Cognition 251 (2024) 105908

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Effortful retrieval of semantic memories induces forgetting of related negative and neutral episodic memories

Jamie Greer [a], Amna Ali [b], Camille Laksman [b], Ringo Huang [b], Mason McClay [b], David Clewett [b,*]

[a] Department of Psychology, Harvard University, United States of America
[b] Department of Psychology, University of California, Los Angeles, United States of America

ARTICLE INFO

Keywords:
Memory
Arousal
Retrieval-induced forgetting
Emotion
Inhibition
Effort

ABSTRACT

Retrieval-induced forgetting (RIF) experiments show that the act of retrieving some recently encoded items from a given conceptual category leads to greater forgetting of competing items from that same category. However, RIF studies using emotional stimuli have produced mixed results, perhaps due to the reinstatement of arousal or negative affect during retrieval practice. To induce forgetting of negative episodic memories more indirectly, we examined if retrieving neutral semantic memories leads to RIF of related negative memories. In two experiments, participants studied eight categorized lists comprised of an equal number of negative and neutral words (Experiment 1) or neutral words preceded by neutral or negative images (Experiment 2). To avoid re-exposing individuals to negative material during retrieval practice, participants then performed a semantic memory retrieval task in which they generated (i.e., completed word-stems for) new neutral words from half of the studied categories. We found that semantic retrieval, or word generation, induced forgetting of recently studied words irrespective of their emotional valence or original emotional context. Additionally, across both experiments, less successful word generation was associated with stronger RIF effects. In Experiment 2, the magnitude of RIF was also correlated with higher subjective ratings of retrieval effort during word generation. Together, these results suggest that even when retrieving neutral semantic memories, effortful retrieval may enhance inhibitory processes that lead to forgetting of both neutral and negative episodic memories.

1. Introduction

Remembering some past events can paradoxically lead to forgetting of related events, a well-established phenomenon known as retrieval-induced forgetting (RIF; Anderson et al., 1994). The first study on RIF demonstrated that the act of retrieving some recently encoded items from a given conceptual category (e.g., *fruit*) led to greater forgetting of competing items from that same category compared to items from categories that did not undergo retrieval practice (Anderson et al., 1994). In this now classic procedure, participants first studied several lists of category-exemplar pairs (Study Phase; e.g., *fruit-apple*, *fruit-orange*, *animal-horse*, *animal-cat*). In the second phase, they retrieved half of the exemplars from half of the studied categories using category and word stem cues (Retrieval Phase; e.g., *fruit-ap___*). Finally, they completed a recall test for all studied items (Test Phase). As expected, memory for unpracticed items from retrieval practiced categories (Rp- items; e.g., *orange*) was worse than memory for practiced items from practiced categories (Rp+ items; e.g., *apple*). Critically, memory for Rp- items was also worse than that of unpracticed items from non-retrieval practiced categories (Nrp items; e.g., *horse* and *cat*), indicating that the retrieval process not only strengthened memory for Rp+ items, but also weakened memory for Rp- items.

The RIF effect is robust and has been replicated across many similar studies using neutral stimuli such as visual objects (Shaw et al., 1995), categorized word lists (Rupprecht and Bäuml, 2016; Storm and Jobe, 2012), sentences (Chan, 2009), personality traits (Dunn and Spellman, 2003), arithmetic facts (Campbell and Thompson, 2012), and motor sequences (Tempel and Frings, 2013). There are several proposed mechanisms of RIF, including competition (or blocking) and inhibition (see Murayama et al., 2014 for a review). The competition-based account suggests that retrieving Rp+ items strengthens their representation in memory, causing them to interfere with retrieval of weaker Rp- items. According to the inhibition account, RIF is caused by the activation of Rp- items during the retrieval phase competing for conscious

* Corresponding author at: Department of Psychology, 5558 Pritzker Hall, University of California, Los Angeles, Los Angeles, CA 90095, United States of America.
E-mail addresses: jamiegreer@g.harvard.edu (J. Greer), dclewett316@g.ucla.edu (D. Clewett).

https://doi.org/10.1016/j.cognition.2024.105908
Received 15 October 2023; Received in revised form 19 July 2024; Accepted 23 July 2024
Available online 1 August 2024

recall with Rp+ items. To reduce memory interference from non-target exemplars, mental representations of the Rp- items are inhibited, making them less likely to be recalled or recognized during a subsequent memory test (Murayama et al., 2014). Although much work has focused on RIF effects using neutral memoranda, it is important to address how emotion may influence the forgetting of aversive memories.

Emotion has a significant influence on memory and, consequently, our ability to forget. For example, highly emotional experiences tend to become our most vivid and enduring memories (Kensinger et al., 2006; Yonelinas and Ritchey, 2015). Although it is generally adaptive to hold onto these memories to guide future behavior, persistent negative memories can disrupt day-to-day life and a subjective sense of well-being. An intriguing question arises as to whether RIF techniques can be harnessed to weaken unwanted emotional memories. There is a growing body of work investigating the application of RIF for emotional stimuli, which has thus far produced mixed results. These inconsistencies may be driven in part by differences in stimuli and methodology across experiments. For example, emotional RIF has been tested using word lists (Bäuml and Kuhbandner, 2007; Dehli and Brennen, 2009; Kobayashi and Tanno, 2013; Koessler et al., 2009; Moulds and Kandris, 2006), images with corresponding verbal labels (Barber and Mather, 2012; Kuhbandner et al., 2009), and autobiographical memories (Barnier et al., 2004; Harris et al., 2010; Hauer and Wessel, 2006; Matsumoto et al., 2021). There is also variability in the types of word stimuli that are used across list learning experiments, such as the use of social and non-social negative and positive words (Amir et al., 2001) or neutral words and three types of emotional words: trauma-specific, negative, and positive words (Blix and Brennen, 2012). While some studies have demonstrated RIF for emotional stimuli (Barnier et al., 2004; Harris et al., 2010; Hauer and Wessel, 2006), others indicate a resistance to RIF for such items (Bäuml and Kuhbandner, 2007; Blix and Brennen, 2012; Kobayashi and Tanno, 2013).

Another possible reason for inconsistent emotional RIF effects is that attempts to retrieve negative memoranda may reduce RIF by reinstating arousal or negative affect. Across several studies, RIF was not observed when participants were re-exposed to negative stimuli during retrieval practice (Blix and Brennen, 2012; Dehli and Brennen, 2009; Kobayashi and Tanno, 2013; Moulds and Kandris, 2006). Likewise, RIF was not observed when a negative mood was induced prior to retrieval through the presentation of emotional images (Bäuml and Kuhbandner, 2007) or a stressful social situation (Koessler et al., 2009). Negative moods may reduce retrieval competition amongst conceptually related exemplars by facilitating narrower item-focused encoding (Bäuml and Kuhbandner, 2007).

At the neural level, acute stress could also reduce RIF by triggering the release of cortisol (Koessler et al., 2009), a hormone that has been shown to impair memory retrieval processes in the lateral prefrontal cortex (LPFC; Arnsten, 2009). Neurobiological theories of RIF are centered upon the idea that inhibitory control processes implemented in the LPFC facilitate the targeted retrieval of specific memories and the suppression of similar non-target memories (Penolazzi et al., 2014; Wimber et al., 2009). Thus, inductions of stress or intense negative emotions during retrieval practice may impair selective retrieval and inhibitory processes that facilitate RIF. In further support of this idea, other work shows that *reducing* negative emotions indeed facilitates rather than disrupts RIF. For example, arousal levels can be reduced at retrieval by presenting participants with neutral verbal labels that had been previously paired with negative images, resulting in RIF for recently studied emotional stimuli (Barber and Mather, 2012; Kuhbandner et al., 2009).

