The Journal of Neuroscience, 0, 2023 • 00(00):000 • 1

Behavioral/Cognitive

Differential Recruitment of Inhibitory Control Processes by Directed Forgetting and Thought Substitution

AQ:au Ryan J. Hubbard[1,2] and Lili Sahakyan[1,2]

[1]Beckman Institute for Advanced Science and Technology, University of Illinois Urbana–Champaign, Urbana, Illinois 61801 and [2]Department of Psychology, University of Illinois Urbana–Champaign, Urbana, Illinois 61801

AQ:A Humans have the ability to intentionally forget information via different strategies, included suppression of encoding (directed forgetting) and mental replacement of the item to encode (thought substitution). These strategies may rely on different neural mechanisms; namely, encoding suppression may induce prefrontally mediated inhibition, whereas thought substitution is potentially accomplished through modulating contextual representations. Yet, few studies have directly related inhibitory processing to encoding suppression, or tested its involvement in thought substitution. Here, we directly tested whether encoding suppression recruits inhibitory mechanisms with a cross-task design, relating the behavioral and neural data from male and female participants in a Stop Signal task (a task specifically testing inhibitory processing) to a directed forgetting task with both encoding suppression (Forget) and thought substitution (Imagine) cues. Behaviorally, Stop Signal task performance (stop signal reaction times) was related to the magnitude of encoding suppression, but not thought substitution. Two complementary neural analyses corroborated the behavioral result. Namely, brain-behavior analysis demonstrated that the magnitude of right-frontal beta activity following stop signals was related to stop signal reaction times and successful encoding suppression, but not thought substitution; and classifiers trained to discriminate successful and unsuccessful stopping in the Stop Signal task could also classify successful and unsuccessful forgetting following Forget cues, but not Imagine cues. Importantly, inhibitory neural mechanisms were engaged following Forget cues at a later time than motor stopping. These findings not only support an inhibitory account of directed forgetting, and that thought substitution engages separate mechanisms, but also potentially identify a specific time in which inhibition occurs when suppressing encoding. AQ:B

Key words: EEG; forgetting; inhibition; memory

Significance Statement

Forgetting often seems like an unintended experience, but forgetting can be intentional, and can be accomplished with multiple strategies. These strategies, including encoding suppression and thought substitution, may rely on different neural mechanisms. Here, we test the hypothesis that encoding suppression engages domain-general prefrontally driven inhibitory control mechanisms, while thought substitution does not. Using cross-task analyses, we provide evidence that encoding suppression engages the same inhibitory mechanisms used for stopping motor actions, but these mechanisms are not engaged by thought substitution. These findings not only support the notion that mnemonic encoding processes can be directly inhibited, but also have broad relevance, as certain populations with disrupted inhibitory processing may be more successful accomplishing intentional forgetting through thought substitution strategies.

Introduction

Forgetting information can be adaptive (Bjork, 1989), and AQ:C while most forgetting occurs unintentionally, individuals can intentionally forget information (Sahakyan et al., 2013; Anderson and Hanslmayr, 2014; Sahakyan, 2022). Intentional forgetting has primarily been studied with the directed forgetting (DF) task, in which study items are followed by cues to remember or forget (Bjork et al., 1968), and items followed by forget cues are remembered less often. Historically, the selective rehearsal account (Bjork, 1970, 1972; MacLeod, 1975; Basden et al., 1993) explained the effect as effortful processing only following cues to remember. However, cognitive neuroscientific research has

Received Apr. 8, 2022; revised Dec. 15, 2022; accepted Jan. 13, 2023.

Author contributions: R.J.H. and L.S. designed research; R.J.H. performed research; R.J.H. analyzed data; R.J.H. wrote the first draft of the paper; L.S. contributed unpublished reagents/analytic tools; L.S. edited the paper.

This work was supported by Beckman Postdoctoral Fellowship to R.J.H. We thank Dr. Dace Apšvalka for guidance and software for conducting the behavioral PLS analysis; and Dr. Jarrod Lewis-Peacock for advice on conducting the decoding analysis.

The authors declare no competing financial interests.

Correspondence should be addressed to Ryan J. Hubbard at rjhubba2@illinois.edu.

https://doi.org/10.1523/JNEUROSCI.0696-22.2023

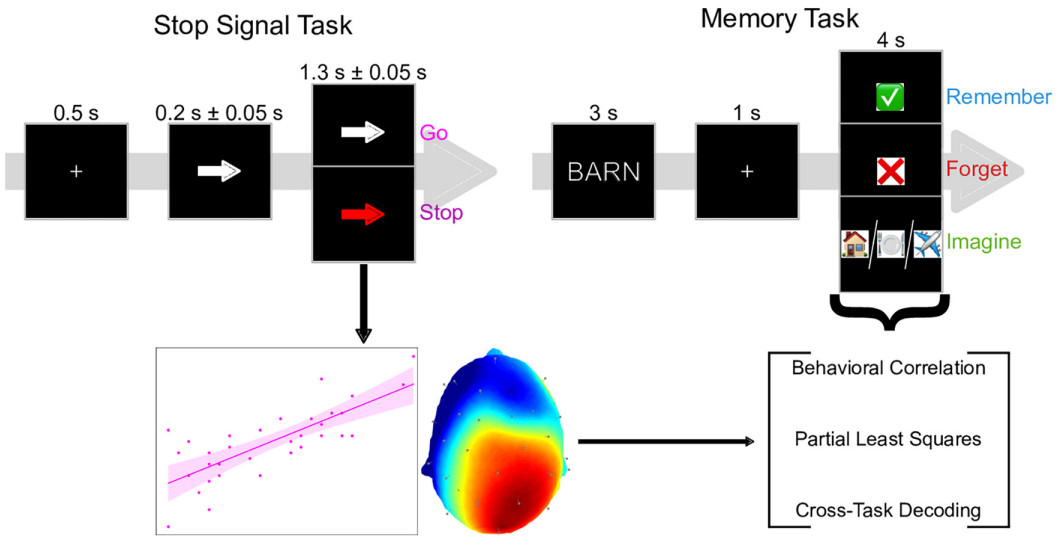

Figure 1. Outline of the experimental procedure and analytical method. Participants first completed the Stop Signal task, in which they responded by pressing the arrow key presented, but stopped this response if the arrow turned red. Afterward, participants completed the DF task, where words were followed by specific memory instructions. Multiple analyses were then conducted to relate behavioral performance across these tasks, as well as engagement of neural processes between successful stopping and intentional forgetting.

identified that active neural processing occurs following cues to forget that is predictive of future forgetting, particularly in the PFC (Wylie et al., 2008; Rizio and Dennis, 2013; Oehrn et al., 2018).

Different strategies to intentionally forget information also engage different mechanisms. When participants are told to think of something else instead of forget information, a strategy of thought substitution, similar rates of forgetting are observed (Sahakyan and Kelley, 2002; Hertel and Calcaterra, 2005), but different neural mechanisms are engaged (Bergström et al., 2009; Benoit and Anderson, 2012; H. Kim et al., 2020). Recently, we developed a modified DF paradigm, including cues to directly suppress encoding ("Forget" cues), as well as cues to perform thought substitution ("Imagine" cues), during encoding (Hubbard and Sahakyan, 2021). Both strategies produced forgetting, but recruited separable neural mechanisms, with only Forget cues eliciting frontal oscillatory activity.

These findings are consistent with other studies using the Think-No-Think (TNT) task to study forgetting, which focuses on retrieval rather than encoding suppression (Anderson and Green, 2001). Retrieval suppression also engages the PFC (Anderson et al., 2004; Depue et al., 2007), which may reflect engagement of top-down inhibitory control mechanisms to suppress mnemonic functions (Anderson et al., 2016; Anderson and Hulbert, 2021). To link prefrontal activity observed during retrieval suppression to inhibitory control mechanisms, researchers have compared neural responses observed in the TNT task to responses observed in the Stop Signal paradigm, a task that directly measures engagement of inhibitory mechanisms to stop actions (Logan and Cowan, 1984; Verbruggen and Logan, 2009). Research using this task has delineated a critical role of the right PFC in action stopping (Chevrier et al., 2007; Aron et al., 2014), and that β oscillations in this region support successful inhibition (Swann et al., 2009; Wagner et al., 2018). Studies using both Stop Signal and TNT tasks have identified that similar neural substrates (Apšvalka et al., 2022) and oscillatory mechanisms (Castiglione et al., 2019) support inhibition of actions and thoughts, suggesting that domain-general prefrontal control is engaged.

The previous work focused on the TNT task, where retrieval must be inhibited. However, no study to date has examined the role of domain-general prefrontal control in the DF task, where encoding is suppressed. Suppressing encoding and stopping a motor action may rely on the same mechanisms (Hourihan and Taylor, 2006; Fawcett and Taylor, 2010), as items that are inhibited in a Stop Signal task are remembered less often (Chiu and Egner, 2015a, 2015b), and prefrontal activity is observed following cues to forget (Oehrn et al., 2018); yet, assuming this to be the case would be a reverse inference, and this hypothesis must be specifically tested.

Here, we test the hypothesis that suppression of encoding relies on domain-general inhibitory mechanisms, and that this engagement is specific to encoding suppression, not thought substitution. We related behavioral performance and neural processing during a Stop Signal task to Forget and Imagine cues in the DF task (Hubbard and Sahakyan, 2021). We used behavioral partial least squares (PLS) and neural decoding techniques to examine the similarity in neural processing between these tasks to directly test the degree to which different strategies of intentional forgetting engaged top-down inhibitory control mechanisms. We predicted that top-down inhibitory control mechanisms engaged by stopping motor actions would also be engaged by Forget cues, but not by Imagine cues.

Materials and Methods

Experimental design and statistical analysis

Participants. The participants were the same as reported by Hubbard and Sahakyan (2021). Three additional participants were dropped for violating the assumptions necessary to accurately calculate stop signal reaction times (SSRTs), resulting in a total of 33 participants in the final analysis. All participants reported normal or corrected vision and had no history of any neurologic or psychiatric disorder. Mean age was 21 years (range 18-30 years), and 22 of the participants were female. The study was approved by the Institutional Review Board of University of Illinois Urbana–Champaign, and all participants provided written informed consent and were debriefed following participation.

