HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 July 15.

Published in final edited form as:

Neuroimage. 2021 July 15; 235: 117983. doi:10.1016/j.neuroimage.2021.117983.

Separation of item and context in item-method directed forgetting

Yi-Chieh Chiu[#a,b], **Tracy H. Wang**[#c], **Diane M. Beck**[a,b], **Jarrod A. Lewis-Peacock**[c], **Lili Sahakyan**[a,b,*]

[a]Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, United States

[b]Beckman Institute for Advanced Science and Technology, United States

[c]Department of Psychology, University of Texas at Austin, United States

[#] These authors contributed equally to this work.

Abstract

Contextual information plays a critical role in directed forgetting (DF) of lists of items, whereas DF of individual items has been primarily associated with item-level processing. This study was designed to investigate whether context processing also contributes to the forgetting of individual items. Participants first viewed a series of words, with task-irrelevant scene images (used as "context tags") interspersed between them. Later, these words reappeared without the scenes and were followed by an instruction to remember or forget that word. Multivariate pattern analyses of fMRI data revealed that the reactivation of context information associated with the studied words (i.e., scene-related activity) was greater whereas the item-related information diminished after a forget instruction compared to a remember instruction. Critically, we found the magnitude of the separation between item information and context information predicted successful forgetting. These results suggest that the unbinding of an item from its context may support the intention to forget, and more generally they establish that contextual processing indeed contributes to item-method DF.

[*]Corresponding author at: Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel St., Champaign, IL 61820, United States. LSAHAKY@illinois.edu (L. Sahakyan).

Credit author statement

Yi-Chieh Chiu collected and analyzed data and contributed to the initial draft of the mansucript.
Tracy H. Wang analyzed data and contributed to manuscript preparation.
Diane M. Beck and Jarrod A. Lewis-Peacock contributed to study design and manuscript preparation.
Lili Sahakyan was the senior investigator and oversaw all aspects of the project.

Declaration of Competing Interest

None.

Data and code availability statement

The behavioral raw trial data and the fMRI data can be accessed on Open Science Framework (https://osf.io/nh3ra/).
The analysis code for this project is available from GitHub repository (https://github.com/LewisPeacockLab/EmoDiF).

Ethics statement

The study was approved by the Institutional Review Board of the University of Illinois (IRB # 18610) and all subjects gave informed written consent in accordance with the Declaration of Helsinki.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi: 10.1016/j.neuroimage.2021.117983.

Keywords

Directed forgetting; Context; Episodic memory; fMRI

1. Introduction

To most people, forgetting is a negative experience that is rarely done on purpose, and often considered a human frailty to be avoided. Yet forgetting is often precisely what we need to do in order to remove outdated or irrelevant information from memory, such as an old password or the hotel room number where we recently stayed. In the extreme, we sometimes have unpleasant or even traumatic experiences that we would prefer to forget. In these circumstances, forgetting can be adaptive (Bjork, 1989, 2011).

Decades of research using a directed forgetting (DF) paradigm (Bjork et al., 1968) confirm that people have worse memory for items followed by a cue to forget (F) that item than a cue to remember (R) the item, suggesting that we can control our cognition voluntarily to impair access to the unwanted information (for reviews, see Anderson and Hanslmayr, 2014; MacLeod, 1998; Sahakyan et al., 2013; Sahakyan and Foster, 2016; Sahakyan, 2021). Whereas directed forgetting for whole lists of items is thought to involve shifts in contextual processing (Sahakyan and Kelley, 2002; Sahakyan et al., 2013), the traditional interpretation of directed forgetting of single items has emphasized passive processes that involve removing the *F* items from rehearsal in working memory (Bjork, 1970; Basden et al., 1993; MacLeod, 1999). According to this view, participants maintain an item in working memory until they receive the memory cue. The *F* cue leads participants to terminate rehearsal and remove the item from working memory, whereas the *R* cue leads participants to continue rehearsal and encode that item in a more elaborate way.

Current views of item-method DF, however, suggest that removal of items from working memory does not happen through passive decay, but rather requires engagement of active, effortful processes aimed at terminating encoding of *F* items (Fawcett and Taylor, 2008, 2010, 2012; Hauswald et al., 2010; Lee et al., 2013; Ludowig et al., 2010; Nowicka et al., 2010; Oberauer, 2018; Oehrn et al., 2018; Paz-Caballero et al., 2004; Reber et al., 2002; Rizio and Dennis, 2013; Wylie et al., 2008). For example, behavioral studies have shown that reaction times on a secondary task that is performed along with a DF task are slower during the execution of the *F* cue (Fawcett and Taylor, 2008, 2010, 2012). There is also reduced processing of other information that is presented in temporal or spatial proximity to *F* items (Taylor, 2005; Taylor and Fawcett, 2011, 2012; Thompson and Taylor, 2015), indicating that intentional forgetting engages cognitive load and is effortful. In addition, evidence from imaging and event-related potentials (ERPs) studies indicates that successful intentional forgetting (*F* items that are subsequently forgotten) recruits additional processes beyond those that are associated with unintentional forgetting (*R* items that are subsequently forgotten) (for a review, see Anderson and Hanslmayr, 2014). For example, during an attempt to forget, prefrontal and parietal regions are more active than during an attempt to remember an item, suggesting that successful forgetting recruits additional resources and may be more demanding (Paz-Caballero et al., 2004; Wylie et al., 2008; Van Hooff And

Ford, 2011; Rizio and Dennis, 2013). Connectivity analyses demonstrate that activity in the right dorsolateral prefrontal cortex (DLPFC) on *F* trials is associated with decreased activity in the left hippocampus, particularly during successful intentional forgetting, suggesting that right PFC exerts inhibitory control over encoding activity in medial temporal lobe (Rizio and Dennis, 2013; Ludowig et al., 2010; Oehrn et al., 2018). In our own recent work using the item-method DF task, we found that the representation of an item in temporal cortex was enhanced on *F* trials compared to *R* trials (Wang et al., 2019). The degree of enhancement was related to forgetting success, and this was explained as an increased attentional focus on the unwanted information that temporarily renders the neural representation vulnerable to memory weakening processes (Ritvo et al., 2019). Overall, growing evidence indicates that item-method DF engages an active process that inhibits ongoing encoding.

While a wealth of research has examined the mechanisms of item-method DF, the primary focus has been on the impairment of individual items, and less attention has been directed to the context that surrounds the items (i.e., the "setting" in which the items are encoded). Events do not take place in a vacuum; they unfold in certain temporal-spatial, social-emotional environments, and these background cues become associated with our memories for individual items. During retrieval, we rely on context cues to search and retrieve appropriate memories, and exclude inappropriate ones. Virtually all memory theories propose that a gradually changing 'mental context' is a critical component for understanding episodic memory (Bower, 1972; Dennis and Humphreys, 2001; Diana et al., 2007; Estes, 1955; Mensink and Raaijmakers, 1988; Howard and Kahana, 2002; Lehman and Malmberg, 2009; Sederberg et al., 2008; Polyn et al., 2009; Raaijmakers and Shiffrin, 1981). Given that memory for events include not only item representations, but also representations of the context in which the events occurred, intentional forgetting processes could be operating on either or both of these components.

The role of mental context is well established in list-method DF (e.g., Sahakyan and Kelley, 2002). In this paradigm, participants study a list of items followed by an *F* or *R* cue that applies to the entire list. A to-be-remembered second list is then studied, followed by a final memory test of items from both lists. According to the context account of list-method DF (Sahakyan et al., 2013; Manning et al., 2016), participants actively shift their mental context in response to an *F* cue, thus distancing themselves from the context of the to-be-forgotten list of items, and allowing themselves to encode the to-be-remembered list in a 'new' mental context. At the time of final test, the retrieval context mismatches the context of the to-be-forgotten list, which impairs retrieval of the context and items from this list. In the Manning et al. (2016) fMRI study, which was aimed at evaluating the context account at the neural level, participants performed a modified version of list-method DF. A first list of words was studied which had task-irrelevant images of scenes interspersed between the words. Following the *F* or *R* cue, participants studied a second list of words which did not contain any scenes. The scenes were used as 'context tags' (Gershman et al., 2013) to decode the mental context signal from the first list and track it afterwards. They observed that in response to the *F* cue, there was a reduction in the neural representation of the encoding context prior to study of the second list that predicted forgetting success. Given that downregulation of contextual information was observed in a list-method DF study, in this

study we sought to test the idea that modulation of contextual information may also contribute to the intentional forgetting of individual items.

