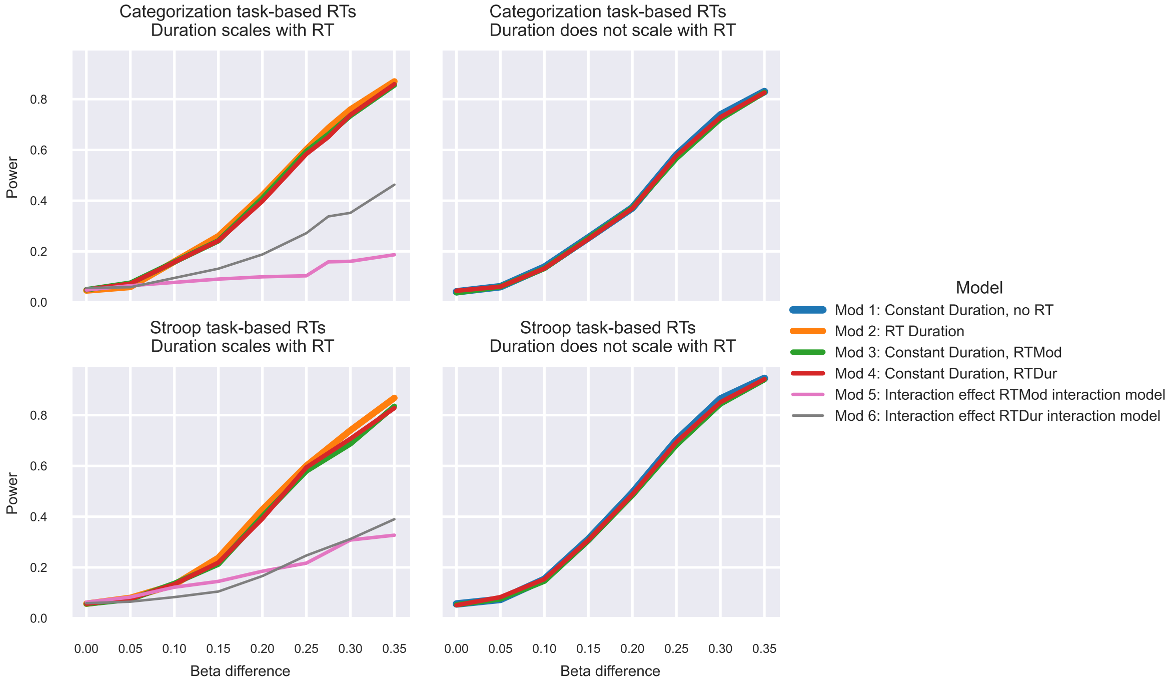


Fig. 4: Power when the RT difference is 0.1s as the condition difference increases. Only models that had controlled error rates in Figure 3 are shown. The two-alternative, forced-choice categorization task RT distribution was used in the top panels (Choice), while Stroop RT distribution was used in the bottom panels. An ISI between 2-4s was used and the inference of interest was for the 1-sample t-test of the condition difference effect with 100 subjects. Model 1, ConsDurNoRT, is the true model for the right column and model 2, RTDur, is the true model for the left column, so both indicate maximal power.

not want to artificially inflate the effect. Even so, it is within the ballpark of the expected true correlations between brain and behavior measures (Marek et al., 2022).

In the previous example, the linear relationship at the group level is, B, the common activation for both conditions and all subjects, but if activation differs between conditions or across subjects the confound will be more complex and can even introduce new artifactual associations into group analyses. To illustrate how false associations can be introduced, relax the assumption that B is the same across subjects, but preserve the assumption that B is the same for both conditions. For example, assume time spent outdoors (TO) is equally related to both conditions through the relationship,

$$B = \gamma_0 + \gamma_1 TO + \epsilon, \tag{2}$$

noting TO is not related to the true difference in activation between conditions or RTs. If the signal durations scale with RTs and ConsDurNoRT is used to estimate condition differences, both an artifactual RT difference effect and an artifactual RT difference by TO interaction are introduced to group-level data. This can be seen by combining equations 1 and 2:

$$\begin{aligned} \hat{\beta}_1 - \hat{\beta}_2 &= (\gamma_0 + \gamma_1 TO) \times \left(\overline{RT}_1 - \overline{RT}_2\right) \\ &= \gamma_0 \left(\overline{RT}_1 - \overline{RT}_2\right) + \gamma_1 TO \left(\overline{RT}_1 - \overline{RT}_2\right). \end{aligned} \tag{3}$$

Due to the interaction effect with TO and the nonlinear impact of RT on the BOLD signal when RTs>2s, adding RT difference as a confound regressor to the group model will not remedy these issues and should be avoided as it may inflate the significance of the false association with TO. The relationship between RT and potential variables of interest will be more complex if the RT difference is correlated with that variable or if the true activation difference and the variable of interest are correlated. This cannot be repaired within group-level analyses, but only by replacing first-level analyses with ConsDurRTDur. This is unsettling news

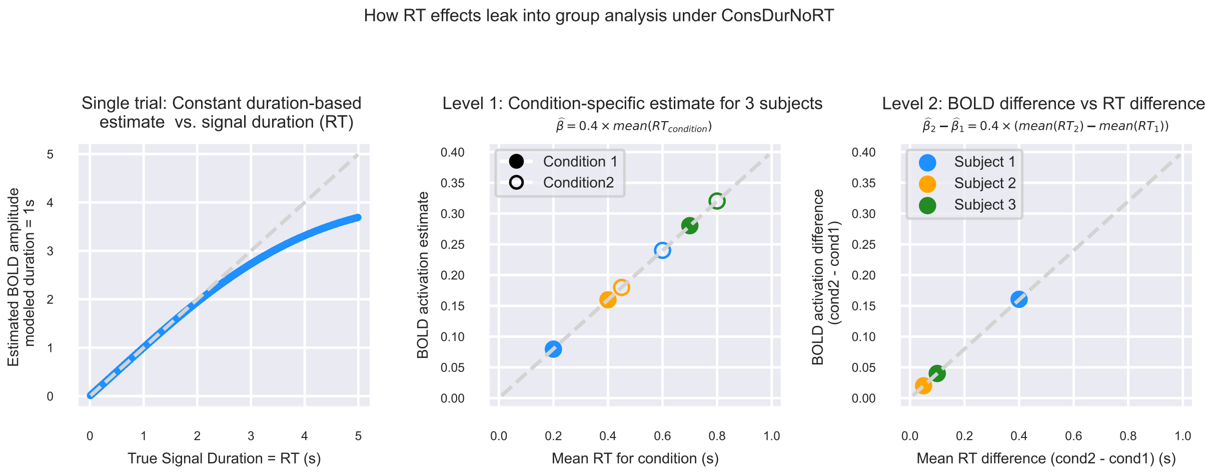


Fig. 5: The relationship between signal duration and activation estimates and how it can cause RT effects to leak into group analyses when using ConsDurNoRT. The left panel shows the relationship between the trial duration (RT) and the trial-specific BOLD activation estimate when a constant duration of 1s regressor is assumed and signal duration scales with RT (blue). The gray dashed line is a line with a slope of 1 and intercept of 0. The true BOLD activation is 1, but the model estimates the BOLD activation to be 1× RT for RTs < 2s. After 2s the linear relationship fails to hold, which explains when an RT modulated regressor will begin to fail as the amplitude is no longer a reasonable replacement for duration. The middle panel is an example of BOLD activation estimates for 3 subjects for 2 conditions when the true activation magnitude is $B = .4$ for both conditions (no condition effect difference) and the RTs are all less than 2s. The BOLD activation estimates are given by $\hat{\beta}_{cond} = .4 \times mean(RT_{cond})$ for each condition and subject. The right panel shows that even though the true condition difference is 0, the estimated condition differences are nonzero and vary according to $\hat{\beta}_2 - \hat{\beta}_1 = .4(mean(RT_2) - mean(RT_1))$, and so the contrast estimates are linearly related to the RT difference.

for those using fMRI activation databases, since the ConsDurNoRT model is typically used to generate activation estimates and, as the next section underlines, the presence of RT effects is widespread.

Widespread RT activation is not specific to task, revisited

Our real data analyses were modeled to included separate condition regressors and a single RT duration regressor, following ConsDurRTDur. A total of 7 tasks, with sample sizes ranging from 86 to 94, were analyzed. The cognitive processes involved in these tasks include attention, temporal discounting, proactive control, reactive control, response inhibition, resisting distraction, and set shifting. Brief descriptions are given in Supplemental Table 1 and more detailed summaries are provided in the Methods section ("Details about tasks involved in real data analysis") and Supplementary Materials. Comparatively, Yarkoni et al. (2009) used tasks including 3-back, decision making, emotion ratings and memory in sample sizes of 50, 102, 26, 35 and 39. Our seven tasks emphasize cognitive control to a greater extent and emotional processing and working memory to a lesser extent, compared to Yarkoni et al. (2009). The focus here is on average RT-related effects across subjects. This effect estimate will be slightly diminished from a full RT effect, since it is adjusted for condition difference. The interpretation is the average within-condition RT effect. Figure 7 shows voxels where the average RT-duration effects were significant across all 7 tasks. Our maps are consistent with Yarkoni et al. (2009), but with a more spatially widespread effects, which may reflect that our sample sizes were larger. In particular, the present comparison demonstrated substantially more signal in the lateral superior parietal cortex.