Altogether, existing work suggests a need to target negative episodic memories more indirectly to avoid inducing emotional states that can disrupt emotional RIF. One such approach is semantic retrieval, a form of retrieval practice where individuals retrieve conceptually related words from semantic memory (i.e., prior knowledge) rather than recently formed episodic memories (Bäuml, 2002). Instead of practicing retrieving studied items, participants generate related words using category words paired with related word stem cues. For example, if participants studied fruit-related words during the Study Phase (e.g., *fruit-apple*, *fruit-orange*), during the Retrieval Phase they would be given cues to retrieve completely new (i.e., extralist) words from the same category (e.g., *fruit-ba___*). Retrieving the new word from semantic memory should still require inhibition of the recently studied words, so long as they are conceptually related, and thereby lead to RIF effects. Consistent with this idea, semantic generation of related negative words has been shown to induce RIF for recently studied negative words (Kobayashi and Tanno, 2015). This finding raises the intriguing possibility that semantic RIF could serve as an effective technique for indirectly reactivating and weakening negative episodic memories. However, one limitation of this study is that semantic retrieval was performed using negative words, which may induce negative affect. *Re*-exposure to negative information poses an issue for not only translating semantic-based retrieval strategies to treat memory-related disorders in the clinic (e.g., post-traumatic stress disorder, PTSD), but also for using forgetting techniques to weaken memories of everyday emotional incidents. From a therapeutic perspective, it would be beneficial to engage forgetting processes without having to explicitly retrieve any unpleasant information.

To address these issues of re-exposing individuals to negative stimuli, the current studies used a *neutral* semantic generation RIF paradigm. Additionally, we examined if the magnitude of emotional RIF is modulated by the difficulty of retrieving neutral words from semantic memory, as prior work suggests that the amount of effort to retrieve rather than retrieval success alone may engage inhibitory mechanisms that facilitate RIF (Storm et al., 2006).

2. Experiment 1

In Experiment 1, participants studied lists composed of both neutral and negative words. We hypothesized that semantic generation of unstudied neutral words would activate memories of previously studied words from the same category, leading to RIF for neutral and negative words alike. We also examined the relationship between the magnitude of RIF and word generation accuracy as an indirect measure of the difficulty of semantic memory retrieval.

2.1. Methods

2.1.1. Participants and design

Eighty-three healthy young adult participants were recruited using Prolific, a website that connects researchers with eligible participants based on a wide range of inclusion criteria. Eligibility criteria for the present study included: (1) ages 18 to 35 years old; (2) United States or Canada nationality; (3) English first language; (4) normal or corrected-to-normal vision; (5) no history of serious head injuries (e.g., concussion); (6) previously participated in a study on Prolific; and (6) have at minimum a 90% approval rate (percentage of studies for which the participant has been approved by the experimenter). Nationality and language requirements were included because the paradigm relies heavily on semantic memory and conceptual relatedness of English words, both of which are deeply tied to language proficiency. All participants provided written informed consent approved by the University of California, Los Angeles, Institutional Review Board and received monetary compensation. The experiment took approximately one hour to complete, and each participant received $9 for their participation.

Sample size was estimated based on comparable sample sizes in studies investigating RIF effects, including manipulations with emotional memoranda (Koessler et al., 2009; Kuhbandner et al., 2009; Storm et al., 2006; range: 48–64 participants. For transparency, we report the full number of participants who consented in the study, even though there were significant exclusions simply based on participant withdrawals from the experiment. A total of twenty-three participants were excluded from analyses for the following reasons: nineteen participants withdrew during the experiment; two participants were rejected because they exited the study immediately after starting; and two participants were excluded because they did not make encoding responses during the Study Phase. A total of 60 participants were included in the final analyses (49 female, 10 male, one gender nonbinary who was assigned female at birth; $M_{age} = 25$, $SD_{age} = 4.57$). Two participants identified as Asian, four as Black or African American, 50 as White, and four as more than one race. Regarding ethnicity, four participants identified as Hispanic, 55 participants identified as non-Hispanic, and one participant did not provide a response.

The current experiment used a 2 × 2 within-subjects design, with semantic retrieval (word category that underwent generation vs. word category that did not undergo generation) and emotional valence (negative word vs. neutral word) as fixed-effects predictors.

2.1.2. Materials

We compiled eight categorized lists of negative and neutral words (see Table 1). Each list consisted of a neutral category name, five neutral and five negative Study words (to be studied in Phase 1 and tested in Phase 3), and five neutral Generation words (to be generated in Phase 2). All categories and exemplars were one word in length and consisted of at least four letters. No two exemplars within a category began with the same first two letters to avoid matching partial word cues in the Generation and Test Phases.

Table 1
The eight categorized word lists used in Experiment 1.

Category Name	Neutral Study Words	Negative Study Words	Generation Words
FAMILY	Adoptive	Abandonment	Brother
	Cousin	Argument	Elder
	Spousal	Divorce	Genetic
	Stepson	Miscarriage	Pregnancy
	Twin	Violence	Relative
INSECT	Caterpillar	Cockroach	Beetle
	Grasshopper	Hornet	Bumblebee
	Moth	Lice	Cricket
	Snail	Tarantula	Dragonfly
	Worm	Wasp	Mantis
MEDICAL	Benign	Contagious	Clinic
	Doctor	Epilepsy	Dental
	Outpatient	Gangrene	Examine
	Pharmacy	Trauma	Nurse
	Prescribe	Tumor	Pediatric
OBJECT	Flashlight	Bomb	Broom
	Glass	Bullet	Fork
	Stapler	Grenade	Scissors
	Toothbrush	Handcuff	Spoon
	Wrench	Whip	Tweezers
OCEAN	Clam	Drown	Boat
	Dive	Marooned	Dock
	Reef	Pollute	Fish
	Snorkel	Shark	Seaweed
	Whale	Wreck	Wave
RELIGIOUS	Altar	Devil	Bible
	Priest	Sinner	Church
	Ritual	Hell	Faith
	Temple	Satanic	Pilgrim
	Worship	Crucify	Rabbi
SCHOOL	Credit	Bully	Class
	Elementary	Cheat	Degree
	Enrollment	Confiscate	Instruction
	Essay	Failure	Professor
	Seminar	Rejected	Student
WEATHER	Atmosphere	Flood	Balmy
	Cloudy	Freezing	Cold
	Forecast	Hurricane	Mist
	Overcast	Monsoon	Rainy
	Snowfall	Scorching	Windy

Each word list included five neutral Study words, five negative Study words, and five neutral Generation words.

Words were selected based on previously compiled semantic categories (Pereira et al., 2018) and supplemented with other words from the expanded ANEW database (Warriner et al., 2013). All words were selected based on normative ratings of valence (how positive or negative a word is) and arousal (the intensity of emotional response) from the expanded ANEW database. Across categories, the three word groups (neutral Study, negative Study, and Generation) were matched in valence and arousal ratings (all category pairwise comparisons p's > 0.05). To ensure words were highly conceptually related to words within their own categories and unrelated to words within other categories, we used Word2Vec, a model that learns word associations and extracts attributes related to word similarities, including semantic relatedness (Mikolov et al., 2013). The experiment was programmed using the PsychoPy builder (Peirce et al., 2019).

2.1.3. Experiment 1 procedure

2.1.3.1. Word study (phase 1). Participants viewed 80 category-exemplar pairs (e.g., *weather-flood*) for six seconds each during the Study Phase (Fig. 1A). The order of category-exemplar pairs was determined by blocked randomization: each eight-item block contained one random exemplar from each category, resulting in 10 blocks total. A fixation cross appeared in the center of the screen during each two-second interstimulus interval. Participants were instructed to think about how typical, or representative, each word was of its corresponding category when it appeared. While viewing each pair, they rated the typicality of the exemplar on a scale from 1 ('not typical at all') to 7 ('very typical'). The purpose of these ratings was to: (1) confirm that the participant was paying attention; (2) ensure that participants were associating exemplars with their category labels; and (3) collect word typicality ratings for subsequent analyses.

2.1.3.2. Distractor task #1. After the Study Phase, participants completed a 45-s distractor task in which they were presented with a rapid stream of 0.5-s left- or right-facing arrow symbols in the middle of the screen ('<' or '>'). Participants were instructed to indicate the direction of each arrow as quickly as possible by pressing the '1' or '2' key on their keyboard. The purpose of the distraction task was to introduce a short delay between the Study and Generation phases and to prevent any additional rehearsal of the final words in the list.