In this experiment, participants first were exposed to a preview phase, where they saw a series of words with images of task-irrelevant scenes embedded between each word presentation. As in Manning et al. (2016), these scenes were used as context tags to identify and track the mental context of this study period. Note that the decoded context (i.e., scene-related neural activity) is not necessarily specific to individual trials, but rather reflects the global encoding context, which changes gradually in response to each presented item. During the subsequent DF phase, all words were re-presented without the accompanying scenes, and an *F* or *R* cue was assigned to each word individually. Given that no scenes were presented during the DF phase, any decoded scene information during this task could be interpreted as reflecting reinstatement of the mental context from the initial preview phase (Gershman et al., 2013).

Although there is considerable behavioral evidence that the mechanisms responsible for item-method DF differ from those of list-method DF (Sahakyan et al., 2013), this inference may be limited by experimental paradigms that prevent a more direct comparison. Measuring context processing separately from item processing has not been possible in item-method DF. This experiment was designed to do just that. To this end, independent data was used to train fMRI pattern classifiers to identify neural signatures of two types of information in ventral visual cortex – *item* information (word-related activity) and *context* information (scene-related activity). Furthermore, because some studies found that directed forgetting effects depend on emotional valence (e.g., Hauswald et al., 2010; Nowicka et al., 2010; but see Taylor et al., 2018; Yang et al., 2012), we used both negative and neutral words. The manipulation of item information and context information was quantified on each trial and related to subsequent recognition outcomes in the final memory test. It is possible that the to-be-forgotten information could be suppressed (as observed in Manning et al., 2016) or enhanced (as observed in Wang et al., 2019) during the attempted forgetting. Our basic hypothesis was that shifting, distancing, or inhibiting the context of to-be-forgotten information could enable successful DF of the item that was studied in that context. This manipulation of contextual information would render the context cues less effective at the time of the test at retrieving the item information or recognizing it as belonging to the study context.

2. Methods

2.1. Participants

Twenty-five right-handed participants between 18 and 35 years of age (11 Male, 14 Female, $M = 23$ yrs, SD = 2.78) were recruited from Champaign-Urbana area in Illinois. Data from one participant was excluded from analyses due to excessive motion during fMRI scanning. We initially examined the data from four pilot participants in order to evaluate the efficacy of the MVPA training phase (i.e. trial counts and stimuli presentation durations) to yield good classification of the stimuli categories. We additionally designed the analytic plan based on these data before performing exploratory analyses. We report the findings based on the $N = 20$ participants whose data were obtained after our analysis pipeline was complete in

Supplementary Fig. 4. There were no differences in the pattern of data obtained between the full sample and the last 20 subjects, and so we report the full $N = 24$ here. All participants provided informed consent and received monetary compensation of $15/h for taking part in the study.

2.2. Design

The present experiment consisted of four phases (Fig. 1): (1) preview phase, (2) directed-forgetting (DF) phase, (3) perceptual localizer (not pictured), and (4) recognition test. The purpose of the preview phase was to "inject" scene related information into the mind, so as to permit decoding of the context signal from the experiment, including the context reinstatement during the subsequent stages of the experiment. The DF phase then assigned memory cues for each previously studied word from the previous preview phase. The localizer phase consisted of a 1-back task performed on various perceptual categories for the purposes of training MVPA categorical classifiers that would be used to quantify item-related and context-related information. Lastly, the recognition test assessed participants' memory of the studied words. Functional magnetic resonance imaging (fMRI) data were collected during the preview, DF, and localizer phases only.

2.3. Preview phase

Preview trials consisted of the presentation of a word (3 s), followed by a triplet of scene images, presented back-to-back (1 s each). Inter-trial intervals (ITIs) consisted of a white fixation cross centered on a black screen (5 s). A total of 30 negative and 30 neutral words (with 180 interspersed scenes) were presented, with the constraint that no more than 3 consecutive trials of the words from the same emotional valence were presented. Each word appeared slightly offset from the center of the screen (4% screen height from the screen center), and participants performed an incidental task, by pressing buttons to indicate whether each word appeared above or below the center of the screen. Participants were told that "words will be separated by images of scenes", which they should "simply view passively". No more than three consecutive trials of the same word location appeared during presentation.

2.4. DF phase

All 60 words from the preview phase were presented again in the DF phase, with an *R* or *F* memory cue was assigned to them, with equal number of words within each valence category receiving each cue. Each DF trial consisted of the presentation of a word (2 s), a black screen (2 s), followed by a memory cue (8.5 s). Note that no scenes were presented during the DF phase. The ITI (0.5 s) consisted of a black screen. Word order differed between the preview and DF phases. Importantly, all words were presented in the same screen location (either above or below the center) as during the preview phase. This was done to avoid any overt changes in study context for the words. No more than three consecutive trials with the same valence or the same memory cue were allowed. Participants were instructed that words followed by an *R* cue will be tested later and that words followed by an *F* cue will not be tested and they should do their best to forget them. The *R* and *F* cues were implemented by two different symbols (see Fig. 1 b) – a blue open circle signified the *R* cue, and an orange circle with a line through the center signified the *F* cue. Symbols,

rather than words, were used as memory cues to avoid confounding the neural analyses, which sought to track the representation of target words both before and after the cue appeared.

2.5. Perceptual localizer

The localizer phase consisted of blocked presentation of exemplars from four stimulus categories – words, scenes, faces, objects, interspersed by rest blocks. Importantly, none of the words and scenes used for the localizer phase appeared during the earlier phases of the experiment, and none of the objects referred to a word from the earlier phases. To maintain attention to the presented stimuli, participants performed a 1-back task. Specifically, each stimulus block consisted of eight unique images of a category plus one repeated image that acted as the target for the 1-back task. Each trial consisted of a stimulus (0.5 s) and an interstimulus-interval of a black screen (1.5 s). Each rest block (10 s) consisted of a black screen with the instruction to "rest for 10 s" in white font. Participant pressed a button on every trial – one button for the presentation of each new exemplar or another button if they detected a repetition of a stimulus. There was a total of six blocked presentations of each stimulus category, resulting in 24 stimulus blocks and 24 rest blocks.

2.6. Recognition phase

Participants were given a recognition test at the end of the experiment, where they were shown studied words and novel lure words. The recognition test list contained 60 old words (30 negative, 30 neutral), along with 30 new words (15 negative, 15 neutral). Participants made judgments that combined study status (old/new) and confidence level (sure/maybe) using a 4-point scale (1 = sure old, 2 = maybe old, 3 = maybe new, 4 = sure new). The trials were self-paced with an average response time of M = 1.70 s (SD = 0.48). Critically, test instructions emphasized that participants should respond "old" to any word they recognize from the study, regardless of the R or F memory cue (i.e., F cue was "canceled"). These instructions are sufficient to dispel any concerns of demands characteristics, as prior work demonstrated that even offering to pay participants $0.50 for each additional F item remembered did not increase memory for these items (Macleod, 1999). This recognition test was given in the scanner, while anatomical scans were acquired (see MRI acquisition, MP-RAGE).

Counterbalancing of the experimental conditions was done by constructing six pseudorandomized versions of the preview and DF phases and administering each version to four participants, who received the same preview and DF instruction sequences. The assignment of words to *R, F,* and lures conditions (in the recognition test) were counterbalanced such that all words involved in the recognition study were assigned to each of these conditions, across the participants. Test order of the words in the recognition test was fully randomized for each participant.

2.7. Stimuli

For the localizer phase, a total of 192 items were shown, comprised of items selected from four stimulus categories (words, scenes, objects, faces), with 48 exemplars of each category. For words, half (24) were negative words (valence M = 2.62, SD = 0.54; arousal M = 5.36,

SD = 0.95) and half (24) were neutral words (valence M = 6.04, SD = 0.62; arousal M = 4.73, SD = 0.70), selected from Affective Norms for English Words (ANEW) database (Bradley and Lang, 1999). All images were presented in color. There were 48 indoor and outdoor scenes (800 × 600 pixels; the same dimensions as was shown in the preview phase), 48 everyday objects (300 × 300 pixels; from Google images), and 48 human faces (24 male, 24 female, 300 × 300 pixels; from Althoff and Cohen, 1999).