To illustrate how the RT-based network overlaps with and does not overlap with condition difference effects, Extended Data Fig. 4 illustrates group average results where only the RT effect was statistically significant (yellow), where only the condition comparison was significant (blue) and the overlap (green) for a single contrast from each task where ConsDurRTDur was used. Using the Stroop task as an example, the expected dMFC activation for incongruent versus congruent is present even though RT was included in

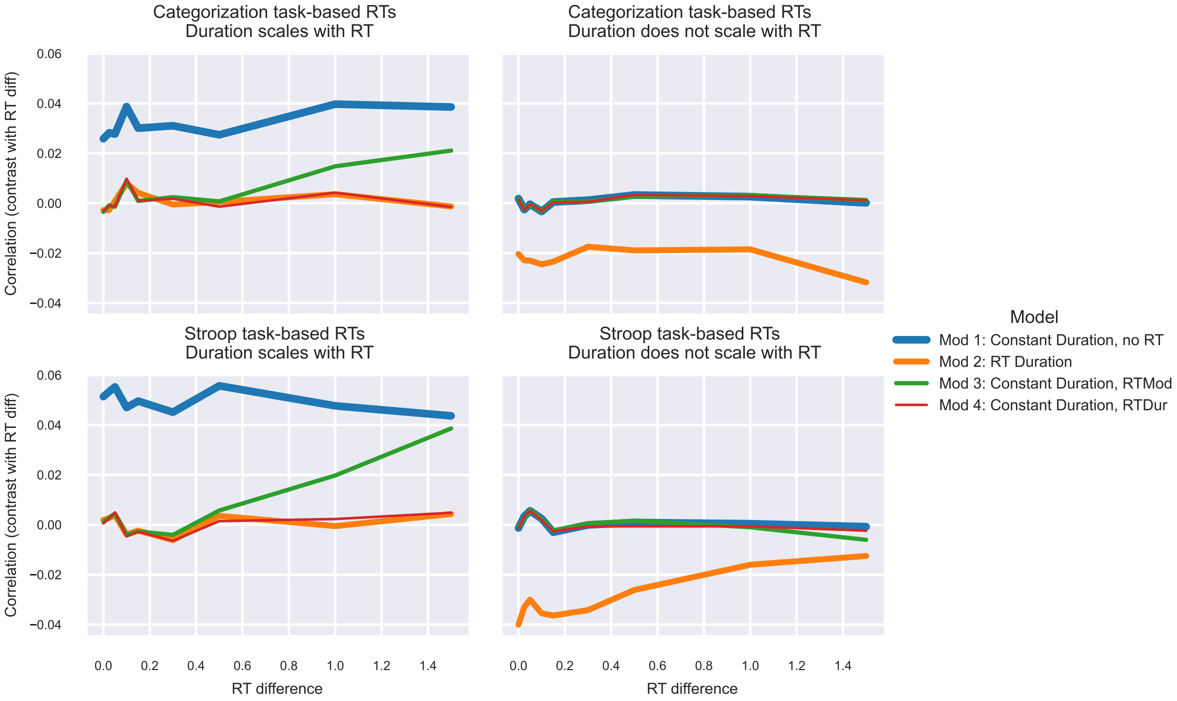


Fig. 6: Correlation between the contrast difference and difference in average condition RT across subjects as a function of the average difference in RT between conditions. Since the correlation is driven by between-subject variability in the *difference* in RT, there is no requirement that RTs differ between tasks and the correlation is constant regardless of the RT difference. In the case of ConsDurRTMod (green line, left panels) the correlation increases when the duration scales with RT for higher RT differences due to increasing model misfit as RTs increase (see Figure 5).

the time series model. Although there is often overlap in the activation for the RT effect and the condition comparison, there are regions where only the condition comparison is active.

Discussion

The problem of potential response time confounds for fMRI activation estimates has been discussed for more than a decade, with little resulting change in how the community approaches analyses and interpretations of fMRI contrasts. There are three takeaways from the present work. First, we propose a modeling approach that can adapt to fit data whether or not activation durations scale with RTs. Importantly this model does not remove the ability to also study RT-specific effects, if they are of interest. Second, this work highlights an important problem that has not been discussed previously: the presence of a between-subject confound of the average RT differences and the potential to introduce artificial associations with variables of interest at the group level. Finally, we replicate the work of Yarkoni et al. (2009) showing widespread RT-related effects that are not task specific.

This work presents a model that can adapt to whether or not signal durations scale with RTs, with limited performance loss. By adding an RT duration regressor to the most commonly used model that only contains condition-specific regressors, ConsDurRTDur removes RT-driven type I errors in average condition comparison effects with a slight reduction in power when the signal durations do not scale with RTs and the RT difference is large. The commonly used ConstDurNoRT model assumes signal durations do not scale with RTs and the RTDur model assumes signal durations must scale with RTs, and both models fail to control error rates when these model assumptions are violated (Figure 3). Interestingly, when RTDur is correct (interaction effect), only a slight power loss results when using ConsDurRTDur, while a 4 regressor interaction model has a large power loss due to collinearity in the model. This implies the ConsDurRTDur model can still be useful in detecting condition differences that arise through interactions that follow the

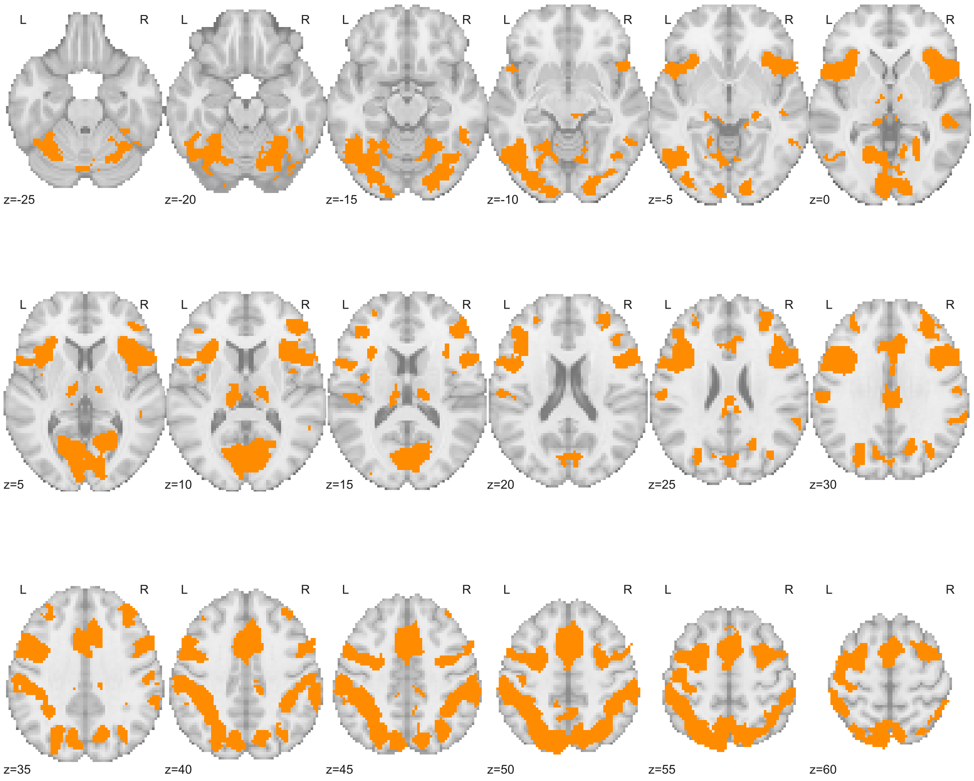


Fig. 7: Conjunction of the average, within-condition RT effect across ANT, DDT, DPX, motor selective stop, stop signal, stroop, and CTS. On average, each analysis included around 91 subjects and maps were corrected for multiple comparisons using a TFCE p-value thresholded at 0.05 using 5000 permutations.

RTDur model, although for regions with a hypothesized interaction, RTDur should be used to maximize power.

We have also uncovered that subject-specific differences in average RT represent an important group-level confound. This confound is only present when using ConsDurNoRT, RTDur or ConsDurRTMod when RT differences are large, whereas the ConsDurRTDur model produces condition difference effects that are free of this confound. Notably, the average difference in RT across subjects does not necessarily impact correlation strength, since correlations are driven by variability, in this case variability in RT differences across subjects. Thus, even if the average RT difference is 0, the first-level models should follow ConsDurRTDur. We have also shown that when signal durations scale with RTs and ConsDurNoRT is used, other false associations can be introduced at the group level. When there is a common association between a variable of interest (e.g., time spent outdoors) and each condition, separately, but no association of this variable with the true condition difference, the ConsDurNoRT condition difference estimates can have false associations with this variable. This is a concerning result for users of neuroimaging databases of condition difference estimates, since ConsDurNoRT is typically used.

Last, our updated RT-effect conjunction analysis across 7 tasks tapping into different mental processes show widespread shared activation in the so-called “task-positive” network, replicating the previous results of Yarkoni et al. (2009). This highlights the generality of the RT effect across tasks, and motivates the need to model these effects across all tasks.