Materials and procedure. An outline of the experimental procedure is presented in Figure 1. The Stop Signal task was designed in concordance with guidelines for accurate measurement of response inhibition (Verbruggen et al., 2019). After informed consent and EEG setup, participants were comfortably seated ~100 cm from a monitor in a quiet room,

where they received instructions for the Stop Signal task. Participants performed four blocks of the task, each block containing 80 trials; 25% of the trials were Stop trials, leading to a total of 240 Go trials and 80 Stop trials. Each trial began with blank screen for 500 ms, followed by a fixation cross for 500 ms, after which a left or right arrow was presented, and participants were instructed to respond as quickly as possible by pressing the matching arrow key on the keyboard. On Stop trials, the arrow turned red after a brief delay, indicating that participants should withhold their keyboard response. The delay between the onset of the arrow and the onset of the Stop stimulus varied based on the performance of the participant: it started at 200 ms, increased by 50 ms if participants successfully stopped their response, and decreased by 50 ms if participants were unsuccessful in stopping. The arrow remained on the screen for a total of 1500 ms.

Following the Stop Signal task, participants took a short break, and then were given instructions for the DF memory task, outlined in Hubbard and Sahakyan (2021). First, participants were given a familiarization period in which they were presented with each of the three separate Imagine prompts, and were given 60 s to visualize and verbally describe a clear mental image that related to the cue. The Imagine cues were selected from previous DF studies that compared Forget instructions with a thought substitution condition in a list-method paradigm (Sahakyan and Kelley, 2002; Delaney et al., 2010). The house prompt corresponded to imagining their childhood home, the silverware prompt corresponded to imagining their high school cafeteria, and the plane prompt corresponded to imagining a recent vacation. Three different imagine prompts (as opposed to a single prompt) were used to increase the chances of engaging in different mental contexts shifts throughout the experiment rather than repeatedly revisiting the same mental context. All of the participants in the study were able to provide vivid images and details in response to the imagination prompts After these instructions, the participants were instructed that they would be presented with a series of words to study, and each word would be followed by an image that would tell them what to do next. If they saw a green check mark (a Remember cue), they should try to remember the word, as their memory for it would be tested later. If they saw a red X (a Forget cue), they should try to forget the word, and their memory for that word would not be tested later. Last, if they saw one of the three Imagine prompts (house, silverware, or plane), they should mentally generate the mental image that they had previously created in the familiarization period, and focus on that, instead of the word they just studied., Memory performance did not differ between the three imagination prompts. Study words were presented centrally for 3 s, followed by a 1 s fixation, and then a 4 s memory cue presentation. Participants viewed 126 words and cues in total, with 42 in each cue condition, and 14 in each of the three Imagine conditions.

Following the encoding phase, the participants performed a recognition task, in which they indicated whether presented words were old or new. They were informed that their memory for the Forget cue items would indeed be tested, and they should respond "old" even if they recall that the word was originally a Forget cue item (i.e., the Forget instructions were canceled at the time of test). Participants were presented with all 126 study words, intermixed with 84 new words, for a total of 210 words in a random order. Each word was presented centrally for 2 s, after which a prompt to make an Old/New response appeared.

The stimuli in the Stop Signal task were filled white arrows pointing to the left or the right, which turned red during stop trials. In the DF task, the stimuli were 210 medium frequency nouns (Kucera & Francis mean word frequency of 43, 4-6 letters in length). The assignment of each word as either an old or new word, as well as to each of the three memory instruction conditions, was randomized for each participant. Picture images, downloaded from www.emojipedia.org, were used to designate the three memory instructions: Remember, Forget, and Imagine. Remember cues were represented by a green check mark, while Forget cues were depicted by a red X. For the Imagine cues, pictures of a house, a silverware set, and an airplane were used.

AQ:E

Stop signal behavioral analysis

Analysis of stop signal responses was conducted following recommendations from Verbruggen et al. (2019). Namely, SSRTs were calculated for each participant using the integration method, and response omissions to Go trials were replaced with the participant's max Go trial RT. As stated in Participants, for 3 participants, mean RTs on unsuccessful Stop trials were numerically larger than mean RTs on Go trials. This violates the assumptions of the independent race model, rendering SSRT estimates unreliable (Band et al., 2003); thus, these 3 participants were removed from the analysis.

Recognition memory behavioral analysis

Recognition memory performance was analyzed with mixed effect logistic regression models predicting whether participants made a correct or incorrect recognition response on trial-level behavioral data. Models were fit by maximum likelihood using the lme4 package in R (Bates et al., 2015), and Wald's z scores were computed for each coefficient to test for significance of fixed effects. Random factors included intercepts for items and slopes and intercepts for participants for the fixed effect of condition. Correlations between random factors were not calculated to ease convergence of the models. To test differences in recognition accuracy by cue condition, cue condition was included as a fixed effect in the model.

Cross-task behavioral analysis

Given that both SSRTs and overall magnitudes of both DF and thought substitution are subject-level measurements as opposed to trial-level measurements, and that the correlation between DF and thought substitution is high because of overall memory performance, leading to interpretation problems in linear regression models because of collinearity, we opted to use correlation methods instead. SSRTs and behavioral performance on the DF task were related using robust correlation methods (Pernet et al., 2013). First, Spearman correlations were used to relate variables, as this method is less sensitive to univariate outliers compared with Pearson correlations (Rousselet and Pernet, 2012). Second, the skipped correlation method was used to detect and remove bivariate outliers in data to be correlated (Rousseeuw, 1984; Rousseeuw and van Driessen, 1999; Hubert et al., 2008). Here, the minimum covariance determinant of the data was computed to estimate multivariate location and scatter, and data points outside the bound defined by the interquartile range were considered outliers and removed. If no outliers were detected, the skipped correlations were equivalent to Spearman correlations between variables. If outliers were detected, skipped Spearman correlations, in which the outliers were removed, were reported. To avoid fallacious interpretations of the presence of interactions based on differences in correlation magnitudes (Nieuwenhuis et al., 2011), we directly statistically tested correlation differences using multiple methods, implemented by the *cocor* package in R (Diedenhofen and Musch, 2015). When reporting statistical differences between correlations, we report Steiger's z (Steiger, 1980) and Zou's 95% CI for differences in the correlation (Zou, 2007). For the memory task, we calculated magnitude of forgetting for each of the cues for each participant by subtracting average performance for Forget cue items or Imagine cue items from average performance for Remember cue items, which we refer to as DF magnitude and TS magnitude, respectively.

Interpretations of behavioral correlations solely based on p values are problematic, and some variance could be shared between the tasks. Therefore, we additionally supplemented the correlation analyses with Bayesian correlation analysis (Ly et al., 2016), which provides a Bayes factor for the correlation. Bayes factors provide estimates of the odds of the alternative hypothesis over the null hypothesis, and thus can provide a more nuanced understanding of the magnitude of effects (Jarosz and Wiley, 2014). Bayes factors are reported in terms of odds in favor of the alternative hypothesis (e.g., $BF_{10} = 5$ would be 5:1 odds). In an attempt to control for shared variance, we also calculated partial correlations, in which the correlation between variables or conditions of interest was calculated after eliminating the effect of the other condition (S. Kim, 2015).

EEG recording and preprocessing

EEG data were recorded from 26 Ag/AgCl electrodes embedded into a flexible elastic cap and distributed over the scalp in an equidistant arrangement. Additional facial electrodes were attached for monitoring

of electro-oculogram) artifacts, including one adjacent to the outer canthus of each eye and one below the lower eyelid of the left eye. Electrode impedances were kept <10 kΩ. Signals were amplified by a BrainVision amplifier with a 16-bit A/D converter, an input impedance of 10 MΩ, an online bandpass filter of 0.016-100 Hz, and a sampling rate of 1 kHz. The left mastoid electrode was used as a reference for online recording; offline, the average of the left and right mastoid electrodes was used as a reference.

Following data collection and offline rereferencing, each raw EEG time series was passed through a 0.2-40 Hz Butterworth filter with a 36 dB/oct roll-off. Filter parameters were chosen *a priori* to remove low-frequency drifts without causing artifacts in ERP analyses (Tanner et al., 2015), as well as to remove high-frequency noise but still allow for examination of beta band activity in time-frequency analyses. The time series was then segmented into epochs ranging from −700 to 1500 ms relative to the onset of the stop signal during the stop signal task, and to each memory cue during the encoding section of the memory task. Epoched data were then submitted to AMICA, an independent component analysis algorithm that decomposes the signal into independent components (ICs) (Palmer et al., 2012), and components with time series and topographies indicative of eye-related activity were removed. Across participants, 1-3 components reflecting ocular activity were removed (mean = 2.1, SD = 0.43). Last, the electro-oculogram-cleaned data were scanned for large voltage deflections (>90 μV), and manually scanned by eye, to remove any epochs with remaining artifacts. Overall, data quality was high, and few trials were removed (mean = 4.4% across subjects, SD = 4.1%). The average number of trials per condition were as follows: Successful Stop, mean = 46 (SD = 8.2); Unsuccessful Stop, mean = 29 (SD = 6.8); Remember cue, mean = 40 (SD = 2.7); Forget cue, mean = 40 (SD = 2.8); and Imagine cue, mean = 40 (SD = 2.5).

Event-related potential (ERP) and event-related spectral perturbation (ERSP) analysis

To determine whether inhibitory processes were engaged during the Stop Signal task and that our results replicated prior research, we examined differences in ERPs, or changes in scalp amplitudes in response to stimuli (Luck, 2014), as well as ERSPs, or changes in oscillatory power in response to stimuli (Makeig, 1993). Before averaging, EEG trials were baseline corrected with a *z* score baseline procedure that reduces potential biases from standard baseline correction procedures (Ciuparu and Mureşan, 2016). For ERP analyses, the time series from −200 to −1 ms before stimulus was extracted from each trial and concatenated; for time-frequency analyses, the time series from −400 to −200 ms before stimulus was extracted. Trials that were identified as artifacts were left out of the baseline. Each trial was then *z*-scored by the average and SD of the concatenated baseline. Separate baseline corrections were performed for successful and unsuccessful stop trials, as well as for the different memory cues.

For ERP analyses, a central channel cluster was created from the average of 5 channels (see Fig. 3*A*) for examining P3 ERP responses. We tested latency differences in the P3 using the relative criterion technique of 50% peak amplitude combined with the jackknife approach (Kiesel et al., 2008; Miller et al., 2009). Specifically, instead of taking within-subject latency measurements, latencies were measured from subsample grand average waveforms, where each participant was omitted from one of the averages. The peak amplitude was identified at the central channel cluster in the time window from 200 to 600 ms for both successful and unsuccessful stops, and the latency at which the amplitude reached 50% of the peak was recorded. Differences in latencies were tested with a *t* test; the critical *t* value was adjusted as described by Ulrich and Miller (2001).