Word stimuli for the preview phase, DF phase, and recognition test phase were drawn from a set that consisted of 45 negative words and 45 neutral words selected from the ANEW database (for valence, negative M = 2.56, SD = 0.57, neutral M = 5.53, SD = 0.59; for arousal, negative M = 5.75, SD = 0.85, neutral M = 4.48, SD = 0.67), equated on concreteness, familiarity, word length, and Kucera and Francis frequency. Scenes used for the preview phase consisted of 180 colored images (800 × 600 pixels) of indoor and outdoor scenes taken from the Fine-Grained Image Memorability (FIGRIM) dataset (Bylinskii et al., 2015). None of the scene images contained words or human presence.

2.8. MRI procedure

fMRI scanning was conducted during the preview (~12 min), DF (~14.5 min), and localizer (~14.5 min) phases only. Each of the scanned phases was separated into two fMRI data collection runs, with a 20 s break separating each run. The preview and DF runs each contained 30 trials; each localizer run contained 12 blocks (3 blocks of each stimulus category). The recognition test phase consisted of one block of 90 self-paced trials, since no functional images and only anatomical MR images were collected.

2.9. fMRI data acquisition

All MR data was collected at the Beckman Institute's Biological Imaging Center at the University of Illinois at Urbana-Champaign on a 3T Siemens Magnetom Prisma scanner with a 64-channel coil. High-resolution T1-weighted structural brain images were acquired using a 3D MP-RAGE (magnetization prepared rapid gradient echo imaging) sequence acquired in a sagittal orientation (echo time = 2.32 ms, repetition time = 2300 ms, spatial resolution of 0.9375 × 0.9375 × 0.9 mm, field of view = 240 mm, flip angle = 8°). Functional brain images were acquired using gradient-echo, echo planar (EPI) sequence, with 38 axial slices collected in ascending order (with a 10% inter-slice gap) parallel to the anterior and posterior commissure (echo time = 25 ms, repetition time = 2000 ms, field of view = 230 mm, voxel size = 2.5 × 2.5 × 3.0 mm, flip angle = 90°).

2.10. fMRI preprocessing

Functional EPI images were preprocessed and analyzed using FSL 5.0 (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) subroutines implemented under MATLAB R2014a. Functional images were realigned to the middle volume of the middle run (fifth overall) to correct for motion, slice-time corrected, and high-pass filtered (128 s) to eliminate slow drift.

2.11. fMRI analysis: multivariate

All multi-voxel pattern analysis (MVPA) procedures were done in native space for each participant using the Princeton MVPA toolbox (https://github.com/PrincetonUniversity/princeton-mvpa-toolbox) and custom code in MATLAB R2014a. Here, we focused on category classifier activity in the ventral temporal cortex. The ventral temporal mask (in MNI space; Montreal Neurological Institute) was defined using boundaries delineated by Grill-Spector and Weiner (2014) and created by merging the temporal fusiform cortex, parahippocampal gyrus, occipital fusiform gyrus, temporal occipital fusiform cortex, and lateral occipital complex regions from the Harvard-Oxford atlas (Frazier et al., 2005; Desikan et al., 2006; Makris et al., 2006). To create subject-specific masks, we co-registered EPI volumes for each subject to their own MPRAGE structural volume using FSL FMRIB's Linear Image Registration Tool (*FLIRT*). We then used FSL FMRIB's Non-linear Image Registration Tool (*FNIRT*) to register structural volumes to MNI space. Individual, native-space ventral temporal masks were created by applying a reversed transformation matrix from EPI to MNI stereotaxic space on the atlas-space ventral temporal mask described above.

Training the classifier: We used MVPA to quantify the degree of face, scene, object, word, and rest category-specific neural activity associated with viewing of each stimulus category during the localizer phase. We trained five binary L2–penalized logistic regression classifiers (with a penalty of 50, based on prior work (Wang et al., 2019)) on faces, scenes, objects, words, and resting activity from the localizer phase. Each of these "one vs. other" category classifiers produced an "evidence score" which is the log-odds for the default category on which the logistic regression classifier was trained. Because the scores for each category were derived from different classifiers, they need not sum to 1, and thus provided more independent assessments of each category than if we had trained a single multinomial classifier. For each stimulus block, we trained and tested the classifier on the preprocessed BOLD data elicited from all 9 images (for a total duration of 9 TRs or 18 s) of each stimulus category. Regressors were shifted forward by 4 s to account for hemodynamic delay. Classifier training used the "leave-one-run-out" cross-validation method on the two localizer runs, in which the classifier is trained on one run, and is tested on the other run, rotating through both runs.

Testing the classifier on DF phase: To decode the DF phase for each participant, the classifiers were trained on both runs from the localizer data (separately for each participant) and then tested on each TR of *R* and *F* trials in the DF phase (for a total of 7 TRs for each trial). The evidence scores for each trial were uncorrected for hemodynamic delay in all figures. All analyses to assess the effects of *R* and *F* cues on memory representations (i.e. classification evidence), including statistical tests and the selection of time windows for analysis were determined using only the first four subjects. Analyses were unchanged for the remaining 20 subjects. Because of this analysis pipeline, we also replicated all analyses with only the final 20 subjects for whom analyses were pre-planned. This resulted in no qualitative changes (relative to the full set of 24 subjects), and therefore we report all 24 subjects in the Results. Any subsequent analyses that were not pre-planned will be labelled as exploratory.

2.12. Visualization of results

GLM and GLM-related surface results are visualized using *FSLeyes* (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) and the SPM12 canonical render. All subcortical findings are visualized over an MPRAGE volume that is comprised of averaging the MPRAGE volumes specific to this dataset.

2.13 . Multilevel modeling

All multilevel modeling analyses were done using R software (R Development Core Team, 2008), fitted with the *lmer* function in the *lme4* package (Bates et al., 2015) as well as the *lmerTest* package (Kuznetsova et al., 2017). 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 slopes were tested using the Mixture Chi-square likelihood-ratio test (Stram and Lee, 1994; 1995).

3. Results

3.1. Behavioral results

Word recognition performance was assessed by fitting receiver-operating characteristic curves specific to each participant and then computing the area under the ROC curve (AUC). This measure allows us to take into account both the hit rates and false-alarm rates across the participants' reported confidence levels (sure old, maybe old, maybe new, sure new), and is widely established as a measure of memory sensitivity (Egan, 1958; Macmillan and Creelman, 2004).

AUC for recognition performance was computed for each participant using the *perfcurve* function in MATLAB for each condition and tested with a repeated-measures ANOVA with the factors of memory cue (*R* vs. *F*) and word type (negative vs. neutral). The results are summarized in Fig. 1 c. Results show a robust main effect of memory cue, such that *F*-cued trials had lower memory sensitivity than *R*-cued trials ($F_{(1,23)} = 18.90$, $p < .001$). Word type (negative/neutral) did not significantly impact recognition ($F_{(1,23)} = 0.59$, $p = .45$), nor did it interact with memory cue ($F_{(1,23)} = 1.07$, $p = .31$).

In sum, behavioral results from the current study replicated robust, canonical DF effects, with no evidence of the effect of word type. For all subsequent analyses, results will be collapsed across word type and focused on the impact of *R* vs. *F* cues.

3.2. Neural results

To assess item and context processing for each trial of the DF phase, we first trained fMRI pattern classifiers, separately for each participant, on data in ventral temporal cortex that were collected during the perceptual localizer phase (Rissman and Wagner 2012; Lewis-Peacock and Norman, 2014; D'Esposito and Postle 2015). Using a cross-validation procedure to assess classification performance of the localizer data, we found that classification accuracy was above chance for all five categories (faces: 82.3 ± 2.5%; scenes: 85.0 ± 2.5%; objects: 68.9 ± 2.6%; words: 70.1 ± 4.4%; and rest: 74.6 ± 2.1%; chance = 20%, all Ps < 0.001, one-sample t-tests). Importantly, during the rest blocks, there was no

systematic identification of scene or face activity (classifier evidence scores of 0.38 and 0.34, respectively, compared to 0.81 for the rest category; Supplementary Fig. 2). After verifying that the localizer data provided sufficient sensitivity to discriminate each category of interest, we then re-trained the classifiers on all localizer data and applied them to new data from the preview and DF phases.