There may be resistance to adding an RT duration regressor to convert ConsDurNoRT to ConsDurRTDur, since RT is the measure of interest in behavioral studies and removing RT effects from condition differences might be argued to be “throwing the baby out with the bathwater”. This argument is paradoxical because if RT is the effect of interest, why would it be ignored in the fMRI model? ConsDurNoRT can find significant results for any of the scenarios in Figure 8, and when the data follow “RT effect only” the results likely reflect false positives as shown in our Type I error simulation results. The ConsDurRTDur model is flexible, fitting signal when durations do or do not scale with RT. ConsDurRTDur can fit data that follow any of the scenarios in Figure 8, without resulting in false positives for the “RT effect only” scenario and with a slight power loss only when there is an interaction between condition and RTs. The flexibility of ConsDurRTDur means it does not align with a specific behavioral model. If a specific link between the underlying cognitive theory and the fMRI model is of interest, this requires motivating why that theory would hold and for which specific regions of interest. Then, for those regions of interest the fMRI model that aligns with the underlying theoretical model should be used. For example, if theory supports that a region would elicit brain activation where neuronal duration scales with RT, then RTDur will have the highest power and should be used for that region rather than ConsDurRTDur. That said, RTDur is not likely appropriate in a whole brain analysis since it is not designed for regions where the signal duration does not scale with RTs as stated in Grinband et al. (2008) and illustrated by elevated Type I errors in our simulations.

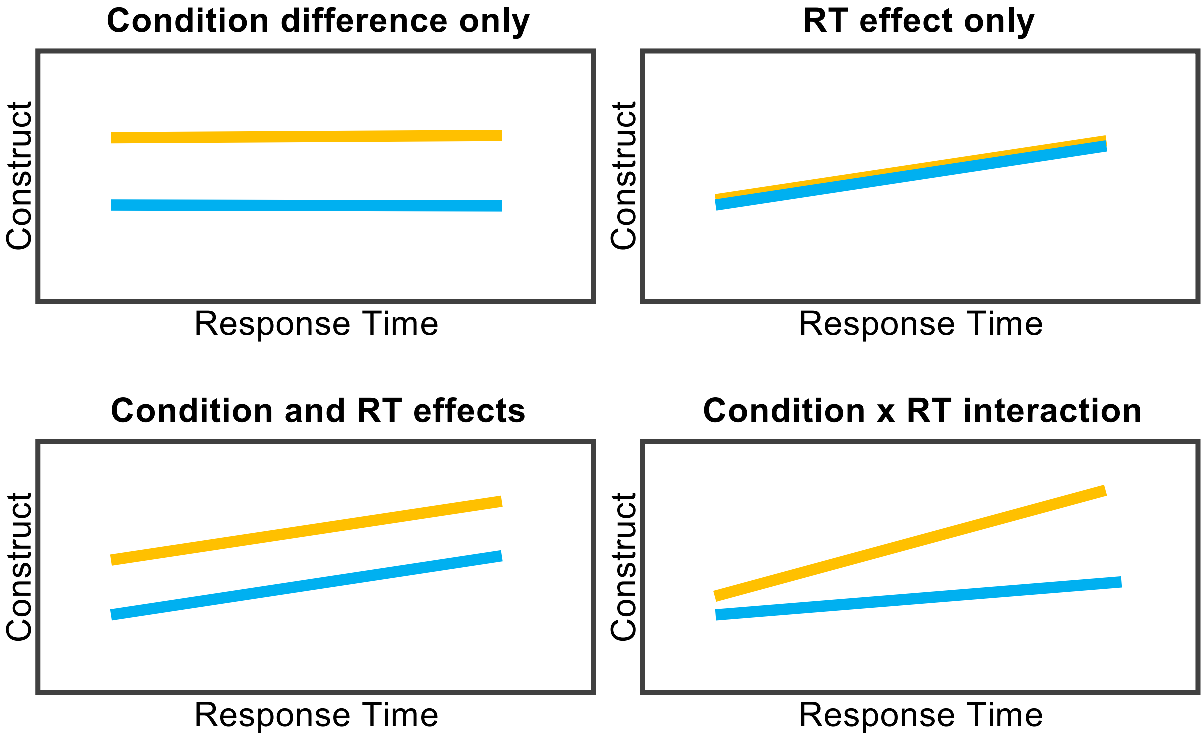


Fig. 8: Examples of the underlying relationship between a behavioral construct and RTs.

Both ConsDurRTMod and ConsDurRTDur assume that mean task activation differs for each condition while the BOLD/RT relationship is the same for both conditions, analogous to a regression model that models two parallel lines corresponding to two conditions. Condition differences, when both conditions involve RTs, will be the same regardless of the value of RT. Conversely, a contrast of a condition involving RTs vs baseline, or a comparisons of conditions where only some of the conditions involve RTs, will vary by RT. Any estimate of these contrasts will reference a specific RT. For example, the contrast interpretation for conditions involving RTs versus baseline correspond to the activation magnitude for a trial with an RT of 0, which is illustrated in the middle panel of Extended Data Fig. 3 where the condition 1 (orange) has a negative task versus baseline effect while the condition 2 (blue line) has a small positive effect in the ConsDurRTDur model. Extended Data Table 1 illustrates examples of contrast interpretations using the

Stroop task. It may be tempting to use orthogonalization so the interpretations do not correspond to an RT of 0, but the following section outlines the limitations of this idea.

A common misunderstanding is that collinearity between regressors is problematic and orthogonalization should be used to remove collinearity or to change parameter interpretation. Orthogonalization is almost never necessary and is often applied incorrectly, leading to confusion in interpretation (Mumford et al., 2015). In the case of ConsDurRTDur and ConsDurRTMod, orthogonalization of the RT regressor can change the interpretations of contrasts so they do not correspond to an RT of 0. Extended Data Table 2 describes various orthogonalization strategies for ConsDurRTDur and ConsDurRTMod and explains how contrast interpretation is impacted and whether orthogonalization is acceptable. For ConsDurRTDur, orthogonalization of the RT regressor with respect to each condition (row 1) is problematic because condition differences are no longer adjusted for RTs. Rows 2 and 3 show how ConsDurRTDur and ConsDurRTMod can orthogonalize the RT regressor with respect to all RT-based trials, but this introduces a between-subject RT confound, which is not acceptable. The last row describes a way to center RTs in ConsDurRTMod that avoids the between-subject RT confound, where all contrasts that previously corresponded to an RT of 0 now correspond to some specific RT, C. This practice is questionable since, if the RT relationship is positive, selecting a larger value for C will inflate the contrast estimate.

Importantly, issues with the contrast interpretations presented in Extended Data Table 1 only arise in analyses of single group means while group comparisons and associations with other variables are not impacted by the reference RT value for a contrast, as long as it is constant across subjects.

The present work provides a modeling framework for moving forward, but it is likely that as researchers start thinking about response times more carefully, the model will need to be further adapted. For example, it is challenging to simply apply the ConsDurRTDur model to the stop signal task since the overt response process is unobservable on some trials (i.e. successful stop trials), and the stop process is generally estimated at a block- or session-level so is unobservable on all individual trials. It is unclear how the absence of this time in the models impacts results. Although it is beyond the scope of this work to find a solution to this problem, this is a future direction we will explore.

This work consists of real data analyses as well as simulated data analyses. Simulations are required in cases where we need to know the ground truth and link the theoretical problems with how these problems might surface in real data analyses (e.g., how strong the results are and whether they persist at the group level). As such, the simulations require specifying a large number of parameters including the RT distribution for each condition, effect size for each condition, stimulus length, ISI, within-subject variance and between-subject variance. Our simulation parameter choices are described in the Methods section and we feel generated realistic simulated data as the results are consistent with similar studies (Yarkoni et al., 2009; Brown, 2011; Grinband et al., 2011).

Methods

This fMRI study described below ("Real Data Analysis") was approved by the Stanford University Institutional Review Board (approval number: 39322) and complies with all ethical regulations.

Models considered

Data generation and modeling

The interstimulus interval (ISI) was sampled from a Uniform distribution and RT was sampled from an ex-Gaussian distribution. For RT, a subject specific mean, m_s, was obtained by sampling an ex-Gaussian with parameters μ_{rt}, σ_{rt} and $1/\lambda_{rt}$. The mean of an ex-Gaussian distribution is the sum of μ and $1/\lambda$ and in our Stroop data the within-subject estimates of μ were approximately 76% of the mean. Using this information the μ and $1/\lambda$ parameters for fast RTs that differed by ΔRT were defined by $\mu_{fast} = .76m_s - .76\Delta RT/2$ and $1/\lambda_{fast} = .24m_s - .24\Delta RT/2$, respectively. The slow RTs were obtained similarly but used addition. This preserved the desired mean RT difference and also allowed the variance to change with the mean RT, since λ contributes to the variance as well. Values of μ_{rt}, σ_{rt} and λ_{rt} were based on our Stroop data and the two-alternative, forced-choice categorization task in Grinband et al. (2008). In both cases distributions were fit to subject-specific data and then parameters were averaged over subjects. The two-alternative, forced-choice categorization task RT distribution was defined by a Gamma distribution with shape parameter = 1.7, beta = 0.49. Sampling from this distribution and fitting an ex-Gaussian to that sample resulted in ex-Gaussian parameters of $\mu_{rt} = 638$, $\sigma_{rt} = 103$, and $1/\lambda_{rt} = 699$ (mean = 1337, sd = 706.5). The Stroop data had faster RTs with less variability, with ex-Gaussian parameters of $\mu_{rt} = 530$, $\sigma_{rt} = 77$, and $1/\lambda_{rt} = 160$ (mean = 690, sd = 177.5). The distribution functions from the Scipy module (v 1.9.1) of Python (v 3.9.7) were used to simulate and estimate the distribution parameters. Trial order was random.