2.1.3.3. Semantic word generation (phase 2). Half (four) of the categories were 'practiced' in the Generation Phase using category and word stem cues for extralist words (i.e., novel words that were not studied in Phase 1; Fig. 1B). These trials were also presented in a blocked randomized order. Each cue was presented for seven seconds and consisted of the first two letters and last letter of the word (e.g., *weather-ra__y*). Although the number of missing letters varied depending on the word, the blank space was always the same length. The four categories used in the Generation Phase were randomized across participants. Each extralist word stem cue was presented three times, and participants were told to type the same word each time the cue was presented if they were able to generate it. Only the third response was used in analyses to compute generation accuracy. Participants were instructed to think of and type the entire correct word (e.g., *rainy*) in the text box at the bottom of their screen. If the seven seconds expired before the participant finished typing, the screen advanced and their current response was recorded.

2.1.3.4. Distractor task #2. After the Generation Phase, the participants completed the same 45-s distractor task again. This was used to introduce a short delay between the Generation and Test phases and to reduce any potential recency effects in the final memory test.

2.1.3.5. Memory test (phase 3). In this phase of the experiment, we

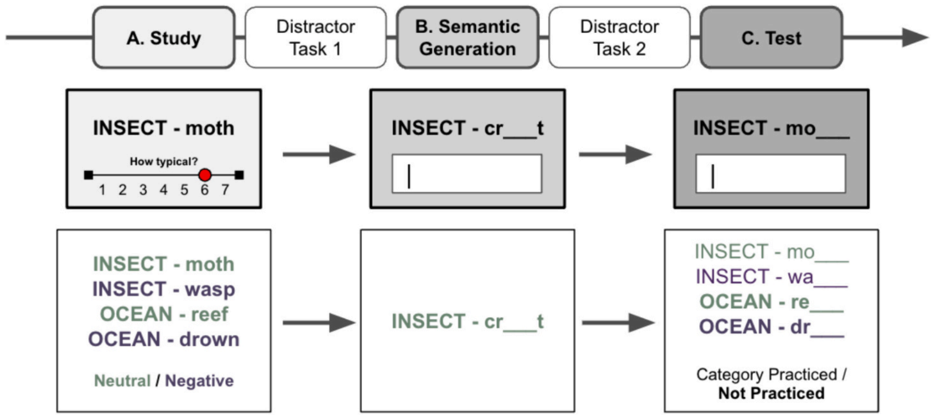

Fig. 1. Experiment 1 semantic retrieval-induced forgetting task. (A) Study Phase. On each trial of the first phase, participants were presented with a word category label in capitalized letters next to an exemplar from that same category in lowercase letters. At this time, participants rated the typicality of each exemplar (how representative the exemplar was of its category) on a scale from 1 to 7. (B) Generation Phase (semantic retrieval). During the second phase, participants were shown half of the word categories that were encountered in the Study Phase along with word stems for new, related exemplars (i.e., extralist words). All extralist words were neutral. On each trial, participants were presented with a category word displayed in capital letters next to a partial word cue. Each trial repeated three times throughout the phase in separate blocks (i.e., all 40 words appeared in one block, and then all 40 words appeared again in the next block, etc.). (C) Test Phase. Memory for all Study Phase words was tested by presenting category and word stem cues in a random order. Categories that were used in the Generation Phase are 'practiced,' while categories that were not used in the Generation Phase are 'not practiced.' In this example, worse performance for the two INSECT items (category practiced) compared to the two OCEAN items (category not practiced) would indicate RIF.

tested participants' memory for all 80 exemplar words from the Study phase (Fig. 1C). Category labels and partial word cues were presented for ten seconds each in a completely randomized order. Since the memory test words were previously studied and would therefore be easier to retrieve than the semantic generation words, only the first two letters (e.g., *weather-fl___*), rather than the first two letters and last letter, were presented as a cue. The participants were instructed to only respond with words that were studied in the first phase of the experiment.

2.1.3.6. Subjective emotion and typicality ratings for all words. Following the entire RIF task, participants provided subjective valence, arousal, and typicality ratings for all words from the experiment. With the exception of Study word typicality ratings, which were collected during Phase 1, all ratings were collected in a Qualtrics survey. The ratings were made using an ordinal scale from one to seven for the following characteristics: valence (1 = 'very negative' to 7 = 'very positive'), arousal (1 = 'not arousing' to 7 = 'very arousing'), and typicality (1 = 'not typical at all' to 7 = 'very typical').

2.1.4. Response coding procedures

2.1.4.1. Cued recall accuracy coding procedure. Word responses from the Test Phase and the last of the three responses for each cue from the Generation Phase were marked as correct or incorrect. Two independent raters (JG and CL) graded each response. Discrepancies were resolved by a third rater (DC).

2.1.4.2. Responses from the semantic retrieval task (generation phase). A response was marked as correct if it was the correct word with correct or incorrect spelling (e.g., *religion-ra___i* being written as *rabi* vs. *rabbi*), the correct word in incorrect form (e.g., plural, such as *school-cl___s* being written as *classes* vs. *class*), or unclear due to spelling error or incompletion based on one of the two previous generations (e.g., *weather-mi___t* being written as *mis* in the third response and *mist* in the second response). A response was also marked as correct if it was not the intended word but still fit the word stem and category (e.g., *school-in___n* being written as *intern* vs. *instruction*). Because the Generation words were not previously studied and instead retrieved from semantic memory, a different within-category word that fit the word stem would technically be correct. A response was marked as incorrect if it was left blank, a word from an incorrect category (e.g., *weather-co___d* being written as *coward* vs. *cold*), an incorrect word that did not fit the word stem when spelled correctly (e.g., *ocean-wa___e* being written as *watre* vs. *wave*), or more than one word (e.g., *school-in___n* being written as *in session* vs. *instruction*). For the main analyses, memory test accuracy was coded as a binary factor variable (1 = correct, 0 = incorrect). To measure the reliability of the response ratings, we calculated the intraclass correlation coefficient (ICC) before discrepancies were resolved. The ICC was 0.92, indicating strong reliability.

2.1.4.3. Responses from the final memory test (test phase). Responses from the final cued recall memory test were graded more stringently than the responses from the Generation Phase because these words were previously studied. A Test response was marked as correct if it was the correct word with correct or incorrect spelling (e.g., *medical-be___* being written as *benine* vs. *benign*) or if it was missing the last letter but showed clear intent of writing the correct response (e.g., *insect-gr___* being written as *grasshoppe* vs. *grasshopper*). A response was also marked as correct if it was missing multiple end letters but with clear intent of writing the correct response (e.g., *ocean-se___* being written as *seawe* vs. *seaweed*), if it was the correct base word in the wrong form (e.g., *ocean-dr___* being written as *drowned* vs. *drown*), or if another suffix was added or omitted (e.g., *weather-sn___* being written as *snow* vs. *snowfall*) with the goal of capturing the presence of memory representations, even if only partially accurate. A response was marked as incorrect if it was left blank, unfinished, or spelled in a way that made the participant's intended response unclear (e.g., *religious-al___* being written as *alt* vs. *altar*). A response was also marked as incorrect if it was an incorrect word regardless of category, including other within-category words (e.g., *weather-hu___* being written as *hurricane* vs. *humidity*) and Study Phase words from a different category (e.g., *ocean-sn___* being written as *snail* vs. *snorkel*). For analyses, memory test accuracy was coded as a binary variable (1 = correct, 0 = incorrect). The ICC was 0.98; thus, inter-rater response coding was very reliable.

2.1.5. Analyses

2.1.5.1. Retrieval-induced forgetting analyses. To test our main hypothesis that the generation of neutral words would induce forgetting of related, previously studied negative and neutral words, we performed generalized linear mixed-effects modeling analyses using the lme4 package in R (Bates et al., 2015). Random effects were specified for participant identity and word identity. Generation (generation, no generation) and valence (negative word, neutral word) were modeled as fixed-effect predictors. A generation-by-word valence interaction term, word frequency, and participant ratings of word typicality were systematically included as fixed effects. Word frequency and typicality were included based on evidence that taxonomic frequency influences RIF (Anderson et al., 1994). Model testing was performed to identify the best-fitting model according to the Bayesian information criterion (BIC). The best model included fixed-effect predictors for generation, valence, and typicality. It had a conditional R^2 of 0.40, which was calculated using the MuMIn package in R (Bartoń, 2023). The model was then tested for negative and neutral words separately. Data visualizations were created using the ggplot2 (Wickham et al., 2023) and sjPlot (Lüdecke, 2023) packages.