For the time-frequency analysis, EEG epochs were first convolved with Morlet wavelets that varied in width (number of cycles) to improve temporal and frequency precisions. The width started at 3 and increased linearly to a width of 7 across a frequency range of 3-30 Hz, resulting in time-frequency bins of 20 ms and 0.5 Hz, respectively. Statistical analyses of comparing successful stops to unsuccessful stops were conducted with nonparametric cluster-based permutation tests (Maris and Oostenveld, 2007). In these tests, *t* tests on differences in power across participants were calculated at each time point, frequency bin, and channel, and significant *t* values that were adjacent in space and time were clustered together. Clusters were characterized by taking the sum of *t* values within the adjacent points. These observed clusters were compared with a permutation distribution, generated by shuffling the condition labels of the data, testing for differences on the shuffled data, finding clusters of differences, and summing the *t* values of the clusters 2000 times. The most extreme cluster was recorded for each permutation, and distributions of these most extreme cluster sums were created for comparison with the observed cluster sums. Reported *p* values represent the percentile ranking of the observed clusters compared with the permutation distribution.

Dipole fitting analysis

Interpreting scalp topographies of ERP and time-frequency effects can be misleading and may vary depending on the scalp coverage, reference used, and individual variability. To determine whether the right frontal cortex was indeed recruited by the Stop Signal task, and to relate this activity to the DF task, we used dipole fitting to localize the source of the Stop Signal activity. Following previously used methods (Wagner et al., 2018; Castigline et al., 2019), the raw EEG data were passed through a 2-60 Hz bandpass Butterworth filter, and rereferenced to the average reference. The data were then segmented into epochs around the onsets of the stop signal −1 to 1.5 s), including both successful and unsuccessful stop trials. These epochs were concatenated and submitted to AMICA for independent component analysis decomposition, resulting in ICs potentially reflecting neural sources. AQ:F

Following this, equivalent current dipoles were estimated for each IC using the DIPFIT2 toolbox in EEGLAB (Oostenveld and Oostendorp, 2002), in which the dipoles were fit to the scalp projections of the ICs using a standardized three-shell boundary element head model. The EEG electrode locations were aligned with the standard MNI brain model for mapping to underlying neural sources. Dipoles were then fit to the ICs by first performing a coarse-grained grid search across the 3D grid of the brain, which provided better starting locations for more fine-grained nonlinear optimization of the dipole positions. IC scalp projections, dipole positions, and dipole source activity were then visually inspected to ensure that the sources were not simply noise, ocular artifact, or positioned outside of the brain. We based our selection of ICs on previous research identifying activity from a right frontal source underlying action stopping (Wagner et al., 2018). For each participant, an IC was selected with a frontal scalp topography, a dipole localized to the right frontal lobe, and a residual variance <85%. We focused our selection on spatial and anatomic constraints to avoid any double dipping. Theoretically, the source activity time course of the selected ICs reflects neural activity specifically from the underlying brain source, but analysis of the source activity was not factored into IC selection. Topographies of subject-level ICs can be found at the OSF repository for the project (see Data availability statement).

For each participant, the selected IC source activity time course was submitted to a time-frequency analysis using wavelet decomposition, identical to the time-frequency analysis described above for the EEG data. The time-frequency data at the source level were then averaged for successful stops, as well as unsuccessful stops, and the difference between condition averages was calculated. As before, significant differences in time-frequency activity were calculated using nonparametric cluster-based permutation tests, in which *t* tests on differences in power across participants were calculated at each time-frequency bin, and the cluster-sum statistics were compared with a permutation distribution created by shuffling the condition labels.

PLS brain-behavior analysis

To relate neural inhibitory responses to behavioral outcomes in a data-driven fashion, we performed behavioral PLS correlation (McIntosh and Lobaugh, 2004; Krishnan et al., 2011). Briefly, PLS is a singular value decomposition (SVD) based analysis that attempts to reduce an X matrix of variables (e.g., voxels or time points of neuroimaging data) into a set of latent variables (LVs) that maximize covariance with a second Y matrix of variables (e.g., behavioral scores). We adopted a similar analysis strategy used in recent TNT studies, where fMRI data were collected and

related to retrieval suppression, for our EEG data (Gagnepain et al., 2017; Apšvalka et al., 2022). Namely, for each participant, we calculated the average difference in power between successful and unsuccessful stop trials at each time-frequency channel bin of their time-frequency data following the onset of the stop signal, and converted this 3D matrix of differences into a row vector. The X matrix for the PLS analysis was created by concatenating each participant's time-frequency data (i.e., the rows of the matrix represented participants, and the columns represented a time-frequency channel bin). The Y matrix had a similar structure, where each row represented a participant, and each column represented behavioral scores from a task (SSRTs, DF magnitude, and TS magnitude).

To perform the analysis, both the X and Y matrices were mean-centered, and each row was normalized to set the sum of squares of all its values to 1 to ensure that overall differences in activity and performance between participants was controlled for. Next, a correlation matrix (R) was computed between the X and Y matrices in which each time-frequency bin measurement was related to each behavioral measurement across participants. SVD was then applied to the correlation matrix R to calculate the LVs that maximized covariance between the X and Y data. Each LV had an associated singular value, representing the degree of explained covariance, as well as a matrix of brain saliences, representing the strength of the relationship between oscillatory power and behavioral performance at each time-frequency channel bin. The statistical significance of the LVs was determined using a permutation testing technique, in which the rows of the X matrix were randomly shuffled to randomly reassign the mapping between neural measurements and behavioral scores, and the SVD was recomputed on the shuffled data to obtain a distribution of singular values. This process was repeated 5000 times, and the p value of the observed LV singular value represented its percentile ranking in the permutation distribution of singular values. Statistical significance of the brain salience scores was determined using bootstrapped sampling, in which resampled data were generated by sampling with replacement and SVD was recomputed on the bootstrapped data 5000 times. The observed brain scores were then divided by the SE of the bootstrapped distribution, resulting in a bootrstrapped standard ratio, equivalent to a z score of the brain score. These bootrstrapped standard ratio values could then be thresholded at 1.96, equivalent to a two-tailed p value of 0.05, to determine significant values.

Cross-task neural decoding

To explicitly test whether the same neural processes engaged by successful action stopping in the Stop Signal task were engaged in the DF task, we used a cross-task multivariate pattern analysis of the EEG data, coined neural decoding (King and Dehaene, 2014; Grootswagers et al., 2017). Here, for each participant's data, a linear discriminant analysis (LDA) classifier was trained on time-frequency data to discriminate successful versus unsuccessful stopping in the Stop Signal task. Specifically, oscillatory power values at each frequency channel bin were concatenated to create a feature vector for both successful and unsuccessful stop trials, and the classifier was trained on these features to discriminate the class (Successful Stop vs Unsuccessful Stop). To improve power and to focus on the specific frontal inhibitory mechanism, the time-frequency data were first averaged in the time domain, centered on the cluster identified by the PLS analysis. Additionally, to ensure the classes were balanced, trials were randomly selected so that the trial numbers in each class were balanced.

Once training was complete, the classifier was tested on time-frequency data from the DF task; specifically, the time-frequency data following cue presentation. Separate analyses were conducted for Forget and Imagine cues. Oscillatory power values from each frequency channel bin were concatenated for each trial, and trials were separated based on future memory success (Remembered vs Forgotten). The rationale here is that, if Forget or Imagine cues engage the same inhibitory mechanisms as in the Stop Signal task to promote later forgetting, then the classifier will be able to discriminate successful versus unsuccessful forgetting in the memory task. Importantly, we tested the classifier on oscillatory data at each time point following the onset of the memory cue, which resulted in a time series of classifier accuracies for each participant and for each memory cue.

To test for statistically significant classification, we performed one-tailed t tests on classification accuracies across participants, with the null hypothesis that accuracy was equal to 50% (chance accuracy). To correct for multiple comparisons, we used cluster-based permutation testing (Maris and Oostenveld, 2007). Specifically, we assumed that inhibitory processing would last for some amount of time, and thus contiguous significant time points in the time series of classification accuracies could be grouped to form clusters by summing their t values. Then, the training data labels were shuffled for each participant, the entire classification procedure was conducted again, and the most extreme cluster in the resulting time-series of classification accuracies was recorded. This was repeated 2000 times to create a permutation distribution of cluster accuracies. The reported p values for the observed classification accuracy clusters represent the percentile ranking compared with the permutation distribution.

Our rationale was that the classifier would learn to discriminate neural signals of successful inhibition (successful Stop trials) from unsuccessful inhibition (unsuccessful Stop trials); however, unsuccessful Stop trials could elicit neural signals related to error monitoring, and the classifier could instead learn to discriminate neural signals related to errors (unsuccessful Stop trials) from trials with no errors (successful Stop trials). To rule this out, we performed an additional decoding analysis where the classifier was instead trained to discriminate successful Stop trials from successful Go trials. Here, participants should not elicit any inhibitory or error-related activity during Go trials. However, this analysis added the technical issue of alignment in time; namely, successful Stop trials are aligned to the onset of the Stop signal, whereas there is no onset of a Stop signal present during Go trials. To correct this, for every Go trial, we took the time the participant responded and subtracted the participant's SSRT, and aligned to this time point. In this way, the participant's response on Go trials should be roughly aligned with the internal "response," or onset of inhibition, on Stop trials. Significant time points of classification were calculated using cluster-based permutation, as described above. For both classification analyses, we plotted the most important channels for the classification by multiplying the LDA classifier weights by the covariance of the data and projecting the transformed weights to the scalp channel locations (Haufe et al., 2014).

Testing the classifier accuracy against chance level is informative to determine whether there is neural signal that differentiates between successful and unsuccessful trials that is shared between the tasks; however, we also wished to explicitly test whether cross-task classification performance differed between Forget cues and Imagine cues. We performed additional tests to directly compare the classifier accuracies between Cue conditions. For testing differences in accuracies, we followed the norms of previous literature and conducted statistical tests using Wilcoxon signed rank tests (Chan et al., 2011; Grootswagers et al., 2017; Linde-Domingo et al., 2019). First, cluster-based permutation tests were conducted similarly as described above. At each time point, the classifier accuracies across participants following Forget cues were statistically compared with the classifier accuracies across participants following Imagine cues, and clusters were formed by summing the z values of contiguous significant time points. Then, the Cue condition labels were shuffled for each participant, the statistical tests across time were conducted again, and the most extreme cluster was recorded. This was repeated 2000 times to create a permutation distribution of clusters. The clusters of the observed data were compared with this permutation distribution. Since cluster-based permutation tests may be less sensitive, we also conducted a second analysis in which we specifically tested differences in averaged classifier accuracies between Cue conditions in the time window in which classifier accuracies differed from chance level.