For the DF phase, we simultaneously decoded both words ("item" information) and scenes ("context" information) to assess how memory cues impacted the neural representation of the item and the context that was associated with that item in the preview phase. Note, that finding evidence of scene activation during the DF phase would indicate the *reinstatement of context-related activity from the preview phase* because scenes were shown in the preview phase but not in the DF phase (see Gershman et al., 2013). The average time courses of item and context decoding are shown separately for *F* trials and *R* trials in Fig. 2 a (30 trials per participant in each cue condition). Prior to the memory cue, there were no differences between *F* and *R* trials for either item or context. However, after the cue, there was a striking divergence such that on *F* trials, item information dropped and context information increased, relative to *R* trials.

To assess these changes statistically, we computed a difference score between average classifier evidence during the pre-cue period (TR 1 to 3; 0 to 6 s) and the post-cue period (TR 5 to 7; 8 to 14 s) of each trial for both item and context information (Fig. 2 b). Because these data are unshifted for hemodynamic lag, the pre-cue period was most influenced by the item presentation and had minimal influence from the cue (which occurred at 4 s), whereas the post-cue period captured the peak response to the cue (approximately 6 s after cue onset). Statistical tests conducted on these difference scores confirmed qualitative patterns observed in Fig. 2a. Item information significantly increased for *R* trials (change score tested against zero, $t(23) = 2.61$, $p = .015$), whereas item information significantly decreased for *F* trials ($t(23) = 5.87$, $p < .001$). These changes in item information were significantly different between *R* trials and *F* trials ($t(23) = 8.64$, $p < .001$).

Context information showed a different pattern of results. Context information was not impacted on *R* trials ($t(23) = 0.12$, $p = .91$), but it was significantly enhanced on *F* trials ($t(23) = 4.16$, $p < .001$). This enhancement of context information was significantly greater for *F* trials compared to *R* trials ($t(23) = 4.21$, $p < .001$).

3.3. Control analyses

In order to address whether the observed increase in context information was a methodological artifact of the decrease in item information – i.e., that the difference in scene and word evidence reflects a negative relationship between classes in a discriminatory classifier, rather than independent neural evidence for each class per se – we conducted two control analyses. First, we removed the word category from classifier training and then retested the DF trials. This analysis confirmed a selective increase of context (scene) information on *F* trials, but not on *R* trials (see Supplementary Fig. 3). The increase in context information, therefore, cannot be an artifact of the reduced item (word) information, as this category was not available to the classifier. Second, as a baseline for comparison against the item (word) and context (scene) decoding results, we assessed classifier evidence

for a trial-irrelevant category (face) during the DF phase. The classifier evidence for the face category remained unchanged across both *R* and *F* trials (Fig. 2 a). This confirms that the pattern of changes observed in word and scene evidence were specific to these two task-relevant categories of information rather than representing a general pattern for all categories on which the classifier was trained. Furthermore, the classifiers identified low levels of both face and scene information during resting periods in the localizer (Supplementary Fig. 2). Thus, the selective increase in scene activity on *F* trials cannot be attributed to the classifier simply identifying "resting" activity on these trials. Together, these control analyses suggest that the rise of context information and the fall of item information on *F* trials reflect independent and task-specific processes. To further evaluate this inference, we then related these neural measures from each trial to subsequent memory outcomes.

3.4. Relating neural evidence to memory outcomes

In an exploratory analysis, we conducted a series of multilevel modeling analyses to assess the relationship between neural evidence and the memory outcome on a trial-by-trial basis. Multilevel approaches are more powerful than the ANOVAs or unilevel regressions, and they are better suited for the nested data that we have (i.e., Jaeger, 2008). Given that recognition accuracy of studied words is the main outcome, it is important to note that high/low accuracy does not mean the same thing for *F* and *R* conditions and needs to be considered along with the goals of the task. Namely, high accuracy in *F* condition implies unsuccessful DF (i.e., the item survived in memory despite the *F* cue), whereas low accuracy means successful DF, consistent with the cue. The reverse is true in the *R* condition, where low accuracy implies memory failure despite the intention to remember (e.g., incidental forgetting).

Recognition accuracy of studied words was analyzed with a multilevel logistic regression, which linked brain activity with whether participants made a correct or incorrect recognition response on a trial-by-trial basis. A total of $n = 1440$ trials entered into the analyses (pulled across $n = 24$ participants), with half of the trials being *R* trials, and the remaining trials being *F* trials. In this analysis, classifier evidence scores for items (words) from the DF phase and Cue were used as a fixed effects, and participants were treated as a random intercept for those fixed effects. We initially tested whether random slopes would contribute significantly to our model by adding a random slope for the fixed effects of Cue and Item evidence. Doing so revealed that the model was not improved by including a random slope for the fixed effects, $Mixture\ \chi^2_{2,1} = 1.43$, $p = .36$, and therefore random slopes were not included in the final model. The variance associated with the random intercept of participants was $\sigma^2 = 0.58$, $SD = 0.76$. There was a significant interaction of Cue × Item Evidence (β€= 0.41, SE = 0.169, $Z = 2.41$, $p = .016$). Namely, on *R* trials ($n = 720$ trials across $n = 24$ participants), higher item evidence was associated with higher memory accuracy (β€= 0.22, SE = 0.002, $Z = 124.6$, $p < .001$). On *F* trials ($n = 720$ trials across $n = 24$ participants), on the other hand, item evidence was not significantly associated with memory accuracy, although higher item evidence was associated with numerically lower memory accuracy (β€= −0.18, SE = 0.100, $Z = 1.82$, $p = .068$).

The same multilevel logistic regression was run to assess the role of context on memory performance. In particular, the analysis was run on recognition accuracy of individual trials using classifier evidence scores for context (scenes) from the DF phase. Random slopes were not included in the final model because the initial analyses revealed that the model was not improved by including a random slope for the fixed effects of Cue and Context evidence, *Mixture* $\chi^2_{2,1} = 1.61$, $p = .33$. The variance associated with the random intercept of participants was $\sigma^2 = 0.60$, $SD = 0.77$.

The results revealed that on *R* trials, context evidence was not associated with memory accuracy ($\beta = 0.11$, SE = 0.099, $Z = 1.09$, $p = .28$). In contrast, on F trials, higher context evidence was associated with significantly worse memory accuracy ($\beta = -0.11$, SE = 0.002, $Z = 61.23$, $p < .001$). Thus, higher context evidence after an *F* cue was associated with a greater likelihood of successful forgetting. However, this effect was not significantly different between the two cue types (β€= −0.23, SE = 0.165, $Z = 1.36$, $p = .173$).

Given that item evidence was not associated with memory accuracy on *F* trials, whereas context evidence did (and the reverse was true for *R* trials), in a final analysis we computed a measure of *Neural Separation* defined by the difference between the context and item classifier evidence scores. A multilevel logistic regression was performed on recognition accuracy of individual trials, using memory Cue and Neural Separation as fixed effects, and participants as random intercepts for those effects (Fig. 2 c). Random slopes for the fixed effects were not included in the final model as they did not improve the model, *Mixture* $\chi^2_{2,1} = 1.71$, $p = .31$. The variance associated with the random intercept of participants was $\sigma^2 = 0.59$, $SD = 0.77$.

There was a significant Cue x Neural Separation interaction (β€= −0.34, SE = 0.166, $Z = 2.07$, $p = .038$). Specifically, in the F condition, higher neural separation between item and context was associated with significantly lower accuracy ($\beta = -0.19$, SE = 0.002, $Z = 112.8$, $p < .001$). In contrast, in the R condition, we did not find a significant relationship between neural separation and recognition accuracy (β€= 0.16, SE = 0.102, $Z = 1.56$, $p = .119$). Thus, higher neural separation was associated with successful DF, although it was not associated with successful recognition in the *R* condition.