Simulated data where signal duration scaled with RT were created with the convolved RT duration regressors (RTDur) and data where signal duration did not scale with RT used the constant duration regressors (ConsDurNoRT). The BOLD activation sizes, $\beta_{i,j}$, for the i^{th} subject for j^{th} condition (j = 1,2) were sampled from a Gaussian distribution, $N(\beta_j, \sigma_b^2)$, where β_j is the true activation magnitude and σ_b^2 is the between-subject variance. The time-series data for the i^{th} subject, of length T, was created according to

$$\mathbf{Y_i} = \mathbf{X}_1\beta_{i,1} + \mathbf{X}_2\beta_{i,2} + \epsilon, \quad \epsilon \sim N(0, \sigma_w^2), \tag{4}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are either the Model 1 or 2 regressors ($T \times 1$) and σ_w^2 is the within-subject variance.

In an effort to choose realistic values for β_1, β_2, σ_w^2 and σ_b^2, we considered the first-level effect size (converting the true β_i to a correlation), second level effect size for a 1-sample t-test (Cohen's D) as well as the ratio of the total mixed effects variance to the within-subject variance. Following the definitions of parameters as given in the model above, the total mixed effects variance for a first-level contrast of parameter estimates is

$$\sigma_{mfx}^2 = \mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'\sigma_w^2 + \mathbf{cc}'\sigma_b^2, \tag{5}$$

where $\mathbf{X}$ and $\mathbf{c}$ are the first-level design matrix (based on models in Figure 2) and contrast of interest (Mumford and Nichols, 2006). The contrast of interest for each model corresponded to condition 2 > condition 1 ($\mathbf{c} = [-1, 1]$ for the 2 regressor models and $\mathbf{c} = [-1, 1, 0]$ for the three regressor model). The ratio of total standard deviation (SD) to within-subject SD is defined by

$$\frac{SD_{total}}{SD_{within}} = \frac{\sqrt{\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'\sigma_w^2 + cc'\sigma_b^2}}{\sqrt{\mathbf{c}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}'\sigma_w^2}}, \tag{6}$$

Our within-subject effect size for condition versus baseline was between 0.07-0.08 (correlation), ratio of total variance to within-subject variance, $\frac{SD_{total}}{SD_{within}}$, ranged between 2-3 and the Cohen's D for the average of task versus baseline across subjects was approximately 0.85.

Each run contained 40 trials of each condition and a time resolution (TR) of 1s. Time course length varied, as it was set to extend 50s past the last stimulus offset. Group analyses included 100 subjects. A total of 1000 data sets were simulated to calculate power and error rates.

Regressors were constructed by convolving boxcar functions with a Double Gamma hemodynamic response function (HRF) using the `spm` HRF within the `compute_regressor` function from the Nilearn (v 0.9.1) module in Python (v 3.9.7).

Least squares regression was used to estimate the models described in Figure 2 at the first-level including a set of cosine basis functions (0.1 Hz cutoff) for high-pass filtering generated with the `cosine_drift` function from Nilearn (v 0.9.1) in Python (v 3.9.7). At the group level, 1-sample t-tests were used to assess type I error and power. A correlation of the average difference in RT between conditions and the fMRI contrast (condition 2 vs condition 1) was estimated for each group analysis.

Since RT and ISI values are random, the contribution of the design matrix, X, to the overall variance varies between samples (Equation 5) and the true effect size was variable. Therefore, to calculate the first-level true effect size 100 data sets were simulated and the partial correlation coefficient for one condition, controlling for the other condition and cosine basis set, was estimated and then averaged over the 100 data sets to serve as the true within-subject effect. The variance ratio, SD_{total}/SD_{within} , was estimated by simulating 100 design matrices. Cohen's D estimates were based on 5000 simulated within-subject model estimates for the task versus baseline contrast.

Real data analysis

Informed consent was obtained from all human participants and participants were paid \$20 per hour for their participation in the MRI sessions and \$10 per hour for participation in practice and setup time. None of the analyses were preregistered. No statistical methods were used to determine sample sizes, although our sample is about twice as large as the data sets involved in a similar analysis in Yarkoni et al. (2009).

Prospective participants for the study were recruited from the Stanford campus and surrounding San Francisco Bay Area using several methods including paper flyers, the Stanford Sona recruitment system, local newspapers ads, the Poldrack Lab website, online resources such as Craigslist, and through email listservs maintained by the Stanford Psychology Department. All recruited participants met the following criteria: have a minimum 8th grade education, speak English fluently, right-handed, have normal or corrected to normal vision and no color-blindness, are between 18-40 years old, have no current diabetes diagnosis, have no history of head trauma with loss of consciousness, cerebrovascular accident, seizures, neurosurgical intervention, stroke, or brain tumor, have no current major psychiatric disorders (including schizophrenia and bipolar disorder) or substance dependence, are not currently using any medication for psychiatric reasons, are not currently pregnant, and have no other contraindications to MRI.

A total of 113 participants were recruited for the study. 3 participants were dropped during their first scan session due to complications in the scanner leaving 110 participants. The mean age was 23.8 years (sd = 5.5) and 71 participants were female. The sample had the following demographic distribution: 40% White, 36% Asian, 10% More than one race, 8% Black or African American, 3% Unknown, 2% Native Hawaiian or Pacific Islander, and 1% American Indian or Alaska Native. The fMRI tasks included: Stroop (Stroop, 1935), Attention Network Test (ANT, Fan et al. (2002)), Dot Pattern Expectancy task (DPX, MacDonald et al. (2005)), Delayed-Discounting task (DDT, Kirby (2009)), cued task-switching task (CTS, Logan and Bundesen (2003)), stop signal task (Logan and Cowan, 1984) and a motor selective stop signal task (DeJong et al., 1995). Brief summaries are provided in Supplemental Table 1 and more detailed descriptions are provided in the Supplementary materials ("Details about tasks involved in real data analysis"). Data were acquired using single-echo multi-band EPI using a GE Discovery MR750 3T scanner and a Nova medical 32-channel head coil. The following parameters were used for data acquisition: TR = 680ms, multiband factor = 8, echo time = 30 ms, flip angle = 53 degrees, field of view = 220 mm , 2.2 × 2.2 × 2.2 isotropic voxels with 64 slices.

Raw task behavior during scans were collected using Experiment Factory (https://www.expfactory.org/; Sochat et al. (2016)) running on Macbook laptops. The raw data can be found at https://openneuro.org/datasets/ds004636/ and a Data Descriptor paper is available at Bissett et al. (2023).

Data were preprocessed in Python using fmriprep 20.2.0 (Esteban et al., 2019). First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. A B0-nonuniformity map (or fieldmap) was directly measured with an MRI scheme designed with that purpose (typically, a spiral pulse sequence). The fieldmap was then co-registered to the target EPI (echo-planar imaging) reference run and converted to a displacements field map (amenable to registration tools such as ANTs) with FSL's fugue and other SDCflows tools. Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. (2002)). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and S (1997), `RRID:SCR_005927`). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as preprocessed BOLD in original space, or just preprocessed BOLD. The BOLD time-series were resampled into standard space, generating a preprocessed BOLD run in MNI152NLin2009cAsym space. First, a reference volume and its skull-stripped version were generated using a custom methodology of fMRIPrep. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al. (2015)) was performed on the preprocessed BOLD on MNI space time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). Corresponding "non-aggresively" denoised runs were produced after such smoothing. These data were used in our time-series analysis models.

Data were analyzed using `FirstLevelModel` from nilearn in Python (version 3.9.7) (Abraham et al., 2014). A double gamma HRF was used for convolution and an AR(1) model addressed temporal auto correlation. Specifically the `spm` HRF setting was used, which follows the HRF from the SPM software package with a 6s delay of response (relative to onset), a 16s delay of undershoot (relative to onset), 1s dispersion of response, 1s dispersion of undershoot, a ratio of response to undershoot of 6 and a 32s long kernel. Regressors were included for each condition, versus baseline, as well as a single RT modulated regressor, similar to the simulation analysis model ConsDurRT. The RT modulated regressor included the uncentered RT values. The contrast of the RT modulated regressor was the contrast of interest in our models and represents the average relationship between BOLD activation and RT within condition, since condition specific regressors were also included. Nuisance regressors in the time-series analysis included the following from the fmriprep output: cosine basis functions (corresponding to a highpass filter cutoff of 128s) and the average time courses for the CSF and WM as estimated by fmriprep.