The best-fitting model did not include a generation-by-word valence interaction term, which was not significant when included. To explore whether RIF was equivalent for neutral and negative items, we fitted a Bayesian logistic mixed model (estimated using MCMC sampling with 4 chains of 2000 iterations and a warmup of 1000) using the rstanarm package in R (Goodrich et al., 2024). The model included the Generation-by-Valence interaction and Typicality, as well as random intercepts for participant identity and word identity. Priors over parameters were all set as normal (mean = 0.00, SD = 5.00; mean = 0.00, SD = 5.00; mean = 0.00, SD = 1.49; mean = 0.00, SD = 5.77) distributions. The model's explanatory power is substantial ($R^2 = 0.28$, 95% CI [0.26, 0.31]).

2.1.5.2. Effects of semantic retrieval success and generated word typicality on memory test performance. In a second set of analyses, we examined if the degree of RIF for negative and neutral words was correlated with successful semantic retrieval (i.e., generation) and the typicality of new words in Phase 2. First, we computed the average word generation success for each category that appeared during Phase 2. Second, we computed the average typicality ratings across each of the word categories that appeared in Phase 2. We assessed generation success and typicality at the category-level rather than at the exemplar-level because the generated words were different from the Study phase exemplars; that is, individual Study Phase words did not correspond to the generation success and typicality of individual exemplars, but rather to all generated word exemplars within these generated word categories. We performed generalized linear mixed-effects modeling analyses, this time using valence, generation success, and generated word typicality as fixed-effect predictors of memory test accuracy. As before, participant identity and word identity were specified as random effects. The best-fitting model according to the BIC also included test word typicality as a fixed effect, but no emotion-by-generation accuracy interaction term. This model had a conditional R^2 of 0.41.

2.2. Results

2.2.1. Subjective ratings for word valence, arousal, and typicality

As a manipulation check, paired *t*-tests were used to compare participants' subjective valence and arousal ratings for the preselected negative and neutral words. Consistent with the normative ratings from the expanded ANEW database (Warriner et al., 2013), participants rated the negative and neutral words as significantly differing in both valence and arousal. Specifically, valence ratings for negative words were significantly lower ($M = 2.16$, $SD = 1.27$) than ratings for neutral words ($M = 4.46$, $SD = 1.22$; $t(59) = 22.68$, $p < .001$, 95% CI [2.09, 2.50]), and arousal ratings for negative words were significantly higher ($M = 2.97$, $SD = 2.00$) than ratings for neutral words ($M = 2.28$, $SD = 1.73$; $t(59) = -3.64$, $p < .001$, 95% CI [−1.06, −0.31]).

A two-tailed paired t-test was also used to compare the participants' typicality ratings for the negative and neutral words from the Study Phase. Typicality ratings for negative words were significantly lower ($M = 4.74$, $SD = 1.79$) than ratings for neutral words ($M = 5.70$, $SD = 1.39$; $t(59) = 15.62$, $p < .001$, 95% CI [0.83, 1.08]), suggesting that negatively valenced words were perceived as being more obscure and less representative of their respective categories than neutral words. See Table 2 for a summary of participant ratings.

2.2.2. Effects of word valence and semantic retrieval on memory test performance

We used generalized linear mixed-effects modeling to first examine if semantic retrieval (i.e., generation) of neutral words induced forgetting of previously studied, related exemplars that were negative or neutral (see Table 3A, Fig. 2). The results revealed a significant main effect of Valence ($\beta = 0.81$, $p = .0084$) on memory test performance, such that recall was better for neutral words than for negative words from the Study Phase. We also found a significant main effect of word generation ($\beta = 0.32$, $p < .001$) on memory test performance, such that memory was worse for words from categories that were practiced, or underwent semantic retrieval, during the Generation Phase compared to words from categories that were not practiced. The effect of Typicality on recall was not significant ($\beta = 0.040$, $p = .14$).

These findings reveal that when controlling for typicality, semantic retrieval induced forgetting of previously studied neutral and negative words, and neutral words were more memorable overall. We then repeated the analysis for neutral (see Table 3B) and negative (see Table 3C) words separately, which showed the same effect of word generation ($\beta = 0.26$, $p = .0098$ for neutral; $\beta = 0.37$, $p < .001$ for negative) on memory test performance. The neutral model also showed a significant effect of typicality ($\beta = 0.11$, $p = .013$) on memory test performance, such that neutral words that were highly typical were more likely to be remembered.

For the Bayesian model, the effect of the Generation-by-Valence interaction (Median = −0.11, 95% CI [−0.37, 0.16]) had a 78.57% probability of being negative (< 0), 54.52% of being significant (< −0.09), and 0.05% of being large (< −0.54). The estimation successfully converged (Rhat = 1.001) and the indices were reliable (ESS = 4358). The interaction had an odds ratio of 0.90 and a credible interval of [0.69, 1.18]. The credible interval included 1, indicating there may be equivalent RIF for negative and neutral words.

2.2.3. Effects of semantic retrieval success and generated word typicality on memory test performance

To test if successful semantic retrieval (i.e., generation) and generated word typicality predicted the magnitude of RIF, we performed separate generalized linear mixed modeling analyses by modeling word valence, word generation success, generated word typicality, and test word typicality as fixed effects (see Table 4, Fig. 3). Participant identity and word identity were modeled as random effects. The results revealed a significant main effect of generation success ($\beta = 0.54$, $p = .034$) on memory test performance, such that memory performance for studied words from practiced categories was higher when word generation success was higher. That is, when a given participant was relatively less successful during the generation phase for a specific category, that participant showed worse recall of the previously studied words from that same category. Again, there was a significant main effect of word

Table 2
Summary of participants' valence, arousal, and typicality ratings (1–7) of neutral and negative words from the Study Phase of Experiment 1.

	Valence	Arousal	Typicality
	Mean (SD)	Mean (SD)	Mean (SD)
Neutral words	4.46 (1.22)	2.28 (1.73)	5.70 (1.39)
Negative words	2.16 (1.27)	2.97 (2.00)	4.74 (1.79)

Table 3
Generalized linear mixed-effects modeling results for A) the effect of valence, generation, and typicality on memory test performance for studied negative and neutral words, B) the effect of generation and typicality on memory test performance for studied neutral words, and C) the effect of generation and typicality on memory test performance for studied negative words.

Fixed Effects	A) Negative and Neutral RIF Model				B) Neutral RIF Model				C) Negative RIF Model			
	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
(Intercept)	−0.56	0.26	−2.13	0.033*	−0.12	0.34	−0.36	0.72	−0.37	0.29	−1.29	0.2
Valence	0.81	0.31	2.64	0.0084**								
Generation	0.32	0.069	4.56	< 0.001***	0.26	0.1	2.58	0.0098**	0.37	0.097	3.82	< 0.001***
Typicality	0.04	0.028	1.46	0.14	0.11	0.045	2.49	0.013*	−0.0048	0.035	−0.14	0.89
Random Effects	Variance	*SD*			Variance	*SD*			Variance	*SD*		
Participant Identity	0.18	0.43			0.22	0.47			0.18	0.43		
Word Identity	1.76	1.33			1.67	1.29			1.96	1.4		

*$p < .05$, **$p < .01$, ***$p < .001$.