4. Discussion

Prior research has made progress towards understanding the neural mechanisms that produce intentional forgetting in the item-method DF paradigm. However, none of the previous studies using this item-method paradigm examined the contribution of contextual information. The role of context processing has received substantial behavioral and neural support in list-method DF (for a review, see Sahakyan et al., 2013; Sahakyan, 2021), and this served as our motivation to examine if similar mechanisms might contribute to item-method DF. In order to assess the role of context in item-method DF we trained fMRI pattern classifiers to discriminate between item-information (using studied words) and contextual information (using trial-irrelevant scenes). We did not find any differences between neutral and emotional items, and we will not be discussing this variable further. However, we observed robust differences in the modulation of item and context information by DF

instructions. Specifically, the instruction to remember was associated with an increase in item processing, with no modulation of context. However, the instruction to forget an item was associated with a down-regulation of the item representation, along with an up-regulation of context information from the initial preview phase.

Furthermore, the magnitude of the neural separation between the context and item signal following the instruction to forget was associated with successful forgetting on a trial-by-trial basis. This dissociation does not reflect an artifact of classification, and thus instead can be interpreted as reflecting a true neural separation of context and item memory. This neural separation was not observed following the instruction to remember, however, where successful remembering was associated only with stronger item processing. Taken together, these results demonstrate a previously unappreciated role for context processing in the intentional forgetting of individual items.

In particular, we propose a new mechanism, which we term the *unbinding hypothesis*, to account for successful item-method DF. The hypothesis is similar in spirit to the one proposed to explain list-method DF (Sahakyan and Kelley, 2002). The fact that the magnitude of neural separation of items from their context observed during the DF phase was a significant predictor of subsequent DF success at the time of final recognition suggests a novel interpretation that the upregulation of contextual information and the concomitant downregulation of item information contribute to successful item-method DF. Together these processes may reflect an active unbinding of an item from its context (Hommel, 2004; Sadeh et al., 2012; Oberauer and Lewandowsky, 2016).

How might we think of this context signal? The signal is clearly scene specific; control analysis that showed that classifier evidence for the task-irrelevant face category remained unchanged across both *R* and *F* trials (Fig. 2a). The selective increase of scene information on *F* trials cannot be explained as a mere artifact of the classification procedure or a general increase of all non-word signals. Moreover, the signal is clearly related to successful forgetting; the observation that the relative degree of item and scene activity on *F* trials was associated with memory outcomes for these items demonstrates that this is a behaviorally relevant neural signal. One possibility, therefore, is that the contextual information from the preview phase was reinstated upon seeing the same words during the DF phase and that reinstatement was then used to facilitate forgetting; upregulation of context was associated with successful forgetting.

To bolster this interpretation, however, it would be necessary to identify the reactivation of specific scene stimuli that initially accompanied each item. Unfortunately, the experiment was not designed to allow for item-specific decoding of scenes. Moreover, participants were not instructed to form explicit associations between the words and specific intervening scenes during the preview phase, so it is unclear whether specific scene reactivation should even be expected. Further experimentation is thus necessary to understand the nature of this context signal, but regardless of its specificity, it is clear from our data that it is meaningfully related to directed forgetting.

These findings also have implications for the debate between the active and passive accounts of DF. Passive accounts posit that successful forgetting occurs from a failure to engage rehearsal processes to strengthen the unwanted memories. Our results show an increase in item-specific processing on *R* trials and a decrease in item-specific processing on *F* trials, which appears consistent with the passive account of DF. However, item processing is only one component of episodic memory, and the evidence of context processing on these trials suggests an alternative interpretation. There was no evidence of context reactivation on *R* trials, suggesting a rehearsal process focused on the most recent encoding episode. However, the reactivation of initial study context on *F* trials implicates focus on the initial memory trace. Active accounts posit that successful forgetting occurs from the engagement of processes that deliberately weaken the unwanted memories. To distinguish between them, it is important to compare the trials where *R*-cued items are subsequently forgotten ("incidental forgetting") and trials where *F*-cued items are subsequently forgotten ("intentional forgetting"). Typically, high accuracy is desired on tests of memory. However, considering the meaning of high and low accuracy in light of the goals of the task suggests that high accuracy for *F* items indicates that these items survived in memory despite the intention to forget them (i.e., unsuccessful DF). Likewise, low accuracy for *F* items indicates successful DF. By focusing on items that did *not* survive in memory, we can distinguish the active and passive accounts of DF. If subsequently forgotten *F* items show a similar neural profile as subsequently forgotten *R* items, this would be consistent with a passive account of DF. However, they do not show the same pattern. Neural separation was not only smaller overall for *R* trials than *F* trials, but greater forgetting for *R* items showed no relationship with neural separation between items and their context. In contrast, greater forgetting for *F* items (i.e. successful DF) was associated with greater neural separation between items and their context, suggesting the presence of an active unbinding process for *F* items that is not present for *R* items.

The current findings add to the growing body of literature indicating that item-method DF recruits active forgetting processes (Ludowig et al., 2010; Fawcett and Taylor, 2012, 2008; Lee et al., 2013; Wang et al., 2019).

Although our main goal of this investigation was to examine the role of contextual processing in item-method DF, we also examined the role of item-level processing. We found that in response to an *F* cue, there was a decrease in the neural signal associated with word processing. Such findings are inconsistent with the historically popular view that intentional forgetting occurs when participants not only terminate rehearsal of *F* items but also initiate selective rehearsal of other *R* items (see also Festini and Reuter-Lorenz, 2017). If in response to an *F* cue, participants use the post-cue period to rehearse previous *R* items, then the classifier evidence for the word information in our study should not have differed between the *F* and *R* trials. That is, participants would be selectively rehearsing previously presented words in either the *F* or *R* case, and this process would be signaled by an increase in word-related brain activity. The results are inconsistent with this account for *F* trials, as the word information decreased following the *F* cue. Thus, to account for the observed DF memory impairment, it is more parsimonious to assume some unbinding mechanism that separates the item from its (reinstated) context in order to forget the item.

Conceptually, the unbinding hypothesis proposed in this investigation is similar to the mechanism proposed to account for list-method DF (Sahakyan and Kelley, 2002). However, we do not claim that it is the sole mechanism responsible for item-method DF as plenty of previous research has confirmed the role of encoding differences between the *R* and *F* items. We merely suggest that item-method DF not only reflects a failure to encode information, but it is also driven by impaired retrieval at test arising from the unbinding of items from their context during DF. Similar ideas have been entertained also in recent behavioral and eye-tracking studies, demonstrating that item-method DF may impair contextual information (Whitlock et al., 2020a; 2020b). Using object-scene pairings, Whitlock et al. demonstrated that the association between the scenes and the object was impaired by DF instructions, and that it was *independent* of item impairment, such that participants could recognize the object (i.e., failure of DF despite the *F* cue), and yet forget which background scene the object was previously paired with. Finally, multinomial modeling analyses that disentangle encoding and retrieval components of memory effects indicate that worse memory of *F* items is driven not only by impaired encoding of *F* items, but also by impaired retrieval, which presumably could be voluntarily controlled (Rummel et al., 2016; Marevic and Rummel, 2018; Marevic et al., 2018).

Our results may seem to be inconsistent with our recent findings in an item-method DF paradigm reported in Wang et al. (2019). In that study, fMRI pattern classifiers identified an increase in item information (faces and scenes in ventral temporal cortex) in response to *F* cues relative to *R* cues, whereas we observed a decrease in item information (words). We entertain two possibilities to address the seeming disparity in these results. First, unlike in the current study, context information was not manipulated or measured in Wang et al. (2019). Thus, it is unclear whether the upregulation of *F* items observed in their study reflects an increased neural representation of the item information per se (specific to face and scene memory items), or whether it might also reflect an upregulation of context information, or some combination of both. Both face and scene stimuli are visually complex and rich in pre-experimental associations and could conceivably become incorporated into the global mental context of the memory episode. In other words, perhaps the upregulation observed by Wang et al. (2019) is more akin to the upregulation of context found in the present study. Alternatively, the different findings for the item processing between these two studies may be due to a critical methodological difference. In the current study, the DF cue was acting on the *second* presentation of the item (the first exposure to the word happened during the preview phase). The second exposure to the item likely reactivated its initial memory trace (which would include both item and context information; Howard and Kahana, 2002; Diana et al., 2007; Hannula et al., 2013), making it the target of the DF cue. In the Wang et al. (2019) study, however, the DF cue was acting on the first presentation of the item, with no prior context to be reactivated, thus perhaps making the item representation itself the target of the DF cue. This interpretation is speculative but an intriguing target for future research.