Subjects were excluded within-task for the following general reasons: missing 1 or more files required to analyze the data, having more than 20% high motion time points (measured by Framewise Displacement > .5 or SD of DVARS > 1.2), having more than 45% missing responses, a subjective poor performance rating assessing high choice and/or omission error rates in at least one condition of the task, and when subjects omitted most of their responses towards the end of the task scan. Specific exclusion for the stop signal tasks are less than 25% successes for stop trials or more than a 75% successful stop rate. For the Delay-Discounting tasks, subjects were excluded if they made the same choice on all trials. Last, if there were exclusions on more than half the tasks for a subject, that subject was completely excluded. Exclusions for each task are reported in the supplemental materials ("Exclusion information by task for real data analysis").

Group models were estimated using Randomise (Smith and Nichols, 2009) and included either a single column of 1s (group mean) or a column of 1s along with the difference in mean RTs. Statistics maps were thresholded, controlling for family-wise error rate, using the Randomise TFCE statistic below 0.05, based on 5000 permutations. Two sided hypotheses were studied using an F-contrast. A conjunction map was constructed by taking the overlap of the thresholded, binarized map for each of the 7 tasks (Nichols et al., 2005).

Data availability

The conjunction map (Figure 7) and separate RT-based activation maps for each task used to generate the conjunction map are available in the form of 1-pvalue maps at https://neurovault.org/collections/13656/. Full, raw fMRI data sets are available on OpenNeuro at https://openneuro.org/datasets/ds004636/.

Code availability

The code for all analyses are shared at https://doi.org/10.5281/zenodo.8083510 (simulations) and https://doi.org/10.5281/zenodo.8083518 (real data analysis).

Acknowledgements

This work was supported by the National Institutes of Health (NIH) Science of Behavior Change Common Fund Program through an award administered by the National Institute for Drug Abuse (NIDA) (UH2DA041713; PIs: Marsch, LA & Poldrack, RA). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

We would also like to thank Daniel Weissman for discussions that helped in the writing of this manuscript.

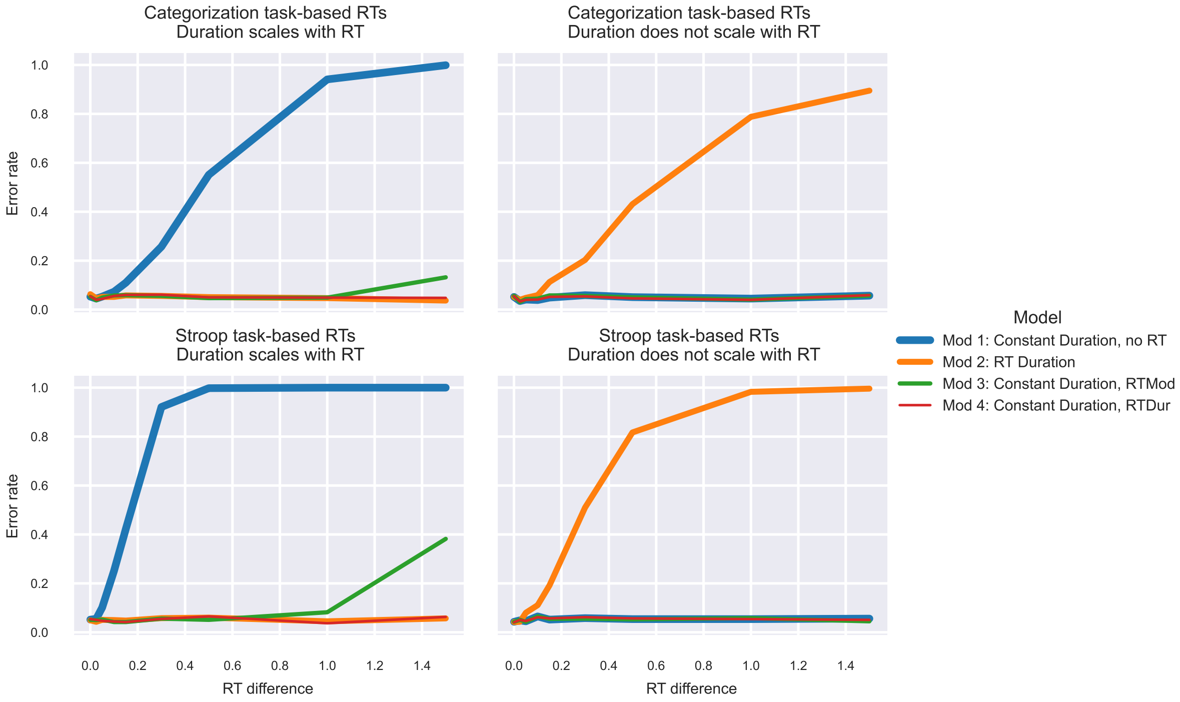
Author contributions

Jeanette Mumford developed the project, wrote all simulation code and much of the real data analysis code, helped with quality assessments of the real data and did most of the writing. Patrick Bissett helped with model construction, interpretation, data collection and writing of the manuscript. Henry Jones, Sunjae Shim and Jaime Rios all helped in the collection, preprocessing, quality assessment and analysis of the real fMRI data. Russell Poldrack helped with model development, interpretation and writing of the manuscript.

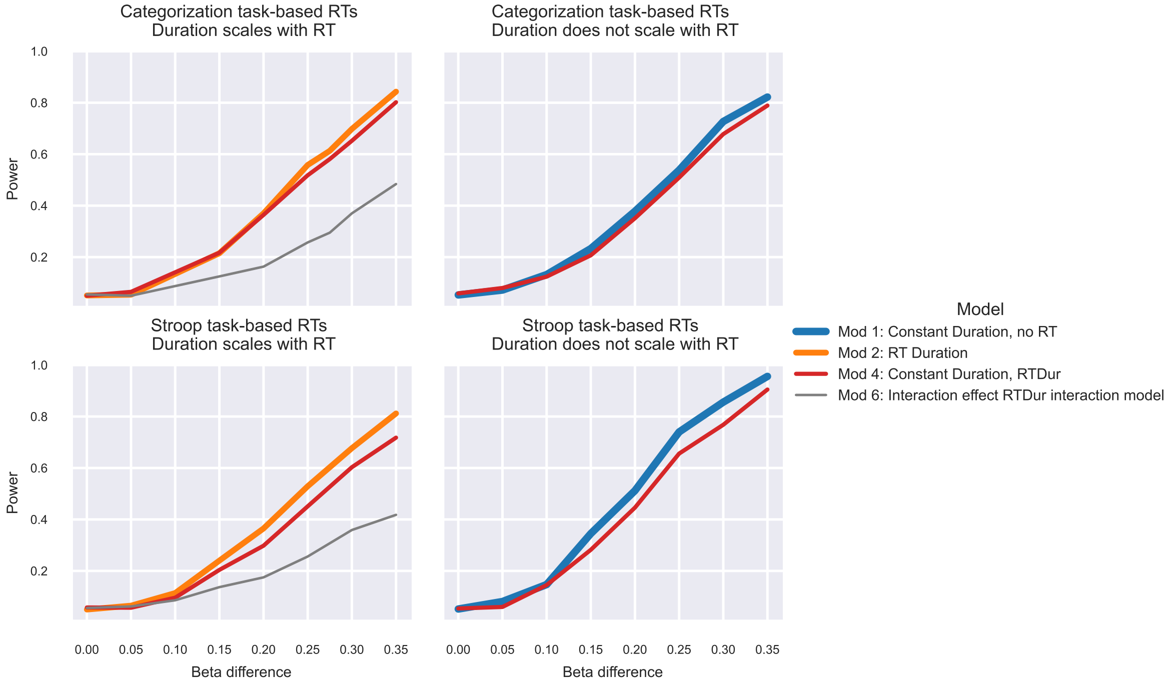
Ethics declarations

The authors declare no competing interests.