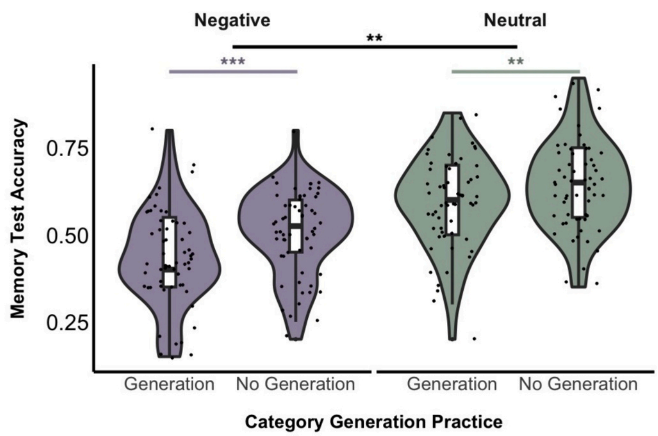

Fig. 2. Experiment 1 results. Memory test accuracy results by valence (negative vs. neutral) and generation (generation vs. no generation). Test accuracy was significantly higher for neutral words (green) than negative words (purple). Test accuracy was also significantly higher for words in 'no generation' categories than 'generation' categories. Each point represents an individual participant's average memory test accuracy. Violin plots show the density distribution of the test accuracy scores for each condition. Box plots summarize key statistical measures: the box represents the interquartile range, the horizontal line represents the median, and the whiskers extend to minimum and maximum values (excluding outliers). **$p < .01$, ***$p < .001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

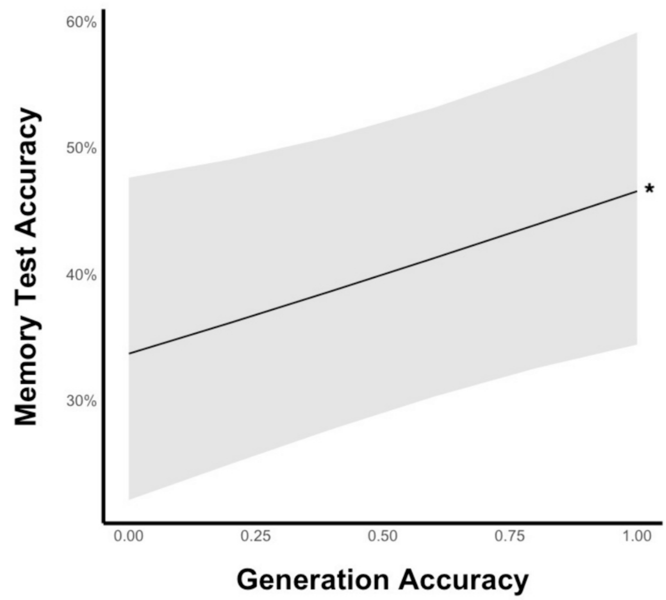

Fig. 3. Experiment 1 results showing relationship between the accuracy of semantic memory retrieval (word generation) and retroactive forgetting. This effect plot visualizes predicted values for memory test accuracy at each value from the generation accuracy term, with all other model predictors held constant. The shaded area represents the 95% confidence band. Lower word generation accuracy predicted lower memory test accuracy scores for recently studied words from those same categories. *$p < .05$.

Table 4
Generalized linear mixed-effects modeling results for effects of semantic retrieval success and generated word typicality on memory test performance for studied negative words.

Fixed Effects	Generation Success Model			
	β	*SE*	*z*	*p*
(Intercept)	−0.8	0.7	−1.14	0.25
Generation Success	0.54	0.25	2.13	0.034*
Valence	0.87	0.33	2.62	0.0089**
Generated Word Typicality	−0.017	0.11	−0.16	0.87
Test Word Typicality	0.038	0.041	0.93	0.35
Random Effects	Variance	*SD*		
Participant Identity	0.16	0.39		
Word Identity	1.94	1.39		

*$p < .05$, **$p < .01$.

valence ($\beta = 0.87$, $p = .0089$) on memory test performance, such that recall was better for neutral words than for negative words. The effect of generated word typicality ($\beta = -0.017$, $p = .87$) on recall was not significant.

2.3. Experiment 1 discussion

In Experiment 1, we found that a semantic RIF paradigm using only neutral exemplars during semantic memory retrieval led to greater forgetting of recently studied and related neutral and negative words. Surprisingly, memory for neutral words was better than memory for negative words. Normally, emotional stimuli stand out more and are better remembered than neutral words when they are presented in mixed lists (see Talmi et al., 2019 for a review). However, high-frequency words tend to be better remembered than low-frequency words in free or cued recall (MacLeod and Kampe, 1996). This difference in word retention in the present study may be due to the neutral and negative words used in Experiment 1 differing in arousal and valence. It is likely not due to differences in typicality alone since the effect of valence remained significant even when typicality was controlled for in the model. One explanation for the lack of an expected emotional memory enhancement is that there is a larger mismatch between the valence of the negative memoranda (e.g., "argument") and their respective category labels (e.g., "family") compared to neutral words (e.g., "cousin"), thereby making the retrieval cues for negative items less effective on the final memory test.

To reduce potential effects of valence-related differences of typicality on memory performance, we performed a second experiment in which we manipulated emotional contexts during the initial Study Phase. Namely, all memoranda of interest were neutral words and we induced emotion by presenting a negative or neutral image before each category-exemplar pair. This procedure may also be a better representation of how negative memories are formed in the real world, given that neutral details are encoded within broader emotional contexts.

In Experiment 1, we also found that lower semantic memory retrieval performance was associated with greater forgetting of previously studied negative and neutral words from those categories. Even when only retrieving neutral information, a difficult semantic retrieval task may thereby promote the inhibition of related negative and neutral memory representations, leading to greater RIF effects. This possibility points to an inhibitory mechanism and mimics the effects observed during an 'impossible retrieval' version of the RIF task, in which the mere act of trying to generate a word induced forgetting of related studied words (Storm et al., 2006; Storm and Nestojko, 2010). However, it is still unclear from Experiment 1 whether worse generation performance is an accurate representation of generation effort.

3. Experiment 2

In Experiment 2, the Study lists were composed of only neutral words. To imbue some categories of words with emotion, Study words were preceded by either a neutral or negative image. To gain insight as to why failures to generate words were associated with greater RIF in Experiment 1, we also obtained subjective ratings of semantic retrieval effort after each Generation trial.

3.1. Methods

3.1.1. Participants and design

Eighty-four healthy young adult participants were recruited via SONA at the University of California, Los Angeles. Although the initial target sample size was lower, a bug in the code was caught partway through data collection that caused an unequal distribution of participants across conditions (i.e., which categories were practiced during Phase 2). To address this issue, we increased the number of participants until the conditions were effectively counterbalanced. We aimed for a convenience sample of 80 participants to achieve a well-balanced design. Eligibility criteria for Experiment 2 was identical to Experiment 1, except for Prolific-specific eligibility, with the addition of not being enrolled in a course taught by the corresponding author. All participants provided written informed consent approved by the University of California, Los Angeles, Institutional Review Board and received monetary compensation. Six participants were excluded from analyses for the following reasons: four participants were excluded because English was not their first language; one participant was excluded because they were not American or Canadian; and one participant was excluded due to data loss. A total of 78 participants were included in the final analyses (60 female, 18 male; $M_{age} = 20.05$, $SD_{age} = 1.97$). Two participants identified as American Indian or Alaskan, 28 as Asian, four as Black or African American, 28 as White, eight as more than one race, and eight participants did not provide a response. Regarding ethnicity, 20 participants identified as Hispanic, 57 participants identified as non-Hispanic, and one participant did not provide a response.

As before, this experiment used a 2 × 2 within-subjects design, with semantic generation (category that underwent generation vs. category that did not undergo generation) and emotional valence (word preceded by negative image vs. word preceded by neutral image) as fixed-effect predictors of memory test performance. To acquire an endogenous measure of arousal and retrieval effort, we also collected pupillometry data during the Study and Generation Phases of the task. However, the pupillometry results are not reported in this paper.

3.1.2. Materials

We compiled eight categorized lists of neutral words (see Table 5). The words were selected from the updated version of the Battig and Montague norms (Van Overschelde et al., 2004). Each list consisted of a neutral category name, ten neutral Study words (to be studied in Phase 1 and tested in Phase 3), and five neutral Generation words (to be generated in Phase 2). The categories from Van Overschelde et al. (2004) contained >15 items, so some of the items were excluded. As in Experiment 1, only one-word category and exemplar names with four or more letters were included, and no two exemplars within a category began with the same first two letters. An exemplar was excluded if it related strongly to another category. To prevent cross-category interference, no more than two exemplars across categories began with the same first two letters, and words that did match were as irrelevant to the other category as possible. To prevent interference from other same-category words, when possible, no chosen category exemplar had the same first two letters as an unchosen category exemplar listed in the category norms (Van Overschelde et al., 2004). Additionally, when possible, exemplars were selected from the expanded ANEW database (Warriner et al., 2013) if their valence ratings were between 3.5 and 6.5

Table 5
The eight categorized word lists used in Experiment 2.