Data availability

The behavioral data, processed neural data, experiment presentation, and analysis files are available at https://osf.io/8s5e7/.

Results

Stop signal behavioral results

Participants performed very few response omissions on Go trials, and close to zero response errors (pressing the incorrect

J_ID: JNEURO Date: 14-February-23 4/Color Figure(s): "F1-F6" ART: JN-RM-0696-22 Page: 6 Total Pages: 14 Comments:

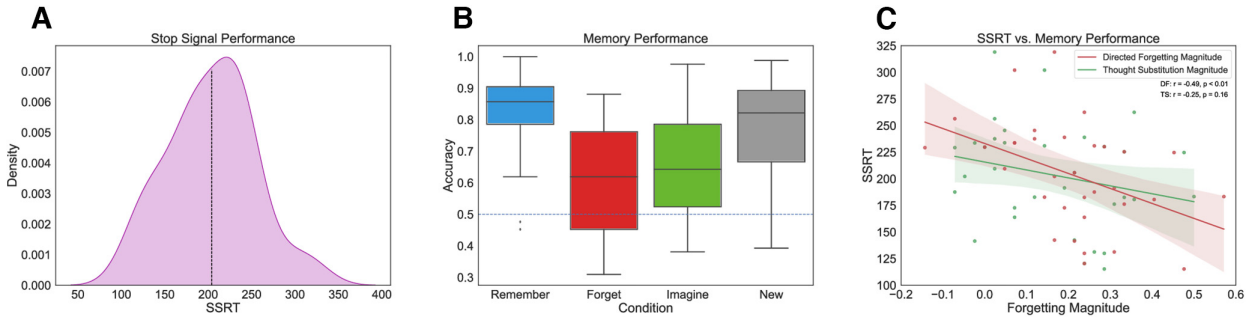

Figure 2. Behavioral results. ***A***, Distribution of SSRTs and mean SSRT (dotted line). ***B***, Recognition memory accuracy for the DF memory task. ***C***, Correlations between SSRT and magnitude of forgetting (Remember - Forget, Remember - Imagine).

directional key), resulting in an average success rate on Go trials of 98%. For Stop trials, the average success rate for stopping was 61% (95% CI of the mean: 58%-64%), and no participant was <25% or >75% stopping. Thus, the assumptions for calculating SSRTs were not violated. The distribution of SSRTs, as well as the mean SSRT, are plotted in Figure 2*A*. The mean SSRT, 205 ms, was in line with other studies using the stop signal task (e.g., Wessel and Aron, 2015).

Recognition memory results

Participants' recognition memory performance for the four types of test items (Remember, Forget, Imagine, and New) are plotted in Figure 2*B*. While a similar figure and analysis were reported in Hubbard and Sahakyan (2021), we recomputed the results with the 3 participants removed for violating the SSRT assumptions. The mixed logit model predicting memory accuracy revealed significant differences between cue conditions; namely, participants had higher memory accuracy for Remember cue items than Forget (β = 1.24, $z = 8.21$, $p < 0.01$) as well as Imagine (β = 1.01, $z = 6.19$, $p < 0.01$) cue items. Additionally, accuracy for Imagine cue items was significantly greater than for Forget cue items (β = 0.24, $z = 2.66$, $p = 0.01$). Thus, we observed a DF effect for both Forget and Imagine items, but the magnitude of the effect was greater for Forget cue items.

Cross-task behavioral results

Given that assumptions were met for calculating SSRTs, and participants demonstrated successful DF and thought substitution memory effects, we tested whether inhibitory efficiency during the Stop Signal task was related to success of intentional forgetting during the DF task. We ran separate analyses for Forget and Imagine cues to determine whether different strategies for intentional forgetting were differentially related to inhibitory efficiency in the Stop Signal task. Forgetting Magnitude scores were calculated as Remember – Forget accuracy (here termed "Directed Forgetting Magnitude") and Remember – Imagine accuracy (here termed "Thought Substitution Magnitude") for each participant, and these Forgetting Magnitude scores were then correlated with SSRTs. Correlations between participant SSRTs and Forgetting Magnitude scores are plotted in Figure 2*C*. Outlier detection methods did not identify bivariate outliers in either correlation; thus, Spearman correlations are reported. The correlation between SSRTs and Directed Forgetting Magnitude was statistically significant ($r_{(31)} = -0.49$, $t = -3.099$, $p = 0.004$). In contrast, the correlation between SSRTs and Thought Substitution Magnitude was not significant ($r_{(31)} = -0.251$, $t = -1.444$, $p = 0.159$). Statistical tests revealed a significant difference between these correlations (Steiger's $z = 2.04$, $p = 0.04$; Zou's 95% CI = 0.012-0.485). Thus, SSRTs were more strongly correlated with Directed Forgetting Magnitude than Thought Substitution Magnitude.

While the correlations significantly differed, the correlation between SSRTs and Thought Substitution Magnitude was in the same direction as the correlation between SSRTs and Directed Forgetting Magnitude, and was perhaps only nonsignificant because of lack of power. Additionally, participant engagement and task attention may have led to shared variance (i.e., participants that paid more attention had higher Directed Forgetting Magnitude and Thought Substitution Magnitude). We conducted further analyses to deal with these issues. First, we used Bayesian correlation analysis to calculate correlation Bayes factors for each of the behavioral correlations. The correlation between SSRTs and Directed Forgetting Magnitude showed moderately strong evidence against the null hypothesis (BF_{10} = 5.54); in contrast, the correlation between SSRTs and Thought Substitution Magnitude showed weak evidence in favor of the null hypothesis (BF_{10} = 0.83). We additionally calculated partial correlations between SSRTs and forgetting magnitudes, in which the variability of one condition was controlled for before calculating the correlation. The partial correlation between SSRTs and Directed Forgetting Magnitude, after controlling for Thought Substitution Magnitude, remained significant ($r_{(31)} = -0.48$, $t = -2.962$, $p = 0.006$). In contrast, The partial correlation between SSRTs and Thought Substitution Magnitude, after controlling for Directed Forgetting Magnitude, was not significant and even flipped direction ($r_{(31)}$ = 0.19, $t = 1.056$, $p = 0.300$). These additional analyses strongly support the notion that SSRTs were related to Directed Forgetting Magnitude, but not Thought Substitution Magnitude.

Stop signal ERPs

We next sought to compare neural processing during the Stop Signal task to processing during the DF task. To this end, we first examined ERP and ERSP responses to stop signals to ensure our results replicated previous studies and further validate the use of the stop signal neural data for a cross-task analysis. Our primary interest in the analysis of ERPs was examining P3 responses, as previous work has demonstrated that P3 latencies are reduced when successfully stopping motor responses (Kok et al., 2004; Wessel and Aron, 2015). ERPs time-locked to the onset of the stop signal and separated into successful and unsuccessful stops are presented in Figure 3*A*. A P3 component from ~175-500 ms over center-parietal channels was clearly observable for both successful and unsuccessful stops, and appeared to peak earlier following successful stops. A jackknife test on peak latencies of the

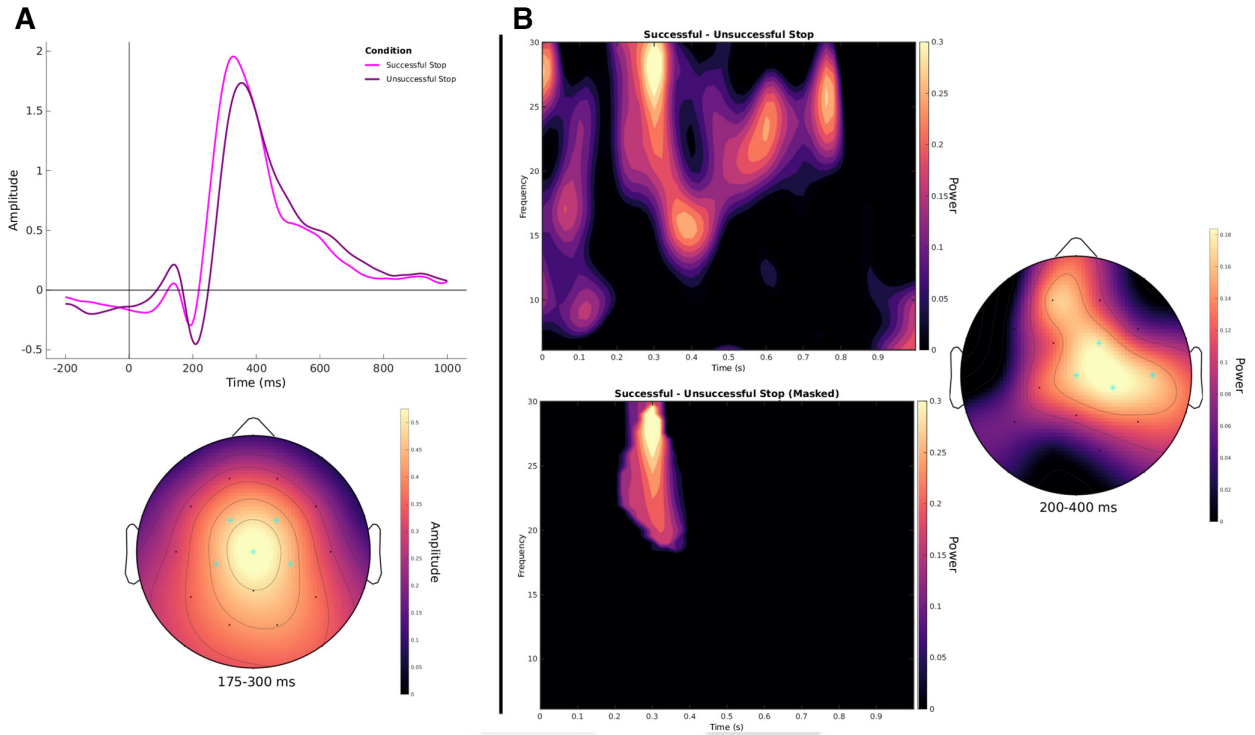

Figure 3. Electrophysiological results of the stop signal task. ***A***, Top, ERPs time-locked to stop signals at a central channel cluster. Bottom, Topography of P3 component. The highlighted channels show the central cluster. ***B***, Top, ERSPs of successful versus unsuccessful stops at a frontal channel cluster. Right, Topography β (18-30 Hz) power from 200 to 400 ms, with highlighted channels showing the frontal cluster. Bottom, Same as in top plot, but statistically masked to highlight the significant frontal β cluster.