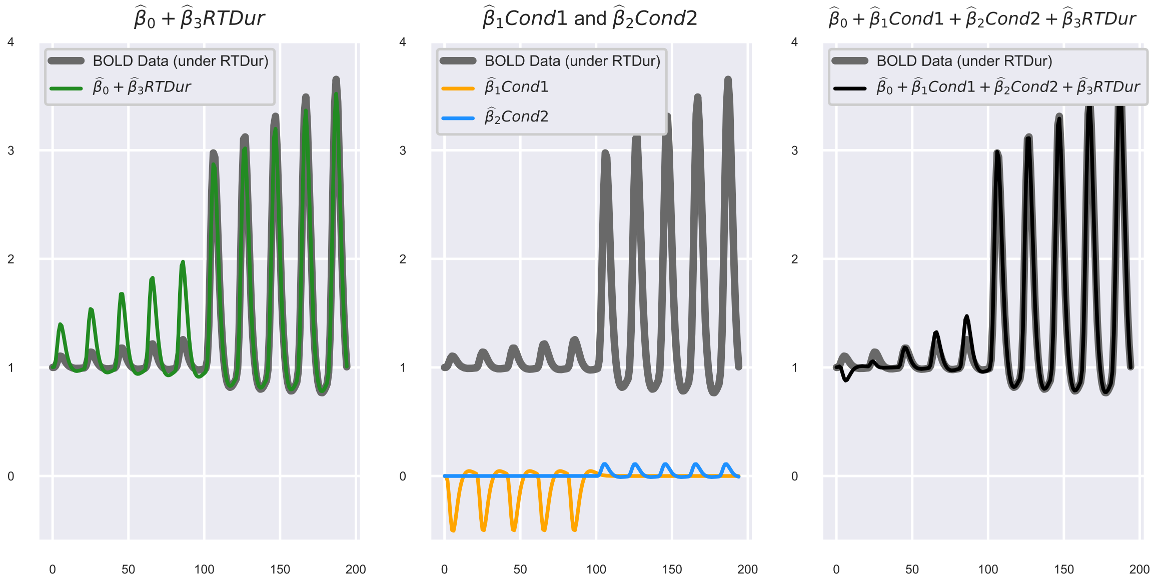
Extended Data



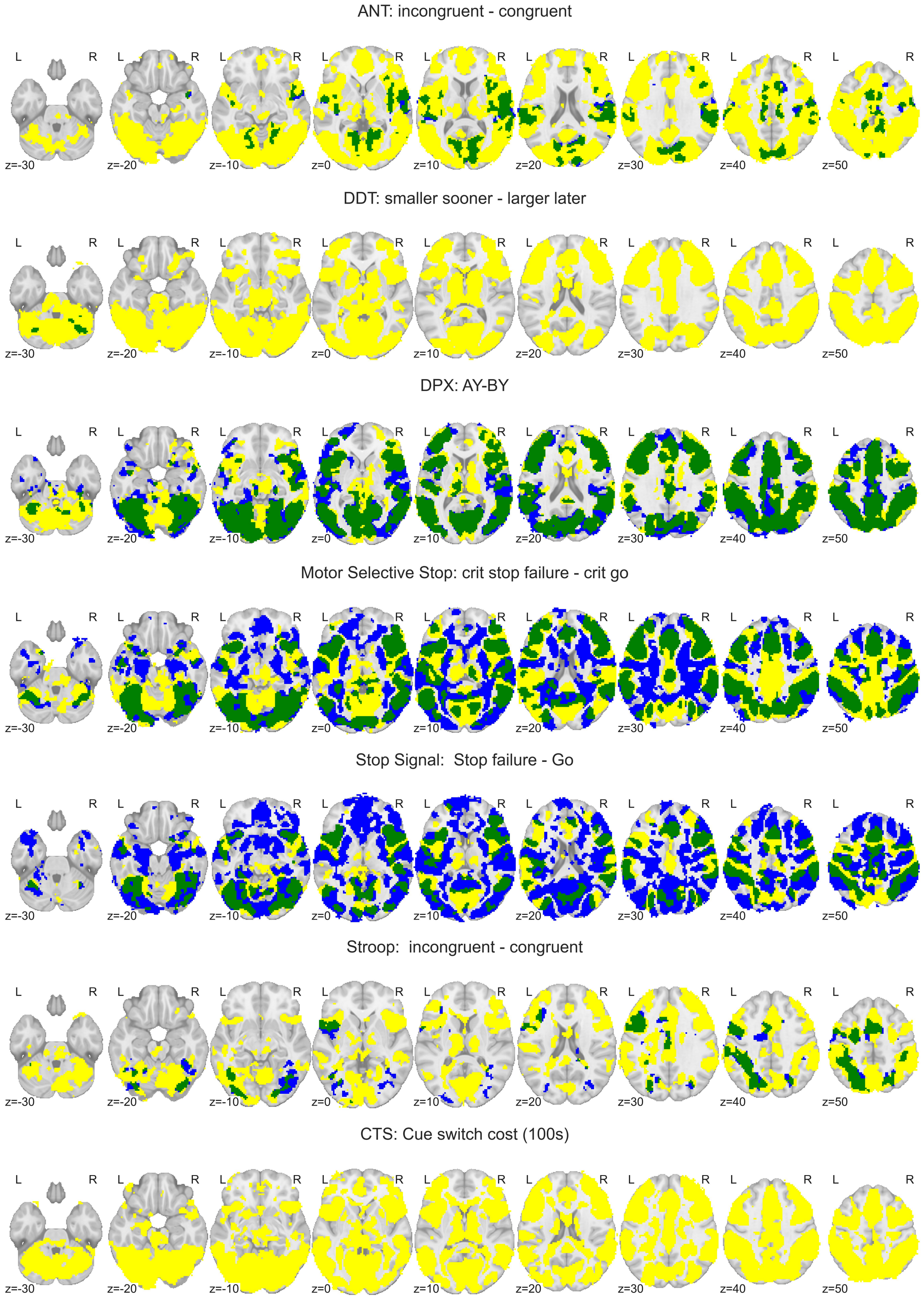
Extended Data Fig. 1: **Type I errors as RT difference between conditions increases (ISI=3-6s).** This figure illustrates that results are similar to when the ISI ranged between 2-4s (results in main manuscript, Fig. 3). The two-alternative, force-choice categorization task (Categorization) RT distribution was used in the top panels, while Stroop RT distribution was used in the bottom panels, both with an ISI between 3-6s and the inference of interest was for the 1-sample t-test of the condition difference effect with 100 subjects. 2500 simulations were used to calculate the error rate.



Extended Data Fig. 2: **Power when the RT difference is 0.8s as the condition difference increases.** Only models that had controlled error rates in Fig. 3 (main text) are shown and since the RT range of this simulation is not within the range where ConsDurRTMod has controlled error rates, it is excluded. The two-alternative, forced-choice categorization task RT distribution was used in the top panels, while Stroop RT distribution was used in the bottom panels. An ISI between 2-4s was used and the inference of interest was for the 1-sample t-test of the condition difference effect with 100 subjects. Model 2, RTDur, is the true model for the left column (orange line) and model 1, ConsDurNoRT, is the true model for the right column (blue line), and the corresponding power curves indicate maximal power. When the signal duration scales with RT (left panel), ConsDurRTDur (red) has some power loss due to model misfit, while the 4 regressor interaction model (gray line) loses considerable power due to collinearity. When the duration does not scale with RT (right panels), the ConsDurRTDur model (red line) has similar power to the true model (blue), illustrating some power loss due to including an RT regressor in the time-series analysis. This power loss is not seen for shorter RT differences (.1s result in main paper Fig. 4).



Extended Data Fig. 3: **Understanding ConsDurRTDur parameter estimates when RTDur model is true.** When the data follow the RTDur model, this implies an RT by condition interaction where the condition effects are 0 when the RT is 0. If using the ConsDurRTDur model to fit the data, the difference between the fitted constant duration regressors is approximately the RTDur difference, albeit with a loss in power, which is illustrated here. The figure illustrates the ConsDurRTDur model fit, $\beta_0 + \beta_1\text{ConsDur1} + \beta_2\text{ConsDur2} + \beta_3\text{RTDur}$, to data based on RTDur model, $BOLD = 1 + 1 * \text{RTDur1} + 4 * \text{RTDur2}$, where the RTs are evenly spaced between .5 and 1.2 for condition 1 and 2.5 and 3.5 for condition 2. The first 5 trials are condition 1 and second 5 trials are condition 2. The first two panels break down components of the ConsDurRTDur model fit to RTDur data, where the first panel is the intercept plus the RTDur effect ($\hat{\beta}_0 + \hat{\beta}_3\text{RTDur}$, green), which overestimates the BOLD for the first condition and underestimates it for the second condition. The middle panel shows how the condition 1 effect from ConsDurRTDur accounts for the overestimation in the first panel and the condition 2 effect accounts for the underestimation. Summing the green, orange and blue lines of the first two panels yields the black line in the third panel, showing similar model fits, which implies the condition difference effect from ConsDurRTDur (peak differences between yellow and blue lines of the middle panel) is a close approximation to the true effect from RTDur. This figure also illustrates that the condition versus baseline effects, alone, from ConsDurRTDur have limited interpretation as they reflect the projected effect when the RT is 0 and can be negative, so they have limited use.



ANT: incongruent - congruent

DDT: smaller sooner - larger later

DPX: AY-BY

Motor Selective Stop: crit stop failure - crit go

Stop Signal: Stop failure - Go

Stroop: incongruent - congruent

CTS: Cue switch cost (100s)

Extended Data Fig. 4: **RT-based activation, RT-adjusted condition comparisons and their overlap.** Comparison of RT-based activation network (yellow), specific condition comparison contrast activation (blue) and their overlap (green). The condition comparison contrast for a given task is described in the panel title. In all cases the ConsDurRTDur model was used and the networks are based off of a randomise TFCE family-wise error corrected $p < 0.05$ for a 2-sided one-sample t-test at the group level.

Contrast type	Example Stroop task)	Interpretation in ConsDurRTDur/ConsDurRTMod
Condition (RTs) vs. baseline	Incongruent vs. baseline	Incongruent activation when RT is 0
Condition (RTs) vs. Condition (RTs)	Incongruent vs. Congruent	Condition difference (same for all RTs)

Extended Data Table 1: **Contrast interpretation examples using the Stroop task.** Contrast interpretation only refers to trials with a specific RT value for condition versus baseline comparisons (or condition comparisons where only some conditions involve RTs) as described by example contrasts in this table.