5. Conclusions

Using fMRI pattern classification of item and context information in an item-method DF paradigm, we established that the instruction to forget is associated with an upregulation of

an item's prior encoding context along with concomitant downregulation of that item's representation, implicating a separation/unbinding of item from its context in response to intentional forgetting. Furthermore, a larger magnitude of that neural separation was associated with successful intentional forgetting. These results contribute an important advance on our understanding of the cognitive processes and neural mechanisms involved in controlled forgetting of individual items, which until now have largely focused on the item and neglected the role of context.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

Funding for this work was provided by the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign and by R01 EY028746 (J.L.-P.). We gratefully acknowledge the support provided by the Biomedical Imaging Center of the Beckman Institute in the conduct of this work.

References

Althoff RR, Cohen NJ, 1999. Eye-movement-based memory effect: A reprocessing effect in face perception. J. Exp. Psychol. Learn. Mem. Cogn 25 (4), 997–1010. [PubMed: 10439505]

Anderson MC, Hanslmayr S, 2014. Neural mechanisms of motivated forgetting. Trends Cogn. Sci 18 (6), 279–282. doi: 10.1016/j.tics.2014.03.002, (Regul. Ed.). [PubMed: 24747000]

Basden BH, Basden DR, Gargano GJ, 1993. Directed forgetting in implicit and explicit memory tests: a comparison of methods. J. Exp. Psychol. Learn. Mem. Cognit 19 (3), 603–616. doi: 10.1037/0278-7393.19.3.603.

Bates B, Machler M, Bolker B, Walker S, 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw 67, 1–48.

Bylinskii Z, Isola P, Bainbridge C, Torralba A, Oliva A, 2015. Intrinsic and extrinsic effects on image memorability. Vis. Res 116 (Pt B), 165–178. doi: 10.1016/j.visres.2015.03.005. [PubMed: 25796976]

Bjork RA, 1970. Positive forgetting: the noninterference of items intentionally forgotten. J. Verbal Learn. Verbal Behav 9 (3), 255–268. doi: 10.1016/S0022-5371(70)80059-7.

Bjork RA, 1989. Retrieval inhibition as an adaptive mechanism in human memory. In: Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving. Lawrence Earlbaum Associates, Inc., p. 476.

Bjork RA, 2011. On the Symbiosis of Remembering, Forgetting, and learning. Successful Remembering and Successful Forgetting. Psychology Press.

Bjork RA, LaBerge D, Legrand R, 1968. The modification of short-term memory through instructions to forget. Psychon. Sci 10 (2), 55–56. doi: 10.3758/BF03331404.

Bower GH (1972). Stimulus-sampling theory of encoding variability. Coding Processes in Human Memory, 3, 85–123.

Bradley MM, Lang PPJ, 1999. Affective norms for english words (anew): instruction manual and affective ratings. Psychol. Tech 0 (C–1). doi: 10.1109/MIC.2008.114.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Killiany RJ, 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31 (3), 968–980. doi: 10.1016/j.neuroimage.2006.01.021. [PubMed: 16530430]

D'Esposito M, Postle BR, 2015. The cognitive neuroscience of working memory. Annu. Rev. Psychol. 66, 115–142 1 3. [PubMed: 25251486]

Diana RA, Yonelinas AP, Ranganath C, 2007. Imaging recollection and familiarity in the medial temporal lobe: a three-component model. Trends Cogn. Sci 11 (9), 379–386 (Regul. Ed.). [PubMed: 17707683]

Egan JP, 1958. In: Recognition Memory and the Operating Characteristic, 32. USAF Operational Applications Laboratory Technical Note, pp. 51–58.

Estes WK, 1955. Statistical theory of spontaneous recovery and regression. Psychol. Rev 62 (3), 145–154. doi: 10.1037/h0048509. [PubMed: 14371893]

Fawcett JM, Taylor TL, 2008. Forgetting is effortful: evidence from reaction time probes in an item-method directed forgetting task. Mem. Cognit 36 (6), 1168–1181. doi: 10.3758/MC.36.6.1168.

Fawcett JM, Taylor TL, 2010. Directed forgetting shares mechanisms with attentional withdrawal but not with stop-signal inhibition. Mem. Cognit 38 (6), 797–808. doi: 10.3758/MC.38.6.797.

Fawcett JM, Taylor TL, 2012. The control of working memory resources in intentional forgetting: evidence from incidental probe word recognition. Acta Psychol. 139 (1), 84–90. doi: 10.1016/j.actpsy.2011.10.001, (Amst).

Festini SB, Reuter-Lorenz PA, 2017. Rehearsal of to-be-remembered items is unnecessary to perform directed forgetting within working memory: support for an active control mechanism. J. Exp. Psychol. Learn. Mem. Cognit 43 (1), 94–108. doi: 10.1037/xlm0000308. [PubMed: 27668484]

Frazier JA, Chiu S, Breeze JL, Nikos Makris M, Lange N, David Kennedy SN, Biederman J, 2005. Article structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. Am. J. Psychiatry 162.

Gershman SJ, Schapiro AC, Hupbach A, Norman KA, 2013. Neural context reinstatement predicts memory misattribution. J. Neurosci 33 (20), 8590–8595. doi: 10.1523/JNEUROSCI.0096-13.2013. [PubMed: 23678104]

Grill-Spector K, Weiner KS, 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nat.Rev. Neurosci 15 (8), 536–548. doi: 10.1038/nrn3747. [PubMed: 24962370]

Hannula DE, Libby LA, Yonelinas AP, Ranganath C, 2013. Medial temporal lobe contributions to cued retrieval of items and contexts. Neuropsychologia 51 (12), 1–11. doi: 10.1016/j.neuropsychologia.2013.02.011. [PubMed: 23142349]

Hauswald A, Schulz H, Iordanov T, Kissler J, 2010. ERP dynamics underlying successful directed forgetting of neutral but not negative pictures. Soc. Cogn. Affect Neurosci. 6 (4), 450–459. doi: 10.1093/scan/nsq061. [PubMed: 20601423]

Hommel B, 2004. Event files: feature binding in and across perception and action. Trends Cogn. Sci 8 (11). doi: 10.1016/j.tics.2004.08.007.

Howard MW, Kahana MJ, 2002. A distributed representation of temporal context. J. Math. Psychol 46 (3), 269–299.

Jaeger TF, 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. J. Mem. Lang 59 (4), 434–446. doi: 10.1016/j.jml.2007.11.007. [PubMed: 19884961]

Kuznetsova A, Brockhoff PB, Christensen RHB, 2017. lmerTest Package: Tests in Linear Mixed Effects Models. J. Stat. Softw 82, 1–26.

Lee YS, Lee HM, Fawcett JM, 2013. Intentional forgetting reduces color-naming interference: evidence from item-method directed forgetting. J. Exp. Psychol. Learn. Mem. Cognit 39 (1), 220–236. doi: 10.1037/a0028905. [PubMed: 22732028]

Lehman M, Malmberg KJ, 2009. A global theory of remembering and forgetting from multiple lists. J. Exp. Psychol. Learn. Mem. Cogn 35 (4), 970–988. [PubMed: 19586264]

Lewis-Peacock JA, Norman KA, 2014. Competition between items in working memory leads to forgetting. Nat. Commun 5 (5768), 1–10. doi: 10.1038/ncomms6768.

Ludowig E, Möller J, Bien CG, Münte TF, Elger CE, Rosburg T, 2010. Active suppression in the mediotemporal lobe during directed forgetting. Neurobiol. Learn. Mem 93 (3), 352–361. doi: 10.1016/j.nlm.2009.12.001. [PubMed: 19969099]

MacLeod CM, 1998. Directed forgetting. In: Golding JM, MacLeod CM (Eds.), Intentional Forgetting: Interdisciplinary Approaches. Erlbaum, Mahwah, NJ, pp. 1–57.