P3 component using the relative criterion technique of 50% peak amplitude confirmed this observation: successful stops led to significantly earlier P3 onsets compared with unsuccessful stops (261 ms vs 284 ms; $t_{(32)} = -4.98$, $p < 0.001$), replicating previous stop signal ERP studies (Kok et al., 2004; Wessel and Aron, 2015).

Stop signal ERSPs

Next, we examined changes in oscillatory activity following stop signal cues to determine whether frontal β mechanisms were engaged when stopping was successful. Previous studies have identified a right frontal oscillatory response in the β frequency range (15-30 Hz) related to successful stopping of motor responses (Wagner et al., 2018; Castiglione et al., 2019). ERSPs time-locked to the onset of the stop signal were calculated for both successful and unsuccessful stops. The difference in oscillatory power between successful and unsuccessful stops is presented in Figure 3*B*. Cluster-based permutation analyses identified a significant positive cluster (greater power for successful stops; $p = 0.039$) with a frontal-central topography, ranging from 200 to 400 ms, in β range (18-30 Hz). Additional significant clusters were a positive cluster (7-15 Hz, 420-750 ms, right frontal-central topography; $p = 0.002$) and a later negative β cluster (15-30 Hz, 750-1000 ms, left posterior topography; $p = 0.008$). The finding of the significant frontal-central β cluster replicated previous results (Wagner et al., 2018; Castiglione et al., 2019), and in conjunction with the P3 results, suggested participants engaged inhibitory processes during the stop signal task.

Dipole fitting results

The topographies of scalp EEG can be difficult to interpret and influenced by reference and processing decisions; and while the previously identified β cluster was somewhat localized to the right front of the scalp, it was distributed centrally as well. Thus, we next sought to determine that the engagement of beta band activity for successful Stop trials compared with unsuccessful Stop trials was localized to the right frontal lobe. Independent component analysis was run on the Stop Signal data for each participant, and equivalent current dipoles were fit to right frontal ICs. Time-frequency decomposition was performed on the IC source activity to determine whether a similar β increase for successful Stop trials was identified at the source level. The results of this analysis are presented in Figure 4. F4

The average scalp topography of the selected ICs, as well as the dipole localizations, are presented in Figure 4*A*. Average ERSPs of the source-level activity showed an increase in beta band activity ranging from ~250 to 500 ms, in β range (19-29 Hz), visualized in Figure 4*B*. Cluster-based permutation statistics identified this cluster as statistically significant ($p = 0.006$). Thus, we replicated the beta band increase for successful Stops compared with unsuccessful Stops, and that this effect was localized to the right frontal cortex.

PLS brain-behavior relationships

Given that we observed successful stopping in the Stop Signal task, successful DF in the DF task, and we replicated previous neural results in our analysis of successful versus unsuccessful stopping. We next examined relationships between the neural data in the Stop Signal task and behavioral performance in both tasks to determine whether individual differences in inhibitory processing were related to behavior. To this end, we used behavioral PLS combined with permutation statistics to identify LVs in the neural data (Successful – Unsuccessful Stop) that explained

J_ID: JNEURO Date: 14-February-23 4/Color Figure(s): "F1-F6" ART: JN-RM-0696-22 Page: 8 Total Pages: 14 Comments:

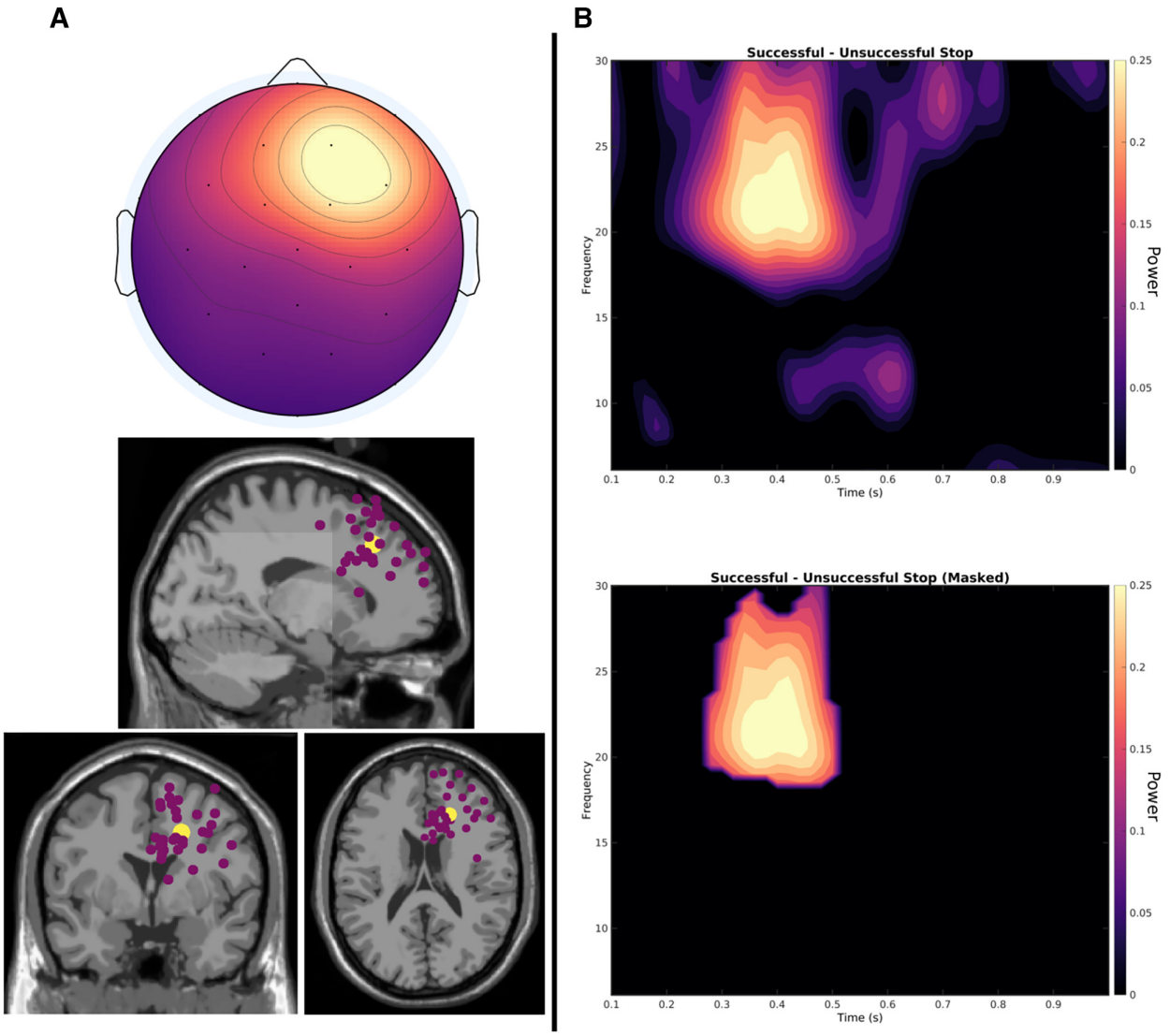

Figure 4. Dipole fitting results. ***A***, Top, Scalp topography showing the averaged scalp projection map of the selected right-frontal ICs across participants. Bottom, Equivalent dipole locations fit to the right frontal ICs for each participant, plotted in the MNI brain. Purple spheres represent participant dipoles. Yellow sphere represents the centroid. ***B***, Top, Average ERSPs of successful versus unsuccessful stops on the IC source-level activity. Bottom, Same as in top plot, but statistically masked to highlight the significant β cluster.

covariance between oscillatory power and behavioral performance. The results of this analysis are presented in Figure 5.

The behavioral PLS analysis resulted in one significant LV (i.e., the singular value of the LV was greater than the permutation distribution; $p = 0.025$). This variable alone explained 71.2% of the covariance between neural oscillatory power and behavioral performance. The significant bins within the time-frequency channel (i.e., the bootstrapped standard ratio at these bins were >1.96) are shown in Figure 4*A*. Remarkably, a cluster very similar to the right frontal-central β cluster that was identified in the time-frequency analysis of the stop signal data were identified by the behavioral PLS analysis as significantly covarying with behavioral performance. The cluster extended from ~160 to 360 ms, and from 18 to 30 Hz in frequency. We note that the entire time-frequency channel data matrix was entered into the behavioral PLS analysis; namely, the cluster was identified in a data-driven fashion and not biased to identify the frontal β cluster in any way.

To specify the relationship between neural activity in this cluster and behavioral performance, average oscillatory power (Successful – Unsuccessful Stop) was extracted from the right frontal-central β cluster identified by the PLS analysis for each participant and correlated with behavioral scores. The same robust correlation methods used for relating behavioral outcomes were used to relate neural activity to behavior. Bivariate outliers were identified in the correlations, and thus skipped Spearman correlations between cluster power and behavioral performance were reported. Oscillatory beta power and SSRTs were significantly negatively correlated ($r_{(31)} = -0.44$, $t = -2.77$, $p = 0.01$), suggesting greater engagement of frontal β mechanisms improved stopping reaction times. Frontal beta power was also significantly positively correlated with magnitude of DF ($r_{(31)} = 0.48$, $t = 3.04$, $p = 0.005$), suggesting that greater engagement of this frontal activity was related to successful suppression of encoding. Importantly, beta power was not significantly correlated with thought substitution magnitude ($r_{(31)} = 0.01$, $t = 0.01$, $p = 0.99$), and the DF and thought