Category Name	Study Words		Generation Words
BIRD	Cardinal	Mockingbird	Falcon
	Crow	Ostrich	Hummingbird
	Duck	Pigeon	Parrot
	Eagle	Seagull	Penguin
	Finch	Woodpecker	Vulture
BUILDING	Clam	Drown	Boat
	Dive	Marooned	Dock
	Reef	Pollute	Fish
	Snorkel	Shark	Seaweed
	Whale	Wreck	Wave
EARTH	Cliff	Mountain	Grass
	Dirt	River	Ocean
	Glacier	Rock	Plain
	Hill	Stream	Tree
	Island	Valley	Volcano
FABRIC	Cloth	Spandex	Rayon
	Linen	Velvet	Silk
	Nylon	Wool	Cashmere
	Polyester	Cotton	Lace
	Satin	Denim	Suede
FISH	Bass	Salmon	Goldfish
	Beta	Snapper	Guppy
	Blowfish	Swordfish	Minnow
	Founder	Trout	Shark
	Halibut	Tuna	Whale
OCCUPATION	Accountant	Janitor	Dentist
	Athlete	Manager	Doctor
	Banker	Professor	Firefighter
	Businessman	Scientist	Lawyer
	Engineer	Secretary	Psychologist
READING	Article	Letter	Book
	Encyclopedia	Magazine	Comic
	Essay	Newspaper	Internet
	Flyer	Novel	Paper
	Journal	Textbook	Website
VEGETABLE	Asparagus	Onion	Beet
	Broccoli	Radish	Pepper
	Celery	Spinach	Potatoes
	Cucumber	Squash	Tomato
	Lettuce	Turnip	Zucchini

Each list included ten neutral Study words and five neutral Generation words. The words were selected from the updated version of the Battig and Montague norms (Van Overschelde et al., 2004).

and arousal ratings were <5.5 (on a scale from 1 to 9). If meeting these criteria was not possible, the exemplar was allocated to the Generation Phase list. Word2Vec was used again to ensure high conceptual relatedness within categories and low relatedness between categories (Mikolov et al., 2013). Finally, four neutral and four negative images were selected from the Nencki Affective Picture System (NAPS; Marchewka et al., 2014). We chose images based on the NAPS valence and arousal ratings and only selected those categorized as People and Faces (e.g., man and woman walking; wounded child). Negative images had average valence and arousal ratings of 2.23 and 7.34; neutral images had average valence and arousal ratings of 5.53 and 4.72 (on a scale from 1 to 9).

3.1.3. *Experiment 2 procedure*

3.1.3.1. Word study (phase 1). Participants viewed 80 category-exemplar pairs for six seconds each during the Study Phase (Fig. 4A) while verbally rating how typical the exemplar was of its category from one to seven. Participants' responses from the Study and Generation Phases were made verbally due to the eye tracker setup. These verbal responses were written down and recorded by the experimenter. For half of the categories, each exemplar was preceded by a negative image with the category name above it. For the other four categories, each exemplar was preceded by a neutral image along with the category name. A given category was always preceded by the same image. While the image and category name were on the screen for two seconds, participants were instructed to think of a simple narrative between the two. The purpose of this encoding task was to reinforce the mnemonic association between the categories and their corresponding negative or neutral images.

3.1.3.2. Distractor task #1. Participants completed a 45-s distractor task in which they were presented with a rapid stream of 0.5-s left- or right-facing arrow symbols in the middle of the screen ('<' or '>'). Participants indicated the direction of each arrow as quickly as possible by pressing the '1' or '2' key on their keyboard.

3.1.3.3. Semantic word generation (phase 2). Two negative (preceded by a negative image) and two neutral (preceded by a neutral image) categories were 'practiced' in the Generation Phase using category and word stem cues for extralist words (Fig. 4B). Each cue was on the screen for six seconds and was presented three times throughout this phase. Participants verbally reported their answers to the experimenter, who then typed the response. After giving each answer, participants had 3.5 s to verbally report a subjective effort rating (i.e., how difficult it was to come up with the correct word) on a scale from 1 ('very easy') to 5 ('very difficult').

3.1.3.4. Distractor task #2. After the Generation Phase, the participants completed the same 45-s distractor task again.

3.1.3.5. Memory test (phase 3). In the final phase of the experiment, we tested participants' memory for all 80 exemplar words from the Study phase (Fig. 4C). Category labels and partial word cues were presented for ten seconds each in a completely randomized order. The participants typed their answers, as pupillometry data was not collected during this phase.

3.1.3.6. Questionnaires. Following the RIF task, participants provided subjective valence and arousal ratings from 1 to 7 for all images from the experiment. They also rated the intrusiveness of the images on their recall and selected which category they believed corresponded to each image during the Study Phase. Finally, the participants completed the following questionnaires: State-Trait Anxiety Inventory (STAI; Spielberger, 1983); Beck Depression Inventory-II (BDI-II; Beck et al., 2011); and Persistent and Intrusive Negative Thoughts Scale (PINTS; Magson

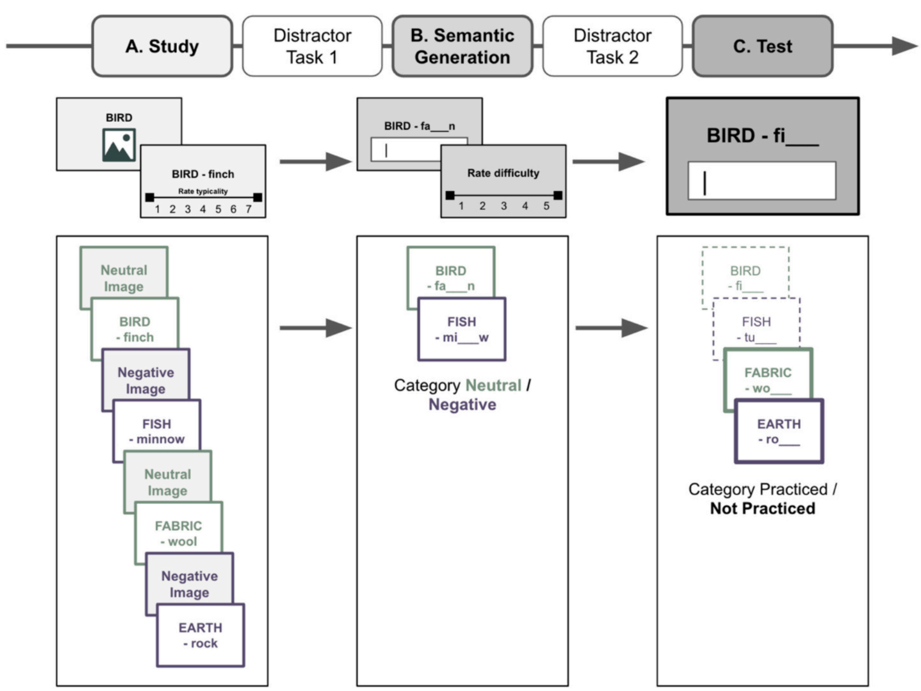

Fig. 4. Experiment 2 semantic retrieval-induced forgetting task. (A) Study Phase. On each trial, participants first viewed a neutral or negative image along with a category label and briefly imagined a narrative connecting them. They then viewed that same category with an exemplar from that category and rated the typicality of the exemplar from 1 to 7. (B) Generation Phase (semantic retrieval). Participants were shown half of the categories (two neutral and two negative) from the Study Phase, this time with word stems for new, related exemplars. On each trial, participants attempted to retrieve the cued word from semantic memory, then rated the word generation difficulty from 1 to 5. Each trial repeated three times throughout the phase. (C) Test Phase. Memory for all Study Phase words was tested by presenting category and word stem cues in a random order. In this example, worse performance for the BIRD and FISH items (category practiced) compared to the FABRIC and EARTH items (category not practiced) would indicate RIF.

et al., 2019). Questionnaire data was collected for exploratory purposes and is not analyzed here.