Macleod C, 1999. The item and list methods of directed forgetting: test differences and the role of demand characteristics. Psychon. Bull. Rev 6, 123–129. [PubMed: 12199306]

Macmillan NA, Creelman CD, 2004. Detection theory: a user's guide. Detection Theory: A User's Guide, 2nd ed. Psychology Press, New York doi: 10.4324/9781410611147.

Makris N, Goldstein JM, Kennedy D, Hodge SM, Caviness VS, Faraone SV, Seidman LJ, 2006. Decreased volume of left and total anterior insular lobule in schizophrenia. Schizophr. Res 83 (2–3), 155–171. doi: 10.1016/j.schres.2005.11.020. [PubMed: 16448806]

Manning JR, Hulbert JC, Williams J, Piloto L, Sahakyan L, Norman KA, 2016. A neural signature of contextually mediated intentional forgetting. Psychon. Bul. Rev 23 (5), 1534–1542. doi: 10.3758/s13423-016-1024-7.

Marevic I, Rummel J, 2018. Retrieval-mediated directed forgetting in the item-method paradigm: the effect of semantic cues. Psychol. Res 3. doi: 10.1007/s00426-018-1085-5.

Marevic I, Arnold NR, Rummel J, 2018. Item-method directed forgetting and working memory capacity: a hierarchical multinomial modeling approach. Q. J. Exp. Psychol. 71, 1070–1180.

Mensink G-J, Raaijmakers JG, 1988. A model for interference and forgetting. Psychological Review 95 (4), 434–455.

Nowicka A, Marchewka A, Jednoróg K, Tacikowski P, Brechmann A, 2010. Forgetting of emotional information is hard: an fMRI study of directed forgetting. Cereb. Cortex 21 (3), 539–549. doi: 10.1093/cercor/bhq117. [PubMed: 20584747]

Oberauer K, 2018. Removal of irrelevant information from working memory: sometimes fast, sometimes slow, and sometimes not at all. Ann. N. Y. Acad. Sci 1–17. doi: 10.1111/nyas.13603. [PubMed: 30230554]

Oberauer K, Lewandowsky S, 2016. Control of information in working memory: encoding and removal of distractors in the complex-span paradigm. Cognition 156, 106–128. doi: 10.1016/j.cognition.2016.08.007. [PubMed: 27552059]

Oehrn CR, Fell J, Baumann C, Rosburg T, Ludowig E, Kessler H, Axmacher N, 2018. Direct electrophysiological evidence for prefrontal control of hippocampal processing during voluntary forgetting. Curr. Biol 28 (18), 3016–3022. doi: 10.1016/j.cub.2018.07.042, e4. [PubMed: 30197086]

Paz-Caballero MD, Menor J, Jiménez JM, 2004. Predictive validity of event-related potentials (ERPs) in relation to the directed forgetting effects. Clin. Neurophysiol 115, 369–377. doi: 10.1016/j.clinph.2003.09.011. [PubMed: 14744579]

Raaijmakers JGW, Shiffrin RM, 1981. Search of associative memory. Psychol. Rev 88 (2), 93–134.

Reber PJ, Siwiec RM, Gitelman DR, Parrish TB, Mesulam M−M, Paller KA, Gitleman DR, 2002. Neural correlates of successful encoding identified using functional magnetic resonance imaging. J. Neurosci. Off. J Soc. Neurosci 22 (21), 9541–9548.

R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3–900051-07–0, URL http://www.R-project.org.

Rissman J, Wagner AD, 2012. Distributed representations in memory: insights from functional brain imaging. Annu. Rev. Psychol 63, 101–128. doi: 10.1146/annurev-psych-120710-100344. [PubMed: 21943171]

Ritvo VJH, Turk-Browne NB, Norman KA, 2019. Nonmonotonic plasticity: how memory retrieval drives learning. Trends Cogn. Sci 23 (9), 726–742. doi: 10.1016/j.tics.2019.06.007, (Regul. Ed.). [PubMed: 31358438]

Rizio AA, Dennis NA, 2013. The neural correlates of cognitive control: successful remembering and intentional forgetting. J. Cogn. Neurosci 25 (2), 297–312. doi: 10.1162/jocn_a_00310. [PubMed: 23066730]

Rummel J, Marevic I, Kuhlmann BG, 2016. Investigating storage and retrieval processes of directed forgetting: a model-based approach. J. Exp. Psychol. Learn. Mem. Cognit 42 (10), 1526–1543. doi: 10.1037/xlm0000266. [PubMed: 26950491]

Sadeh T, Maril A, Bitan T, Goshen-Gottstein Y, 2012. Putting Humpty together and pulling him apart: accessing and unbinding the hippocampal item-context engram. Neuroimage 60 (1), 808–817. doi: 10.1016/j.neuroimage.2011.12.004. [PubMed: 22200724]

Sahakyan L, Foster NL, 2016. In: Dunlosky J, Tauber S (Eds.), The Need For Metaforgetting: Insights from Directed Forgetting. Oxford Handbook of Metacognition, pp. 341–356.

Sahakyan L, 2021. Current Perspectives on Directed Forgetting. In: Kahana MJ, Wagner A (Eds.), Journey to the Heart of Directed Forgetting, Eds. (in press). Oxford Handbook of Human Memory: Foundations and Applications.

Sahakyan L, Delaney PF, Foster NL, Abushanab B, 2013. List-method directed forgetting in cognitive and clinical research: a Theoretical and methdological review. In: The Psychology of Learning and Motivation, 59. Academic Press, pp. 131–190. doi: 10.1017/CBO9781107415324.004.

Sahakyan L, Kelley CM, 2002. A contextual change account of the directed forgetting effect. J. Exp. Psychol. Learn. Mem. Cogn 28 (6), 1064–1072. doi: 10.1037/0278-7393.28.6.1064. [PubMed: 12450332]

Sederberg PB, Howard MW, Kahana MJ, 2008. A context-based theory of recency and contiguity in free recall. Psychol. Rev 115, 893–912. [PubMed: 18954208]

Stram DO, Lee JW, 1994. Variance components testing in the longitudinal mixed effects model. Biometrics 50, 1171–1177. [PubMed: 7786999]

Stram DO, Lee JW, 1995. Corrections to “Variance components testing in the longitudinal mixed effects model” by Stram, D. O. and Lee, J. W.; 50: 1171–1177 (1994). Biometrics 51, 1196.

Taylor TL, 2005. Inhibition of return following instructions to remember and forget. Q. J. Exp. Psychol 58 (4), 613–629. doi: 10.1080/02724980443000115.

Taylor TL, Fawcett JM, 2011. Larger IOR effects following forget than following remember instructions depend on exogenous attentional withdrawal and target localization. Atten. Percept. Psychophys 1790–1814. doi: 10.3758/s13414-011-0146-2. [PubMed: 21618066]

Taylor TL, Fawcett JM, 2012. Does an instruction to forget enhance memory for other presented items? Conscious. Cogn 21 (3), 1186–1197. doi: 10.1016/j.concog.2012.05.002. [PubMed: 22687390]

Taylor TL, Quinlan CK, Vullings KCH, 2018. Decomposing item-method directed forgetting of emotional pictures: equivalent costs and no benefits. Mem. Cognit 46, 132–147.

Thompson KM, Taylor TL, 2015. Memory instruction interacts with both visual and motoric inhibition of return. Atten. Percept. Psychophys 77 (3), 804–818. doi: 10.3758/s13414-014-0820-2. [PubMed: 25592783]

Van Hooff JC, Ford RM, 2011. Remember to forget: ERP evidence for inhibition in an item-method directed forgetting paradigm. Brain Res. 1392, 80–92. doi: 10.1016/j.brainres.2011.04.004. [PubMed: 21514571]

Wang TH, Placek K, Lewis-Peacock JA, 2019. More is less: increased processing of unwanted memories facilitates forgetting. J. Neurosci 38 (17), 3551–3560. doi: 10.1523/JNEUROSCI.2033-18.2019.