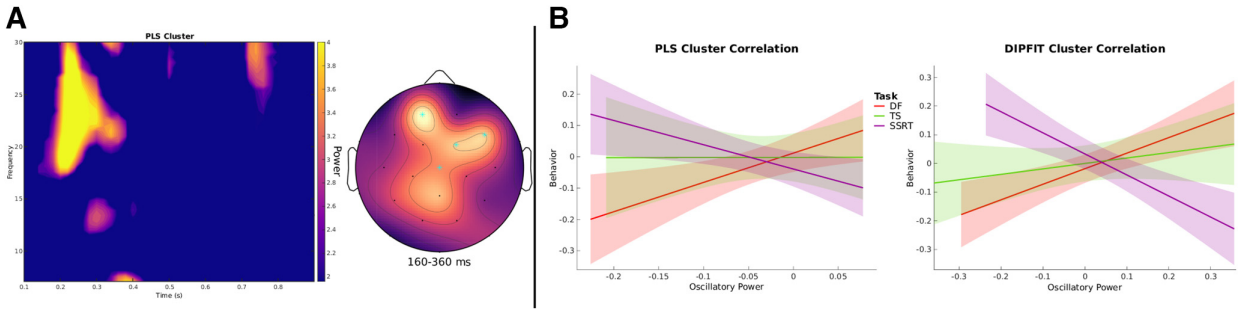

Figure 5. PLS results. ***A***, Left, Time-frequency plot of PLS cluster at frontal channel cluster, showing beta activity. Right, Topography of PLS cluster from 18-30 Hz and 160-360 ms, showing frontal cluster. ***B***, Left, Subject-level correlations between behavioral scores (SSRT, DF Magnitude, and TS Magnitude) and oscillatory power extracted from identified PLS cluster. Right, Same as in left, but oscillatory power was extracted from the dipole source-level oscillatory power.

substitution correlations significantly differed (Steiger's $z = 3.54$, $p < 0.001$; Zou's 95% CI = 0.214-0.717). We additionally performed Bayesian correlation analyses to determine the Bayes factors of the correlations between neural activity and behavior. These analyses reported moderately strong evidence against the null hypothesis for the correlations between neural activity and DF ($BF_{10} = 5.77$), as well as SSRTs ($BF_{10} = 3.88$). In contrast, the correlation between oscillatory power and thought substitution showed evidence in favor of the null hypothesis ($BF_{10} = 0.39$). In sum, engagement of frontal β mechanisms following stop signals was related to SSRTs and suppression of encoding, but was unrelated to forgetting because of thought substitution.

Given that the PLS analysis converged with the time-frequency analysis of the Stop Signal data, and we replicated the localization of this activity to a right frontal source using dipole fitting, we sought to determine whether the same relationships to behavior would be found using the source-level oscillatory activity. We extracted average oscillatory power from the source-level IC activity for each subject (Successful – Unsuccessful Stop), focused on the significant cluster found by the cluster-based permutation tests, and correlated this activity with behavioral scores. The skipped Spearman correlations between oscillatory activity and DF ($r_{(31)} = 0.61$, $t = 4.32$, $p < 0.001$), as well as SSRTs ($r_{(31)} = -0.59$, $t = -4.10$, $p < 0.001$) were significant. In contrast, the correlation between oscillatory activity and thought substitution was not ($r_{(31)} = 0.20$, $t = 1.11$, $p = 0.14$), and the DF and thought substitution correlations significantly differed (Steiger's $z = 3.32$, $p < 0.001$; Zou's 95% CI = 0.173-0.679). Additionally, the Bayesian correlation analyses reported strong evidence against the null hypothesis for the correlation between oscillatory power and DF ($BF_{10} = 26.96$), as well as SSRTs ($BF_{10} = 127.46$), whereas the Bayes factor for thought substitution showed evidence for the null hypothesis ($BF_{10} = 0.61$). These results corroborate the PLS results and provide even stronger evidence that engagement of frontal β mechanisms predicted faster stopping in the stop signal task and more effective suppression in the DF task, but was not predictive of thought substitution magnitude.

Cross-task neural decoding

The behavioral PLS results showed that engagement of neural inhibitory processes during the Stop Signal task was related to successful encoding suppression. However, this result by itself does not directly demonstrate that the same process or pattern of neural activity is engaged when a Forget cue is presented, only that individuals who showed greater engagement of frontal β mechanisms during the Stop Signal task also showed more successful suppression of encoding. To specifically examine the recruitment of the same frontal mechanisms during the DF task, we performed a cross-task neural decoding analysis, in which an LDA classifier was trained to discriminate successful versus unsuccessful stopping in the Stop Signal task using neural oscillatory features in the 160-360 ms time window (where the time-frequency and behavioral PLS analyses identified differences between these conditions). This classifier was then tested to discriminate successful versus unsuccessful forgetting following Forget cues, as well as Imagine cues, at each time point following the cue. Compared with fMRI, neural decoding techniques using EEG lack spatial specificity, but have the strength of identifying when in time neural mechanisms may be engaged. Therefore, this analysis not only allowed us to test whether suppression of encoding or thought substitution recruited the same neural mechanisms as action inhibition, but also identify when this recruitment occurred. Importantly, this analysis was conducted within-subjects and at the single-trial level, allowing for greater evidence of shared neural recruitment beyond subject-level or block-level decoding.

The time series of classifier accuracies for the first classifier analysis, in which classifiers were trained on neural data to discriminate Successful Stops from Unsuccessful Stops, are presented in Figure 6*A*, along with the topographic plot of the transformed classifier weights (Haufe et al., 2014). Cluster-based permutation testing was used to determine time windows of significantly above-chance classifier accuracy. A significant cluster was identified when decoding memory outcome following Forget cues ($p = 0.012$). The cluster spanned ~420-560 ms. In contrast, no significant clusters were found when decoding memory accuracy following Imagine cues. While the importance map of the classifier weights was broadly distributed, it showed some resemblance to the PLS cluster identified. A cluster-based permutation test comparing classifier accuracies across cue conditions identified two significant small clusters in the same time window, separated by 2 time points ($p = 0.045$; $p = 0.037$), demonstrating significantly greater classification accuracy following Forget cues compared with Imagine cues. A signed rank test comparing average classifier accuracies between conditions from 420-560 ms also significantly identified greater accuracy following Forget cues ($z = 3.01$, $p = 0.002$). F6

To ensure that the successful classification was not simply because of identification of error signals following Unsuccessful Stops and unsuccessful Directed Forgetting, a second classifier analysis was performed in which classifiers were trained to discriminate Successful Stops from Successful Go trials. The results of this analysis are shown in Figure 6*B*. A significant cluster, similar to the cluster found in the previous analysis, was identified when decoding memory outcome following

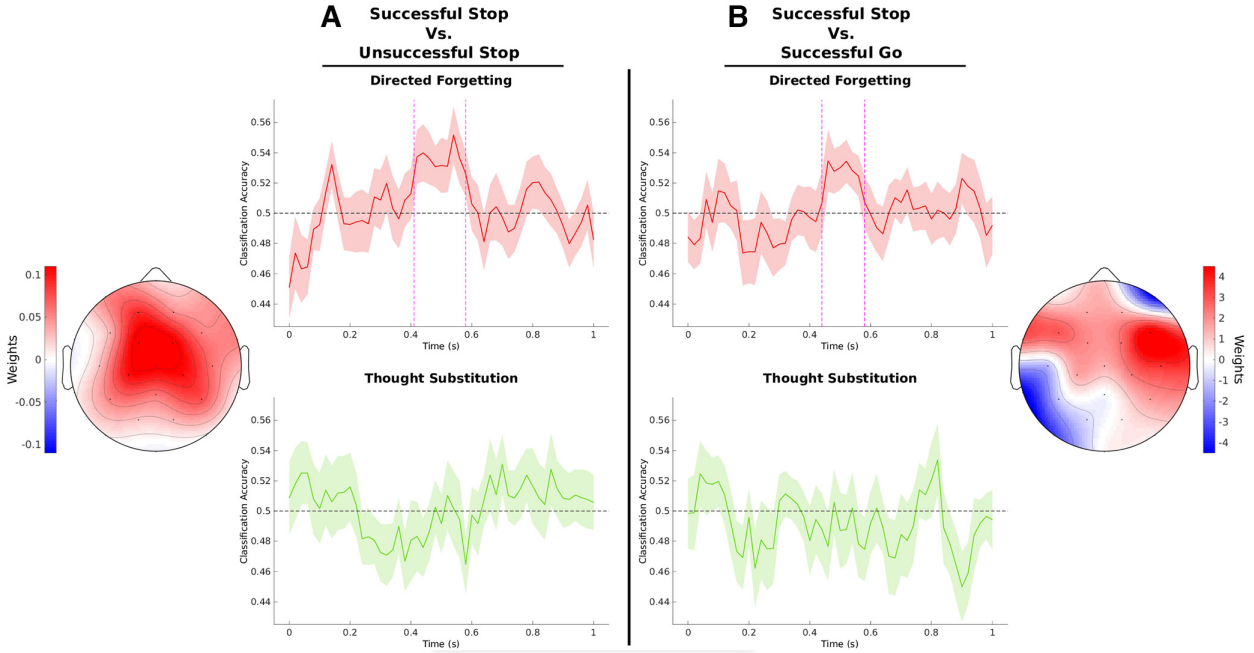

Figure 6. Cross-task decoding results. LDA classifiers trained to discriminate successful versus unsuccessful stopping in the stop signal task were used to discriminate successful versus unsuccessful forgetting following memory cues. Transparent shading represents SE. Topographic maps represent classifier weights. ***A***, Time series of classifier accuracies for classifiers trained to discriminate successful versus unsuccessful Stop trials. ***B***, Same as in ***A***, but for classifiers trained to discriminate successful Stop versus successful Go trials.

Forget cues, although accuracy was lower and the cluster was shorter in time ($p = 0.035$). This is potentially because of the difference in classifier weights, as the topographic map showed large differences in importance from the initial analysis, or potentially because of the procedure to align trials in time leading to some inaccuracies in trial alignment. As in the first analysis, no significant clusters were found when decoding memory outcome following Imagine cues. A cluster-based permutation test comparing classifier accuracies across cue conditions did not identify a significant cluster; however, a signed rank test comparing average classifier accuracies between conditions from 420-560 ms also significantly identified greater accuracy following Forget cues ($z = 2.23$, $p = 0.027$). Thus, classifiers trained on neural data from the Stop Signal task could successfully classify memory outcomes when given neural data following Forget cues, but could not successfully classify memory outcomes when given neural data following Imagine cues, suggesting that direct suppression of encoding recruits frontal inhibitory mechanisms, whereas thought substitution does not.

Discussion

In the current study, we directly tested the degree to which different strategies of intentional forgetting during encoding, namely, direct suppression and thought substitution, engage top-down inhibitory control mechanisms to produce forgetting. Recent research has generally converged on the notion that suppression of retrieval recruits the same inhibitory control mechanism, indexed by a right frontal β oscillatory signal, as suppression of other actions (Castiglione et al., 2019; Apšvalka et al., 2022). While similar signals have been observed following DF cues (Rizio and Dennis, 2013; Oehrn et al., 2018; Hubbard and Sahakyan, 2021), to interpret this as inhibitory control would reflect a fallacious reverse inference, as right frontal activity may reflect many different cognitive processes (Poldrack, 2006). Additionally, few studies have investigated the neural mechanisms underlying thought substitution, and no study to date has probed differences in engagement of inhibitory control mechanisms between these strategies. The present data support the view that direct suppression of mnemonic encoding engages domain-general inhibitory mechanisms, while forgetting mediated by thought substitution does not.