3.1.4. Response coding procedures

3.1.4.1. Cued recall accuracy coding procedure. Word responses from the Generation and Test Phases were marked as correct or incorrect by two independent raters (JG and AA). Discrepancies were resolved by a third rater (DC). A response was marked as correct if it was the correct word regardless of spelling or form (e.g., singular vs. plural), if there were enough letters for the grader to infer that the correct word was being typed, and if the response was a similar word with the same meaning (e.g., *occupation-ba__r* being written as *bankteller* vs. *banker*). Since the Generation words were not previously studied and instead retrieved from semantic memory, a different within-category word that fit the word stem was also marked as correct for Generation Phase words (e.g., *bird-pe__n* being written as *pelican* vs. *penguin*). The ICC was 0.97 for the Generation Phase and 0.96 for the Test Phase, indicating excellent reliability.

3.1.5. Analyses

3.1.5.1. Retrieval-induced forgetting analyses. To investigate whether the generation of neutral words induced forgetting of related, previously studied neutral words from categories that had been associated with negative and neutral images, the same generalized linear mixed-effects model was used in Experiment 2 that was used in Experiment 1. Random effects were specified for participant identity and word identity. Generation (generation, no generation), category valence (preceded by neutral image, preceded by negative image), and typicality were modeled as fixed-effect predictors. A generation-by-category valence interaction term was added back into the model, but as in Experiment 1 analyses, the model was outperformed by one without the interaction according to the BIC. The final model had a conditional R^2 of 0.37. It was then tested for negative and neutral category words separately.

As in Experiment 1, the best-fitting model did not include a generation-by-word valence interaction term, which was not significant when included. We again fitted a Bayesian logistic mixed model (estimated using MCMC sampling with 4 chains of 2000 iterations and a warmup of 1000) to explore whether RIF was equivalent for neutral and negative items. The model included the Generation-by-Valence interaction and Typicality, as well as random intercepts for participant identity and word identity. Priors over parameters were all set as normal (mean = 0.00, SD = 5.00; mean = 0.00, SD = 5.00; mean = 0.00, SD = 1.79; mean = 0.00, SD = 5.77) distributions. The model's explanatory power is substantial (R^2 = 0.26, 95% CI [0.24, 0.28]).

3.1.5.2. Effect of semantic retrieval success on memory test performance. To replicate the results from Experiment 1, we examined if the degree of RIF was correlated with successful semantic retrieval. Again, we computed average word generation success for each category that appeared during the Generation Phase and performed generalized linear mixed-effects modeling analyses using generation success, category valence, and test word typicality as fixed-effect predictors of memory test accuracy. Participant identity and word identity were specified as random effects. The model had a conditional R^2 of 0.37.

3.1.5.3. Effect of semantic retrieval effort on memory test performance. We also examined if the degree of RIF was correlated with the amount of subjective word generation effort during Phase 2. To achieve this, we computed the average word generation effort for each category that appeared during Phase 2. We performed generalized linear mixed-effects modeling analyses using generation effort, category valence, and test word typicality as fixed-effect predictors of final memory test accuracy. Participant identity and word identity were modeled as random effects. The model had a conditional R^2 of 0.37.

3.1.5.4. Relationship between semantic retrieval success and subjective effort. Finally, we sought to investigate the relationship between word generation success, a binary factor denoting whether participants successfully generated words, and subjective word generation effort, a factor with five ordinal levels ranging from 1 to 5. To explore the association between these variables, we performed generalized linear mixed-effects modeling analyses using generation effort as a fixed-effect predictor of generation success. The analysis was conducted on each trial of the Generation Phase rather than the averages used in the memory test analyses.

3.2. Results

3.2.1. Effects of category valence and semantic retrieval on memory test performance

We used generalized linear mixed-effects modeling to examine if semantic retrieval (i.e., generation) of neutral words induces forgetting of previously studied, related neutral exemplars that were associated with negative or neutral images (see Table 6A, Fig. 5). The results revealed a significant main effect of valence (β = 0.18, p = .0045) on memory test performance, such that recall was better for words that had been preceded by neutral images than those preceded by negative images. We also found a significant main effect of word generation (β = 0.28, p < .001) on memory test performance, such that memory was worse for words from categories that were practiced, or underwent semantic retrieval, during the Generation Phase compared to those that were not practiced. The effect of typicality was also significant (β = −12, p < .001), such that words that were rated as being more typical of their categories were easier to remember than words with lower typicality.

For the Bayesian model, the effect of the Generation-by-Valence interaction (Median = 0.02, 95% CI [−0.22, 0.26]) has a 56.17% probability of being positive (> 0), 28.48% of being significant (> 0.09), and 0.00% of being large (> 0.54). The estimation successfully converged (Rhat = 1.000) and the indices are reliable (ESS = 3377). The interaction had an odds ratio of 1.02 and a credible interval of [0.80, 1.29] for the Generation-by-Valence interaction. The credible interval included 1, suggesting equivalent RIF for words from negative and neutral categories.

These results replicate the findings from Experiment 1 showing that semantic retrieval induced forgetting of previously studied words from neutral and negative contexts, and words from neutral contexts were more memorable overall. Again, we repeated the analysis for neutral (see Table 6B) and negative (see Table 6C) words separately, which showed the same effect of word generation (β = 0.29, p < .001 for neutral; β = 28, p = .0014 for negative) on memory test performance.

3.2.2. Effect of semantic retrieval success on memory test performance

To replicate the results from Experiment 1, we performed generalized linear mixed modeling analyses by modeling semantic retrieval (i.e., generation) success, category valence, and test word typicality as fixed effects, and participant and word identity as random effects (see Table 7, Fig. 6A). The main effect of generation success on memory (β = 0.97, p = .0021) was significant, replicating the finding from Experiment 1 that worse semantic retrieval was correlated with worse recall of the previously studied words. The typicality of the test words was also a significant predictor of memory test accuracy (β = 0.12, p = .0022), such that higher typicality ratings were associated with better recall performance. Category valence was not a significant predictor of final recall (β = 0.14, p = .12).

3.2.3. Effect of semantic retrieval effort on memory test performance

To test if subjective semantic retrieval effort influences RIF, we performed generalized linear mixed modeling analyses by modeling

Table 6
Generalized linear mixed-effects modeling results for: (A) the effect of category valence, generation, and typicality on memory test performance for studied neutral words preceded by negative and neutral images, (B) the effect of generation and typicality on memory performance for studied neutral words preceded by neutral images, and (C) the effect of generation and typicality on memory performance for studied neutral words preceded by negative images.

Fixed Effects	A) Negative and Neutral RIF Model				B) Neutral RIF Model				C) Negative RIF Model			
	β	*SE*	*z*	*p*	*β*	*SE*	*z*	*p*	*β*	*SE*	*z*	*p*
(Intercept)	−0.11	0.22	−0.51	0.61	−0.15	0.28	−0.54	0.59	0.055	0.26	0.21	0.83
Category Valence	0.18	0.062	2.84	0.0045**								
Generation	0.28	0.062	4.52	< 0.001 ***	0.29	0.088	3.34	< 0.001 ***	0.28	0.087	3.19	0.0014**
Typicality	0.12	0.028	4.34	< 0.001 ***	0.16	0.039	3.96	< 0.001 ***	0.089	0.037	2.37	0.018*
Random Effects	Variance	*SD*			Variance	*SD*			Variance	*SD*		
Participant Identity	0.19	0.43			0.25	0.5			0.16	0.39		
Word Identity	1.72	1.31			1.65	1.28			1.65	1.28		

$^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

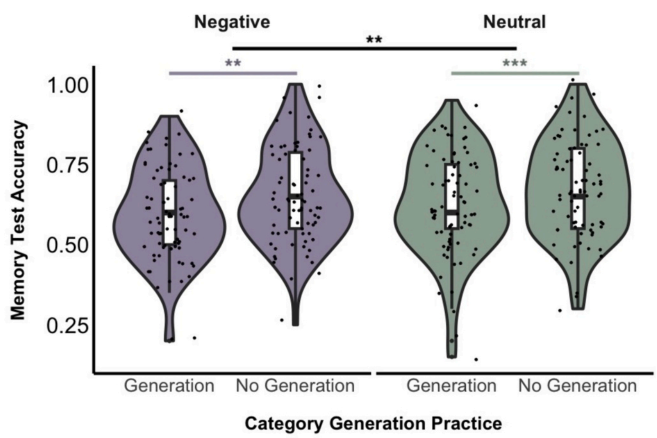

Fig. 5. Experiment 2 results showing effects of semantic retrieval on the retroactive forgetting of neutral words encoded under neutral or emotional contexts. Memory test accuracy results by valence (negative vs. neutral) and generation (generation vs. no generation). Test accuracy was significantly higher for neutral words (green) than negative words (purple). Test accuracy was also significantly higher for words in 'no generation' categories than 'generation' categories. Each point represents an individual participant's average memory test accuracy. Violin plots show the density distribution of the test accuracy scores for each condition. Box plots summarize key statistical measures: the box represents the interquartile range, the horizontal line represents the median, and the whiskers extend to minimum and maximum values (excluding outliers). The data points, violins, and box plots reflect raw data, while the statistical significance bars indicate the model-derived results. $^{**}p < .01$, $^{***}p < .001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Generalized linear mixed-effects modeling results for effect of generation success, category valence, and test word typicality on memory test performance for studied neutral words preceded by negative and neutral images.