Whitlock J, Lo Y, Chiu Y, Sahakyan L, 2020a. Eye movement analyses of strong and weak memories and goal-driven forgetting. Cognition 204. doi: 10.1016/j.cognition.2020.104391.

Whitlock J, Chiu Y, & Sahakyan L (submitted). 2020 Directed forgetting in associative memory: dissociating item and associative impairment.

Wylie GR, Foxe JJ, Taylor TL, 2008. Forgetting as an active process: an fMRI investigation of item-method-directed forgetting. Cereb. Cortex 18 (3), 670–682. doi: 10.1093/cercor/bhm101. [PubMed: 17617657]

Yang W, Liu P, Xiao X, Li X, Zeng C, Qui J, Zhang Q, 2012. Different neural substrates underlying directed forgetting for negative and neutral images: an event-related potential study. Brain Res. 1441, 53–63. [PubMed: 22285435]

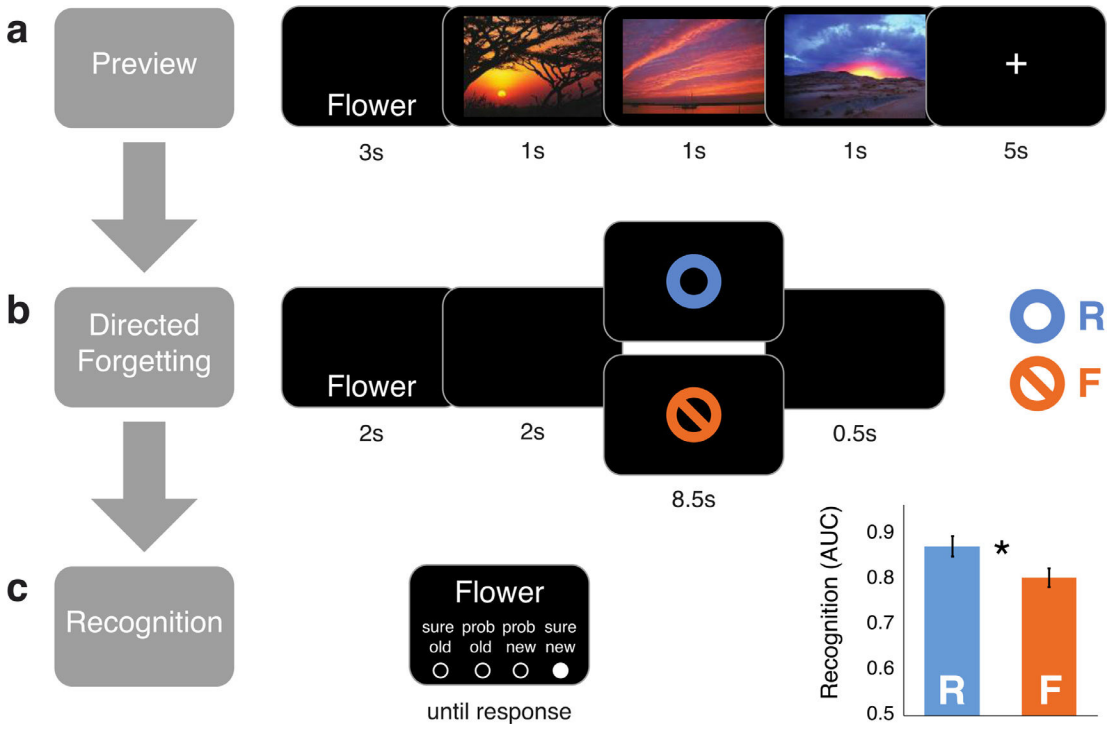

Fig. 1.

Study overview. (a) Participants underwent an initial preview phase in which they studied each word followed by a new set of three task-irrelevant scenes. Participants were asked to indicate whether the word was presented on the top or bottom half of the screen, and no response was required for the scenes. (b) Words were presented again, but without the scenes, followed by a memory cue (*R*: remember, *F*: forget). (c) At the end of the experiment a recognition confidence test was presented for all studied items plus novel foils. Memory recognition scores (AUC: area under the ROC) are shown for *R* and *F* items. Error bars indicate 95% CI of the difference score between remember and forget conditions. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

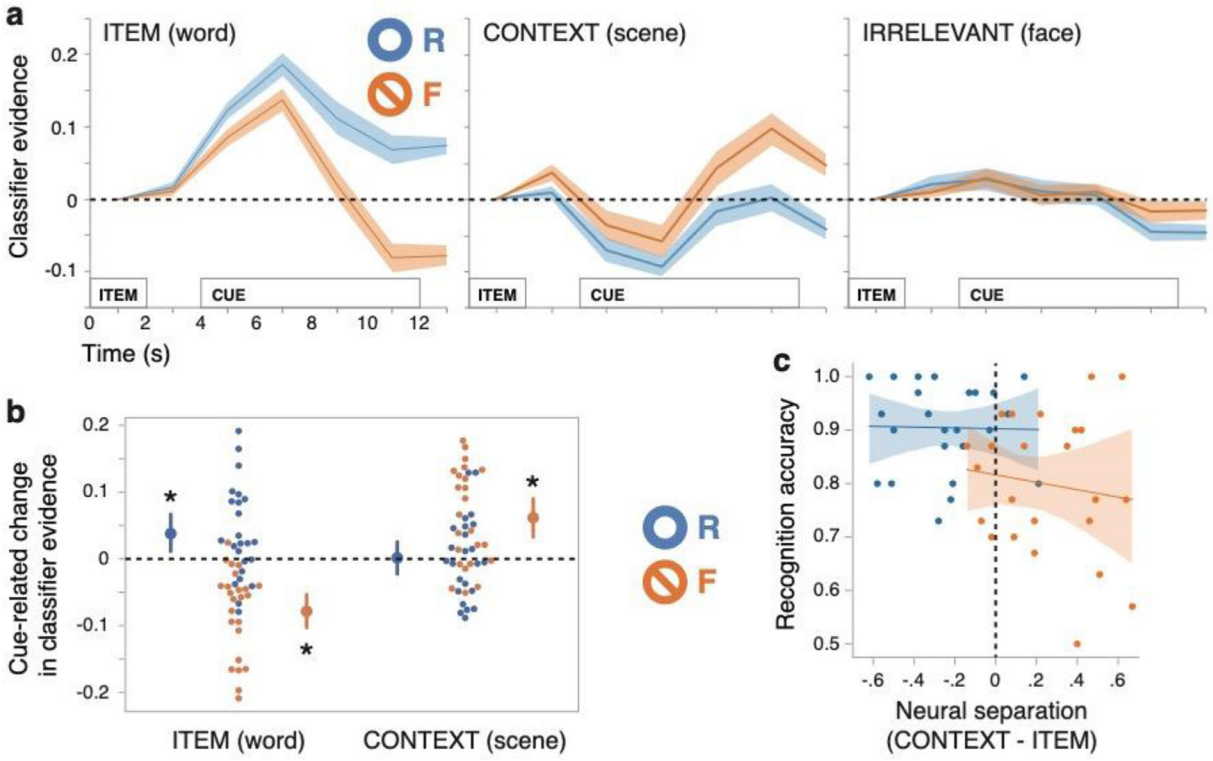

Fig. 2.

Decoding of fMRI data from DF task ($n = 24$). (a) Trial-averaged classifier evidence time courses are shown for item (word), context (scene), and irrelevant (face) categories. Values are not shifted to account for hemodynamic lag but are baselined relative to the first scan of each trial. Ribbon thickness indicates SEM of the difference between *R* and *F* conditions. *R*: Remember; *F*: Forget. Item and DF cue presentations are diagrammed along the horizontal axis. (b) Cue-related changes in classifier evidence were computed by subtracting the pre-cue scores (0 to 6 s) from the post-cue scores (8 to 14 s) on each trial. Small circles represent individual participants, and point estimates indicate group mean with 95% confidence intervals. * $P < .05$. (c) Relationship between behavioral accuracy and post-cue neural separation (context minus item). Data are visualized by averaging across participants, with error bars representing 95% confidence intervals, although multilevel modeling was done on individual trials.