Three main points of new evidence from out study corroborate this view. First, behaviorally, participants who showed better engagement of inhibition in the Stop Signal task (i.e., faster SSRTs) also showed greater magnitude of forgetting because of encoding suppression, but did not show a greater magnitude of thought substitution-mediated forgetting. Importantly, this result was unlikely to be driven by individual differences in generic task performance or attention (Goodhew and Edwards, 2019), as this explanation would be unable to explain the observed interaction in the correlations, and follow-up analyses using partial correlations demonstrated that the relationship between SSRTs and encoding suppression remained after controlling for the thought substitution magnitude. Second, behavioral PLS identified that β oscillatory activity over right-frontal and central sites was related to both speed of stopping during the Stop Signal task, as well as magnitude of encoding suppression in the memory task, but was unrelated to the magnitude of forgetting following thought substitution. This was corroborated by a dipole fitting analysis that identified a right frontal source for the β oscillatory activity which was also related to speed of stopping and magnitude of encoding suppression. Finally, a cross-task neural decoding analysis revealed that classifiers trained on neural data to discriminate successful versus unsuccessful stopping could also predict successful forgetting following Forget cues, but were unable to predict forgetting

following Imagine cues. These results not only demonstrate that suppression during encoding recruits domain-general inhibition, but also that while thought substitution is an effective strategy for forgetting, it relies on different neural mechanisms than frontally mediated inhibition.

These results strongly argue against the selective rehearsal explanation of item-method DF, which posits that Remember cues lead to additional processing of items, while Forget cues simply do not, and memory for these items passively decays (MacLeod, 1975; Basden et al., 1993). Instead, our results are in line with work suggesting that DF at encoding is an active process (Anderson and Hanslmayr, 2014), involving active stopping of mnemonic encoding processes to produce forgetting. This active stopping is likely accomplished by engagement of the right frontal cortex (Wylie et al., 2008; Rizio and Dennis, 2013; Wang et al., 2019). Indeed, studies incorporating brain stimulation techniques in a list-method DF paradigm have identified a role of the right PFC in successful forgetting (Silas and Brandt, 2016; Imbernón et al., 2022). Here, we expand on previous studies, and are the first study to directly tie this frontal activity to top-down inhibitory mechanisms with cross-task analyses. While EEG scalp topographies are not strictly indicative of specific neural generators, our dipole fitting analysis suggested that the observed β oscillatory activity was localized to the right frontal cortex and was related to both motor inhibition and DF success, a finding consistent with research implicating the right PFC in domain-general inhibitory control (Aron et al., 2004, 2014; Depue et al., 2016; Wessel and Aron, 2017). In sum, our results support the notion that the right PFC is recruited when inhibitory control is required, whether it be stopping a motor action, retrieval of a memory, or mnemonic encoding.

Our results also provide novel evidence that the onset of inhibitory control following a Forget cue occurs later than following a cue to cease a motor action. Whereas motor inhibition engaged right frontal mechanisms within 200 ms, intentional forgetting engaged this mechanism ~400 ms after the onset of the cue. This differed from studies of retrieval suppression using the TNT paradigm, where engagement of frontal mechanisms seemed to occur at roughly the same time across tasks (Castiglione et al., 2019). One potential explanatory difference between our results and the TNT results is that inhibition of retrieval of already studied information may occur more rapidly than inhibition of the mnemonic encoding process. Another potential difference is that, in the TNT task, retrieval suppression of a particular item occurs multiple times within the experiment. Over repeated trials, the need for inhibitory control is reduced, and activation in right PFC declines (Kuhl et al., 2007; Wimber et al., 2015; Apšvalka et al., 2022). This could potentially lead to faster engagement of inhibition over time, although this has not been demonstrated empirically. In our study, we measured inhibition of encoding at the first encounter of the item, and thus engagement of inhibition may have taken longer. Future studies could potentially use brain stimulation, such as transcranial magnetic stimulation, over right frontal sites at different time points following a Forget cue to more accurately identify the timing of engagement of inhibition, or specifically examine bursts of beta activity, as was recently investigated in a Stop Signal task (Hannah et al., 2020).

We found little evidence that thought substitution cues elicited frontal inhibition mechanisms, although it is of course possible that thought substitution does recruit some form of inhibition, but to a much lesser extent than direct suppression that we were unable to detect in the current study. Interestingly, similar rates of forgetting were still observed following these cues compared with Forget cues, although Forget cues were more effective than Imagine cues. What, then, is the mechanism of forgetting engaged by thought substitution? One likely possibility is that thought substitution produces a shift in context during encoding of the items, leading to later forgetting. During encoding, representations of mental context, which can include external environmental features as well as internal mental thoughts, become associated with items and can be used as a cue to retrieve items from memory (Howard and Kahana, 2002; Polyn et al., 2009). If the mental context is shifted from the ongoing experimental context, then retrieval of items when the mental shift occurred may be impacted. Indeed, in list-method DF studies, instituting a mental context change instead of a direct instruction to forget induces similar rates of forgetting (Sahakyan and Kelley, 2002; Delaney et al., 2010), and we demonstrated that this can occur at the item level as well (Hubbard and Sahakyan, 2021). Recent theoretical advances suggest context may indeed shift, creating event boundaries and segmentation of ongoing experience into episodes (DuBrow et al., 2017), which may occur following thought substitution. Future fMRI work comparing item-method DF and thought substitution may elucidate the specific mechanisms that underlie this strategy.

Our results could have important clinical implications, and may provide greater understanding to research involving DF in special populations. For instance, older adults generally show reduced effects of DF cues compared with younger adults (Titz and Verhaeghen, 2010), but exhibit larger forgetting effects when a thought substitution instruction is given instead (Sahakyan et al., 2008). Older adults also show higher SSRTs than younger adults on the Stop Signal task (Kramer et al., 1994; Rush et al., 2006) and less right frontal cortical engagement following stop signals (Coxon et al., 2016). Similarly, patients with schizophrenia also exhibit longer SSRTs and reduced right frontal activity following stop signals (Hughes et al., 2012), but only individuals at risk for schizophrenia with positive schitzotypy symptoms showed reduced effects of forget cues in an item-method DF experiment (Sahakyan et al., 2020). These patterns of results suggest that deficits in domain-general inhibitory processing, mediated by the right frontal cortex, can lead to reduced efficacy of direct suppression of encoding. However, individuals in these populations may retain the ability to successfully forget information using a thought substitution strategy. Directly comparing the success of these strategies in different populations may not only further elucidate the mechanisms underlying methods of intentional forgetting, but also highlight which strategies are impaired versus retained, potentially leading to more successful clinical interventions down the line.

References

Anderson MC, Green C (2001) Suppressing unwanted memories by executive control. Nature 410:366–369.

Anderson MC, Hanslmayr S (2014) Neural mechanisms of motivated forgetting. Trends Cogn Sci 18:279–292.

Anderson MC, Hulbert JC (2021) Active forgetting: adaptation of memory by prefrontal control. Annu Rev Psychol 72:1–36.

Anderson MC, Bunce JG, Barbas H (2016) Prefrontal–hippocampal pathways underlying inhibitory control over memory. Neurobiol Learn Mem 134:145–161.

Anderson MC, Ochsner KN, Kuhl B, Cooper J, Robertson E, Gabrieli SW, Glover GH, Gabrieli JDE (2004) Neural systems underlying the suppression of unwanted memories. Science 303:232–235.

Apšvalka D, Ferreira CS, Schmitz TW, Rowe JB, Anderson MC (2022) Dynamic targeting enables domain-general inhibitory control over action and thought by the prefrontal cortex. Nat Commun 13:274.

Aron AR, Robbins TW, Poldrack RA (2004) Inhibition and the right inferior frontal cortex. Trends Cogn Sci 8:170–177.

Aron AR, Robbins TW, Poldrack RA (2014) Inhibition and the right inferior frontal cortex: one decade on. Trends Cogn Sci 18:177–185.

Band GP, Van Der Molen MW, Logan GD (2003) Horse-race model simulations of the stop-signal procedure. Acta Psychol (Amst) 112:105–142.

Basden BH, Basden DR, Gargano GJ (1993) Directed forgetting in implicit and explicit memory tests: a comparison of methods. J Exp Psychol 19:603–616.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Soft 67:1–48.

Benoit RG, Anderson MC (2012) Opposing mechanisms support the voluntary forgetting of unwanted memories. Neuron 76:450–460.

Bergström ZM, de Fockert JW, Richardson-Klavehn A (2009) ERP and behavioural evidence for direct suppression of unwanted memories. Neuroimage 48:726–737.

Bjork RA (1970) Positive forgetting: the noninterference of items intentionally forgotten. J Verbal Learn Verbal Behav 9:255–268.

Bjork RA (1972) Theoretical implications of directed forgetting. In: Coding processes in human memory(Melton AW, Martin E, eds), pp 217–325. Washington, DC: Winston.

Bjork RA, Laberge D, Legrand R (1968) The modification of short-term memory through instructions to forget. Psychon Sci 10:55–56.

Castiglione A, Wagner J, Anderson M, Aron AR (2019) Preventing a thought from coming to mind elicits increased right frontal beta just as stopping action does. Cereb Cortex 29:2160–2172.

Chan AM, Halgren E, Marinkovic K, Cash SS (2011) Decoding word and category-specific spatiotemporal representations from MEG and EEG. Neuroimage 54:3028–3039.

Chevrier AD, Noseworthy MD, Schachar R (2007) Dissociation of response inhibition and performance monitoring in the stop signal task using event-related fMRI. Hum Brain Mapp 28:1347–1358.

Chiu YC, Egner T (2015a) Inhibition-induced forgetting: when more control leads to less memory. Psychol Sci 26:27–38.

Chiu YC, Egner T (2015b) Inhibition-induced forgetting results from resource competition between response inhibition and memory encoding processes. J Neurosci 35:11936–11945.

Ciuparu A, Mureşan RC (2016) Sources of bias in single-trial normalization procedures. Eur J Neurosci 43:861–869.

Coxon JP, Goble DJ, Leunissen I, Van Impe A, Wenderoth N, Swinnen SP (2016) Functional brain activation associated with inhibitory control deficits in older adults. Cereb Cortex 26:12–22.