Model	Orthogonalization procedure	Implication of orthogonalization for contrasts in Extended Data Table 1	Acceptable?
ConsDurRTDur	Replace RTDur with residual from: RTDur $\sim$ Cond1 + Cond2	RT adjustment has been neutralized, rendering the contrast estimates to be equivalent to ConsDurNoRT.	No
ConsDurRTDur	Replace RTDur with residual from: RTDur $\sim$ all_trials	Contrasts that correspond to RT of 0 now correspond to the mean RT of the run. Between-subject RT confound has been introduced.	No
ConsDurRTMod (all RTs<2s)	Center RTs by mean(RT), across run	Contrasts that previously corresponded to an RT of 0 now correspond to the mean RT of the run. Between-subject RT confound has been introduced.	No
ConsDurRTMod (all RTs<2s)	Center RTs by the same constant, C, in all runs (any C within range of RTs is fine)	Contrasts that previously corresponded to an RT of 0 now correspond to an RT of C	Maybe

Extended Data Table 2: **Impact of orthogonalization on contrast interpretation.** Orthogonalization examples that describe how orthogonalization may be carried out and the implication on the interpretation of the original model's contrast estimates (Extended Data Table 1). Note that "cond1" and "cond2" are constant duration regressors for each condition and "all_trials" is a constant duration regressor including all trials with RTs.

References

Donders, F.C.: Over de snelheid van psychische processen [on the speed of psychological processes]. Acta Psychologica **30**, 412–431 (1969)

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

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D.: Conflict monitoring and congnitive control. Pschological Review **108**(3), 624–652 (2001)

Botvinick, M., Nystrom, L., Fissell, K., Carter, C., Cohen, J.: Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature **402**(6758), 179–181 (1999)

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

Kerns, J.G., Cohen, J.D., MacDonald, A.W., Cho, R.Y., Stenger, V.A., Carter, C.S.: Anterior Cingulate Conflict Monitoring and Adjustments in Control. Science **303**(5660), 1023–1026 (2004) https://doi.org/10.1126/science.1089910.

Logothetis, N.K.: What we can do and what we cannot do with fMRI. Nature **453**(7197), 869–878 (2008) https://doi.org/10.1038/nature06976

Savoy, R., Bandettini, P., Weisskoff, R., Kwong, K., Davis, T., Baker, J.: Pushing the temporal resolution of fmri: studies of very brief visual stimuli, onset variablity and asynchrony, and stimulus-correlated changes in noise. Proceedings SMR Third Annual Meeting, Nice, 450 (1995)

Jezzard, P., Matthews, P., Smith, S.: Functional MRI, an Introduction to Methods. Oxford University Press Inc., New York (2001)

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., Hirsch, J.: Detection of time-varying signals in event-related fMRI designs. NeuroImage **43**(3), 509–520 (2008) https://doi.org/10.1016/j.neuroimage.2008.07.065 .

Ratcliff, R.: A Theory of Memory Retrieval. American Psychological Association **85**(2) (1978)

Yarkoni, T., Barch, D.M., Gray, J.R., Conturo, T.E., Braver, T.S.: BOLD Correlates of Trial-by-Trial Reaction Time Variability in Gray and White Matter: A Multi-Study fMRI Analysis. PLoS ONE **4**(1), 4257 (2009) https://doi.org/10.1371/journal.pone.0004257 .

Grinband, J., Savitskaya, J., Wager, T.D., Teichert, T., Ferrera, V.P., Hirsch, J.: The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. NeuroImage **57**(2), 303–311 (2011) https://doi.org/10.1016/j.neuroimage.2010.12.027 .

Carp, J., Kim, K., Taylor, S., Diamond-Fitzgerald, K., Weissman, D.: Conditional differences in mean reaction time explain effects of response congruency, but not accuracy, on posterior medial frontal cortex activity. Frontiers in Human Neuroscience (2010) https://doi.org/10.3389/fnhum.2010.00231 .

Yeung, N., Cohen, J.D., Botvinick, M.M.: Errors of interpretation and modeling: A reply to Grinband et al. NeuroImage **57**(2), 316–319 (2011) https://doi.org/10.1016/j.neuroimage.2011.04.029 .

Brown, J.W.: Medial prefrontal cortex activity correlates with time-on-task: What does this tell us about theories of cognitive control? NeuroImage **57**(2), 314–315 (2011) https://doi.org/10.1016/j.neuroimage.2011.04.028 .

Grinband, J., Savitskaya, J., Wager, T.D., Teichert, T., Ferrera, V.P., Hirsch, J.: Conflict, error likelihood, and RT: Response to Brown & Yeung et al. NeuroImage **57**(2), 320–322 (2011) https://doi.org/10.1016/j.neuroimage.2011.04.027 .

Nachev, P.: The blind executive. NeuroImage **57**(2), 312–313 (2011) https://doi.org/10.1016/j.neuroimage.2011.04.025 .

Carp, J., Fitzgerald, K.D., Taylor, S.F., Weissman, D.H.: Removing the effect of response time on brain activity reveals developmental differences in conflict processing in the posterior medial prefrontal cortex. NeuroImage **59**(1), 853–860 (2012) https://doi.org/10.1016/j.neuroimage.2011.07.064 .

Poldrack, R.A., Mumford, J.A., Nichols, T.E.: Handbook of Functional MRI Data Analysis. Cambridge University Press, New York (2009)

Ratcliff, R., Murdock, B.B.: Retrieval processes in recognition memory. Psychological Review **83**, 190–214 (1976)

Weissman, D.H., Carp, J.: The Congruency Effect in the Posterior Medial Frontal Cortex Is More Consistent with Time on Task than with Response Conflict. PLoS ONE **8**(4), 62405 (2013) https://doi.org/10.1371/journal.pone.0062405

Dubois, J., Adolphs, R.: Building a Science of Individual Differences from fMRI. Trends in Cognitive Sciences **20**(6), 425–443 (2016) https://doi.org/10.1016/j.tics.2016.03.014

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., Dosenbach, N.U.F.: Reproducible brain-wide association studies require thousands of individuals. Nature **603**, 654–660 (2022) https://doi.org/10.1038/s41586-022-04492-9

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

Mumford, J.A., Nichols, T.: Modeling and inference of multisubject fMRI data. IEEE Engineering in Medicine and Biology Magazine **25**(2), 42–51 (2006) https://doi.org/10.1109/MEMB.2006.1607668

Fan, J., McCandliss, B.D., Sommer, T., Raz, A., Posner, M.I.: Testing the Efficiency and Independence of Attentional Networks. Journal of Cognitive Neuroscience **14**(3), 340–347 (2002) https://doi.org/10.1162/089892902317361886

MacDonald, A., Goghari, V., Hicks, B., Flory, J., Carter, C., Manuck, S.: A convergent-divergent approach to context processing, general intellectual functioning. Neuropsychology **19**(6), 814–821 (2005)

Kirby, K.N.: One-year temporal stability of delay-discount rates. Psychonomic Bulletin & Review **16**(3), 457–462 (2009) https://doi.org/10.3758/PBR.16.3.457

Logan, G., Bundesen, C.: Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? Journal of Experimental Psychology: Human Perception and Performance **29**(3), 575–599 (2003)

Logan, G.D., Cowan, W.B.: On the Ability to Inhibit Simple and Choice Reaction Time Responses: A Model and a Method. Journal of experimental psychology **10**(2), 276–291 (1984)

DeJong, R., Coles, M., Logan, G.: Strategies and mechanisms in nonselective and selective inhibitory motor control. Journal of Experimental Psychology: Human Perception and Performance **21**(3), 498–511 (1995)

Sochat, V., Eisenberg, I., Enkavi, A., Li, J., Bissett, P, Poldrack, R: The Experiment Factory: Standardizing Behavioral Experiments. Frontiers in Psychology **7**, 610 (2016) https://doi.org/10.3389/fpsyg.2016.00610

Bissett, P.G, Eisenberg, I.W., Shim, S., Rios, J.A.H., Jones, H.M., Hagen, M.P, Enkavi, A.Z, Li, J.K, Mumford, J.A., MacKinnon, D.P., Marsch, L.A, Poldrack, R.P: Cognitive tasks, anatomical MRI, and functional MRI data evaluating the construct of self regulation. preprint at https://www.biorxiv.org/content/10.1101/2023.09.27.559869v1 (2023)

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., Gorgolewski, K.J.: fMRIPrep: A robust preprocessing pipeline for functional MRI. Nature Methods **16**(1), 111–116 (2019) https://doi.org/10.1038/s41592-018-0235-4

Greve, D.N., Fischl, B.: Accurate and robust brain image alignment using boundary-based registration. NeuroImage **48**(1), 63–72 (2009)

Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage **17**(2), 825–841 (2002) https://doi.org/10.1006/nimg.2002.1132

Cox, R.W., S, H.J.: Software tools for analysis and visualization of fMRI data. NMR in Biomedicine **10**, 171–178 (1997)

Pruim, R.H.R., Mennes, M., Rooij, D., Llera, A., Buitelaar, J.K., Beckmann, C.F.: ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage **112**, 267–277 (2015) https://doi.org/10.1016/j.neuroimage.2015.02.064

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G.: Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics **8**, 14 (2014) https://doi.org/10.3389/fninf.2014.00014 24600388

Smith, S., Nichols, T.: Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage **44**(1), 83–98 (2009) https://doi.org/10.1016/j.neuroimage.2008.03.061

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

Supplementary Materials

Details about tasks involved in real data analysis

Table S1: fMRI task summaries

Name	**Description**	**N**	**Age mn(sd)**	**N Female**
Attention Network Test (ANT)	Tests three aspects of attention or "attentional networks": alerting, orienting, and executive control	91	24(5)	60
Delay-Discounting Task (DDT)	Measure of temporal discounting, the tendency for people to prefer immediate monetary rewards over delayed rewards	86	24(6)	57
Dot Pattern Expectancy (DPX)	Measure of individual differences in cognitive control including proactive and reactive control modes	91	24(5)	61
Motor Selective Stop Signal	Measures the ability to engage response inhibition selectively to specific responses	91	24(5)	63
Stop-Signal Task	Measure of response inhibition	91	24(5)	60
Stroop	Measure of cognitive control perhaps including resisting distraction or attentional filtering	94	24(5)	62
Cued Task-Switching Task (CTS)	Indexes the processes involved in reconfiguring the cognitive system to support a new task	94	24(5)	61

The Attention Network Test (ANT) is a task designed to test three attentional networks: (1) alerting, (2) orienting, and (3) executive control. The ANT combines attentional and spatial cues with a flanker task (a central imperative stimulus is flanked by distractors that can indicate the same or opposite response to the imperative stimulus). On each trial a spatial cue is presented, followed by an array of five arrows presented at either the top or the bottom of the computer screen. The subject must indicate the direction of the central arrow in the array of five. The cue that precedes the arrows can be non-existent, a center cue, a double cue (one presented at each of the two possible target locations), or a spatial cue that deterministically indicates the upcoming target location. Each network is assessed via reaction times (RTs). The alerting network contrasts performance with and without cues, the orienting network contrasts performance on the task with or without a reliable spatial cue, and executive control (conflict) is measured by assessing interference from flankers.

The Dot Pattern Expectancy (DPX) task measures individual differences in cognitive control. Participants are presented with a cue made up of dots. This cue can be a valid cue – referred to as A (e.g., ":") – or an invalid cue – referred to as B (e.g., ".."). Next a probe is presented, also made up of a simple dot formation. This probe can be valid (X) or invalid (Y). Participants are instructed to respond to valid probe and cue combinations (targets – AX combinations) with a key press (e.g., "x") and all others (non-targets) with a different key press (e.g., "m").

The Delay-Discounting Task (DDT) is a measure of temporal discounting, the tendency for people to prefer smaller, immediate monetary rewards over larger, delayed rewards. Participants complete a series of 27 questions that each require choosing between a smaller, immediate reward (e.g., $25 today) versus a larger, later reward (e.g., $35 in 25 days). The 27 items are divided into three groups according to the size of the larger amount (small, medium, or large). Modeling techniques are used to fit the function that relates time to discounting. The main dependent measure of interest is the steepness of the discounting curve such that a more steeply declining curve represents a tendency to devalue rewards as they become more temporally remote.

The cued task-switching task indexes the control processes involved in reconfiguring the cognitive system to support a new stimulus-response mapping. In this task, subjects are presented with a task cue followed by a colored number (between 1-4 or 6-9). The cue indicates whether to respond based on parity (odd/even), magnitude (greater/less than 5), or color (orange/blue). Trials can present the same cue and task, or can switch the cue or the task. Responses are slower and less accurate when the cue or task differs across trials (i.e., a switch) compared to when the current cue or task remains the same (i.e., a repeat).

The Stop-Signal Task is designed to measure motor response inhibition, one aspect of cognitive control. On each trial of this task participants are instructed to make a speeded response to an imperative "go" stimulus except on a subset of trials when an additional "stop signal" occurs, in which case participants are instructed that they should make no response. The Independent Race Model describes performance in the Stop-Signal Task as a race between a go process that begins when the go stimulus occurs and a stop process that begins when the stop signal occurs. According to this model, whichever independent process reaches completion first determines the resulting behavior; earlier completion of the go process results in an overt response (i.e., stop-failure), whereas earlier completion of the stop process results in successful inhibition. The main dependent measure, stop-signal reaction time (SSRT), can be computed such that lower SSRT indicates greater response inhibition. One variant of the task measures proactive slowing, the tendency for participants to respond more slowly in anticipation of a potential stopping signal. This variant often uses multiple probabilities of a stop signal (e.g., 20% and 40%) to manipulate participants' expectancies about the likelihood of a stop signal occurring. The extent of slowing in the higher compared to the lower stop probability conditions is an index of proactive slowing/control.

The motor selective stop-signal task measures the ability to engage response inhibition selectively to specific responses. In this task, cues are presented to elicit motor responses (e.g., right hand responses, left hand responses). A stop-signal is presented on some trials, and subjects must stop if certain responses are required on that trial (e.g., right hand responses) but not others (e.g., left hand responses) if a signal occurs. In contrast to a simple stop-signal task in which all actions are stopped when a stop-signal is presented, this task aims to be more like stopping in "the real world" in that certain motor actions must be stopped (e.g., stop pressing the accelerator at a red light) but others should proceed (e.g., steering the car and/or conversing with a passenger). Commonly, stop-signal reaction time (SSRT), the main dependent measure for response inhibition in stopping tasks, is prolonged in the motor selective stopping task when compared to the more canonical simple stopping task. This prolongation of SSRT is taken as evidence of the cost of engaging inhibition that is selective to specific effectors or responses.

The Stroop task is a seminal measure of cognitive control. Successful performance of the task requires the ability to overcome automatic tendencies to respond in accordance with current goals. On each trial of the task, a color word (e.g., "red", "blue") is presented in one of multiple ink colors (e.g., blue, red). Participants are instructed to respond based upon the ink color of the word, not the identity of the word itself. When the color and the word are congruent (e.g., "red" in red ink), the natural tendency to read the word facilitates performance, resulting in fast and accurate responding. When the color and the word are incongruent (e.g., "red" in blue ink), the strong, natural tendency to read must be overcome to respond to the ink color. The main dependent measure in the Stroop task is the "Stroop Effect", which is the degree of slowing and the reduction in accuracy for incongruent relative to congruent trials.

Exclusion information by task for real data analysis

Table S2: Exclusion information for Attention Network task.

Incomplete data	**Subject omitted (issues with behav. > 50% of tasks)**	**High motion >20% total volumes**	**No response on >45% of trials**	**Stopped performing task at end of scan**	**Poor performance (subjective)**
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	1
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	1	0	0

Table S3: Exclusion information for Delay-Discount task.

Incomplete data	**Subject omitted (issues with behav. > 50% of tasks)**	**High motion >20% total volumes**	**No response on >45% of trials**	**Stopped performing task at end of scan**	**Poor performance (subjective)**	**Made same choice on all trials**
1	1	0	0	0	0	0
1	1	0	0	0	0	0
1	1	0	0	0	0	0
1	1	0	0	0	0	0
1	1	0	0	0	0	0
1	1	0	0	0	0	0
1	1	0	0	0	0	0
1	1	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
0	1	0	1	0	0	0
0	1	0	0	0	0	0
0	1	0	0	0	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1
0	0	0	0	0	0	1

Table S4: Exclusion information for Dot Pattern Expectancy task.

Incomplete data	**Subject omitted (issues with behav. > 50% of tasks)**	**High motion >20% total volumes**	**No response on >45% of trials**	**Stopped performing task at end of scan**	**Poor performance (subjective)**
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	1
0	1	0	0	0	1
0	1	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	0	0	1	0
0	0	0	0	0	1
0	0	0	0	0	1

Table S5: Exclusion information for Motor Selective Stop Signal task.

Incomplete data	Subject omitted (issues with behav. > 50% of tasks)	High motion >20% total volumes	No response on >45% of trials	Stopped performing task at end of scan	Poor performance (subjective)	>75% stop success rate	<25% stop success rate
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	1	0	1	0	0	1	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	1

Table S6: Exclusion information for Stop Signal task.

Incomplete data	**Subject omitted (issues with behav. > 50% of tasks)**	**High motion >20% total volumes**	**No response on >45% of trials**	**Stopped performing task at end of scan**	**Poor performance (subjective)**	**>75% stop success rate**	**<25% stop success rate**
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	1	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0

Table S7: Exclusion information for Stroop task.

Incomplete data	**Subject omitted (issues with behav. > 50% of tasks)**	**High motion >20% total volumes**	**No response on >45% of trials**	**Stopped performing task at end of scan**	**Poor performance (subjective)**
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
0	1	0	0	0	1
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	0	1	0	1

Table S8: Exclusion information for Cued Task Switching task.

Incomplete data	**Subject omitted (issues with behav. > 50% of tasks)**	**High motion >20% total volumes**	**No response on >45% of trials**	**Stopped performing task at end of scan**	**Poor performance (subjective)**
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	1	0	0	0	0
1	0	0	0	0	0
0	1	0	1	0	0
0	1	0	0	0	1
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0