Delaney PF, Sahakyan L, Kelley CM, Zimmerman CA (2010) Remembering to forget: the amnesic effect of daydreaming. Psychol Sci 21:1036–1042.

Depue BE, Curran T, Banich MT (2007) Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. Science 317:215–219.

Depue BE, Orr JM, Smolker HR, Naaz F, Banich MT (2016) The organization of right prefrontal networks reveals common mechanisms of inhibitory regulation across cognitive, emotional, and motor processes. Cereb Cortex 26:1634–1646.

Diedenhofen B, Musch J (2015) cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 10:e0121945.

DuBrow S, Rouhani N, Niv Y, Norman KA (2017) Does mental context drift or shift? Curr Opin Behav Sci 17:141–146.

Fawcett JM, Taylor TL (2010) Directed forgetting shares mechanisms with attentional withdrawal but not with stop-signal inhibition. Mem Cognit 38:797–808.

Gagnepain P, Hulbert J, Anderson MC (2017) Parallel regulation of memory and emotion supports the suppression of intrusive memories. J Neurosci 37:6423–6441.

Goodhew SC, Edwards M (2019) Translating experimental paradigms into individual-differences research: contributions, challenges, and practical recommendations. Conscious Cogn 69:14–25.

Grootswagers T, Wardle SG, Carlson TA (2017) Decoding dynamic brain patterns from evoked responses: a tutorial on multivariate pattern analysis applied to time series neuroimaging data. J Cogn Neurosci 29:677–697.

Hannah R, Muralidharan V, Sundby KK, Aron AR (2020) Temporally-precise disruption of prefrontal cortex informed by the timing of beta bursts impairs human action-stopping. Neuroimage 222:117222.

Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110.

Hertel PT, Calcaterra G (2005) Intentional forgetting benefits from thought substitution. Psychonom Bull Rev 12:484–489.

Hourihan KL, Taylor TL (2006) Cease remembering: control processes in directed forgetting. J Exp Psychol 32:1354.

Howard MW, Kahana MJ (2002) A distributed representation of temporal context. J Math Psychol 46:269–299.

Hubbard RJ, Sahakyan L (2021) Separable neural mechanisms support intentional forgetting and thought substitution. Cortex 142:317–331.

Hubert M, Rousseeuw PJ, van Aelst S (2008) High-breakdown robust multivariate methods. Statist Sci 92–119.

Hughes ME, Fulham WR, Johnston PJ, Michie PT (2012) Stop-signal response inhibition in schizophrenia: behavioural, event-related potential and functional neuroimaging data. Biol Psychol 89:220–231.

Imbernón JJ, Aguirre C, Gómez-Ariza CJ (2022) Selective directed forgetting is mediated by the lateral prefrontal cortex: preliminary evidence with transcranial direct current stimulation. Cogn Neurosci 13:77–86. AQ:G

Jarosz AF, Wiley J (2014) What are the odds? A practical guide to computing and reporting Bayes factors. J Problem Solving 7:2.

Kiesel A, Miller J, Jolicóur P, Brisson B (2008) Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring methods. Psychophysiology 45:250–274.

Kim H, Smolker HR, Smith LL, Banich MT, Lewis-Peacock JA (2020) Changes to information in working memory depend on distinct removal operations. Nat Commun 11:1–14.

King JR, Dehaene S (2014) Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn Sci 18:203–210.

Kim S (2015) ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Commun Stat Appl Methods 22:665–674.

Kok A, Ramautar JR, de Ruiter MB, Band GP, Ridderinkhof KR (2004) ERP components associated with successful and unsuccessful stopping in a stop-signal task. Psychophysiology 41:9–20.

Kramer AF, Humphrey DG, Larish JF, Logan GD (1994) Aging and inhibition: beyond a unitary view of inhibitory processing in attention. Psychol Aging 9:491–512.

Krishnan A, Williams LJ, McIntosh AR, Abdi H (2011) Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage 56:455–475.

Kuhl BA, Dudukovic NM, Kahn I, Wagner AD (2007) Decreased demands on cognitive control reveal the neural processing benefits of forgetting. Nat Neurosci 10:908–914.

Linde-Domingo J, Treder MS, Kerrén C, Wimber M (2019) Evidence that neural information flow is reversed between object perception and object reconstruction from memory. Nat Commun 10:1–13.

Logan GD, Cowan WB (1984) On the ability to inhibit thought and action: a theory of an act of control. Psychol Rev 91:295–327.

Luck SJ (2014) An introduction to the event-related potential technique. Cambridge, MA: Massachusetts Institute of Technology.

Ly A, Verhagen J, Wagenmakers EJ (2016) Harold Jeffreys's default Bayes factor hypothesis tests: explanation, extension, and application in psychology. J Math Psychol 72:19–32.

MacLeod CM (1975) Long-term recognition and recall following directed forgetting. J Exp Psychol 1:271.

Makeig S (1993) Auditory event-related dynamics of the EEG spectrum and effects of exposure to tones. Electroencephalogr Clin Neurophysiol 86:283–293.

Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG-and MEG-data. J Neurosci Methods 164:177–190.

McIntosh AR, Lobaugh NJ (2004) Partial least squares analysis of neuroimaging data: applications and advances. Neuroimage 23:S250–S263.

Miller J, Ulrich R, Schwarz W (2009) Why jackknifing yields good latency estimates. Psychophysiology 46:300–312.

Nieuwenhuis S, Forstmann BU, Wagenmakers EJ (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. Nat Neurosci 14:1105–1107.

Oehrn CR, Fell J, Baumann C, Rosburg T, Ludowig E, Kessler H, Hanslmayr S, Axmacher N (2018) Direct electrophysiological evidence for prefrontal control of hippocampal processing during voluntary forgetting. Curr Biol 28:3016–3022.e4.

Oostenveld R, Oostendorp TF (2002) Validating the boundary element method for forward and inverse EEG computations in the presence of a hole in the skull. Hum Brain Mapp 17:179–192.

Palmer JA, Kreutz-Delgado K, Makeig S (2012) AMICA: an adaptive mixture of independent component analyzers with shared components. Swartz Center for Computatonal Neursoscience. University of California San Diego, Tech. Rep.

Pernet CR, Wilcox RR, Rousselet GA (2013) Robust correlation analyses: false positive and power validation using a new open source matlab toolbox. Front Psychology 3:606.

Poldrack RA (2006) Can cognitive processes be inferred from neuroimaging data? Trends Cogn Sci 10:59–63.

Polyn SM, Norman KA, Kahana MJ (2009) A context maintenance and retrieval model of organizational processes in free recall. Psychol Rev 116:129–156.

Rizio AA, Dennis NA (2013) The neural correlates of cognitive control: successful remembering and intentional forgetting. J Cogn Neurosci 25:297–312.

Rousseeuw PJ (1984) Least median of squares regression. J Am Statist Assoc 79:871–880.

Rousseeuw PJ, van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223.

Rousselet GA, Pernet CR (2012) Improving standards in brain-behavior correlation analyses. Front Hum Neurosci 6:119.

Rush BK, Barch DM, Braver TS (2006) Accounting for cognitive aging: context processing, inhibition or processing speed? Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 13:588–610.

Sahakyan L (2022) Current perspectives on directed forgetting. In: Oxford handbook of human memory(WagnerA, Kahana M, eds). Oxford: Oxford UP.

Sahakyan L, Kelley CM (2002) A contextual change account of the directed forgetting effect. J Exp Psychol 28:1064.

Sahakyan L, Delaney PF, Goodmon LB (2008) Oh, honey, I already forgot that: strategic control of directed forgetting in older and younger adults. Psychol Aging 23:621–633.

Sahakyan L, Delaney PF, Foster NL, Abushanab B (2013) List-method directed forgetting in cognitive and clinical research: a theoretical and methodological review. In: Psychology of learning and motivation, Vol 59, pp 131–189. San Diego: Academic.

Sahakyan L, Kwapil TR, Jiang L (2020) Differential impairment of positive and negative schizotypy in list-method and item-method directed forgetting. J Exp Psychol 149:368–381.

~~Schindler S, Kissler J (2018) Too hard to forget? ERPs to remember, forget, and uninformative cues in the encoding phase of item-method directed forgetting. Psychophysiology 55:e13207.~~

Silas J, Brandt KR (2016) Frontal transcranial direct current stimulation (tDCS) abolishes list-method directed forgetting. Neurosci Lett 616:166–169.

Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87:245–251.

Swann N, Tandon N, Canolty R, Ellmore TM, McEvoy LK, Dreyer S, DiSano M, Aron AR (2009) Intracranial EEG reveals a time-and frequency-specific role for the right inferior frontal gyrus and primary motor cortex in stopping initiated responses. J Neurosci 29:12675–12685.

Tanner D, Morgan-Short K, Luck SJ (2015) How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. Psychophysiology 52:997–1009.

Titz C, Verhaeghen P (2010) Aging and directed forgetting in episodic memory: a meta-analysis. Psychol Aging 25:405–411.

Ulrich R, Miller J (2001) Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. Psychophysiology 38:816–827.

Verbruggen F, et al. (2019) A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. eLife 8: e46323.

Verbruggen F, Logan GD (2009) Models of response inhibition in the stop-signal and stop-change paradigms. Neurosci Biobehav Rev 33:647–661.

Wagner J, Wessel JR, Ghahremani A, Aron AR (2018) Establishing a right frontal beta signature for stopping action in scalp EEG: implications for testing inhibitory control in other task contexts. J Cogn Neurosc 30:107–118.

Wang TH, Placek K, Lewis-Peacock JA (2019) More is less: increased processing of unwanted memories facilitates forgetting. J Neurosci 39:3551–3560.

Wessel JR, Aron AR (2015) It's not too late: the onset of the frontocentral P3 indexes successful response inhibition in the stop-signal paradigm. Psychophysiology 52:472–480.

Wessel JR, Aron AR (2017) On the globality of motor suppression: unexpected events and their influence on behavior and cognition. Neuron 93:259–280.

Wimber M, Alink A, Charest I, Kriegeskorte N, Anderson MC (2015) Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. Nat Neurosci 18:582–589.

Wylie GR, Foxe JJ, Taylor TL (2008) Forgetting as an active process: an fMRI investigation of item-method–directed forgetting. Cereb Cortex 18:670–682.

Zou GY (2007) Toward using confidence intervals to compare correlations. Psychol Methods 12:399–413.

AQ:H

AQ:I

AQ:J