Fixed Effects	Generation Success Model			
	β	*SE*	*z*	*p*
(Intercept)	−0.85	0.36	−2.38	0.017*
Generation Success	0.97	0.31	3.08	0.0021**
Category Valence	0.14	0.088	1.54	0.12
Test Word Typicality	0.12	0.038	3.07	0.0022**
Random Effects	Variance	*SD*		
Participant Identity	0.18	0.42		
Word Identity	1.72	1.31		

$^*p < .05$, $^{**}p < .01$.

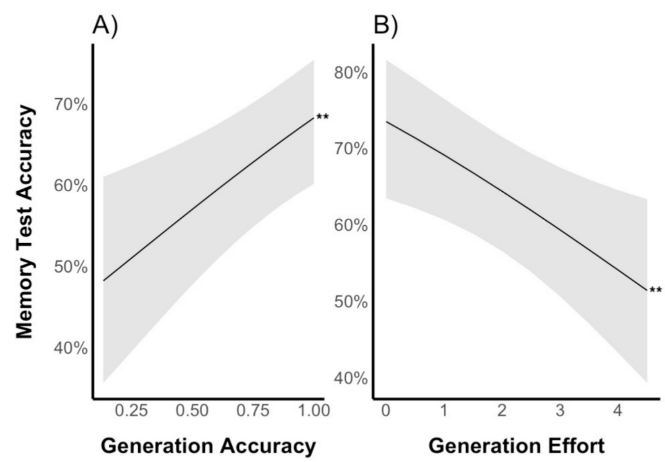

Fig. 6. Experiment 2 results showing relationship between the accuracy and effort of semantic memory retrieval (word generation) and retroactive forgetting. (A) This effect plot visualizes predicted values for memory test accuracy at each value from the generation accuracy term, with all other model predictors held constant. Lower generation accuracy predicted lower memory test accuracy scores for items from the initial Study Phase, replicating the result from Experiment 1. (B) This effect plot visualizes predicted values for memory test accuracy at each value from the generation effort term. Higher generation effort during semantic retrieval predicted lower memory test accuracy scores. Shaded areas represent 95% confidence bands. $^{**}p < .01$.

Table 8
Generalized linear mixed-effects modeling results for effect of generation effort and word category valence on memory test performance for studied neutral words preceded by negative and neutral images.

Fixed Effects	Generation Effort Model			
	β	*SE*	*z*	*p*
(Intercept)	0.38	0.33	1.15	0.25
Generation Effort	−0.21	0.08	−2.7	0.0070**
Valence	0.16	0.088	1.8	0.072
Test Word Typicality	0.12	0.038	3.02	0.0025**
Random Effects	Variance	*SD*		
Participant Identity	0.18	0.42		
Word Identity	1.72	1.31		

$^{**}p < .01$.

word generation effort, category valence, and test word typicality as fixed effects, and participant and word identity as random effects (see Table 8, Fig. 6B). The results revealed a main effect of generation effort ($\beta = -0.21$, $p = .0070$) on memory performance, such that higher category-specific effort ratings during semantic retrieval were associated with worse memory performance for previously studied words from those corresponding categories on the final memory test. The effect of test word typicality on memory was significant ($\beta = 0.12$, $p = .0025$), such that higher typicality ratings were associated with better recall performance. There was also a marginally significant effect of category valence ($\beta = 0.16$, $p = .072$), such that neutral words were more likely to be remembered than negative words.

3.2.4. Relationship between semantic retrieval success and subjective effort

To test if word generation success was correlated with subjective word generation effort, we performed generalized linear mixed modeling analyses by modeling word generation effort as a fixed effect, and participant and word identity as random effects. The results revealed a main effect of generation effort ($\beta = -2.25$, $p < .001$) on generation performance, such that higher effort ratings during semantic retrieval were associated with worse performance during word generation.

4. General discussion

Aversive memories can have a negative and lasting impact on well-being. Although much research has explored methods of weakening or updating emotional memories to remove their sting, most require individuals to recall different aspects of the original negative event. This explicit retrieval of emotional information can be distressing and induce negative affect. It can be especially problematic for people coping with highly arousing traumatic memories because it may discourage some from seeking or remaining in treatment. In this laboratory experiment, we examined if retrieving neutral semantic memories could lead to the indirect forgetting of related emotional episodic memories. Consistent with our prediction, we found that the semantic RIF paradigm using only neutral exemplars led to greater forgetting of recently studied and related neutral and negative words. We also found that failures to retrieve information from semantic memory were associated with greater forgetting of negative and neutral words. Importantly, these semantic RIF effects were also observed for neutral words encoded in neutral or negative contexts, suggesting that retroactive forgetting occurs irrespective of the intrinsic emotional significance of recently encoded information. Finally, we also found that highly effortful semantic retrieval processes track with the magnitude of RIF, consistent with models of retrieval inhibition.

Our main finding that semantic retrieval processes induce RIF of negative material extends prior work in several important ways. One major limitation of existing emotional RIF methods is that they rely on the presentation of negative material during retrieval practice (Blix and Brennen, 2012; Dehli and Brennen, 2009; Kobayashi and Tanno, 2013, 2015; Moulds and Kandris, 2006). In addition to potentially inducing negative moods, performing retrieval practice on negative memories can also reduce the likelihood of RIF (Koessler et al., 2009). While semantic retrieval has been shown to induce forgetting of negative words, this work is limited in that participants still had to generate negative memories (Kobayashi and Tanno, 2015). We expand upon these findings by demonstrating that retrieving *neutral* semantic information also induces emotional RIF.

We also found that more effortful and unsuccessful retrieval attempts from semantic memory correspond with greater retrieval-induced forgetting, an effect that aligns with inhibitory accounts of RIF. According to these frameworks, attempting to remember one item activates other related items, facilitating the inhibition of those related non-target items and increasing the likelihood of forgetting (Anderson, 2003). As in the 'impossible retrieval' RIF task in which attempting but failing to generate still induces RIF, our results suggest the existence of an inhibitory mechanism (Storm et al., 2006; Storm and Nestojko, 2010). The current findings expand upon prior work by showing that semantic generation does not need to be impossible, per se. Rather, it just needs to be sufficiently difficult to engage strong inhibitory processes in memory. Studies implementing an impossible retrieval design also typically pre-define and create separate categories of items that are impossible or possible to retrieve (Storm et al., 2006). We were able to examine how unsuccessful retrieval and increased effort have a linear effect on forgetting because it was theoretically possible to generate the solution to all words, which varied in difficulty. Thus, we show that forgetting corresponds with both the magnitude of retrieval success and the effort exerted during semantic retrieval.

A potentially more parsimonious explanation of these effort-related findings is that people who tried hard during the experiment did better both at generating words and at remembering words compared to those individuals who exerted less effort throughout the experiment. Yet, this idea is unlikely to explain the results. Because the linear mixed models included participant as a random intercept, we were able to account for individual differences in encoding strategies and effort across participants. Thus, the most appropriate interpretation of our results is that *when* a given individual showed more effort generating words from a certain category, they also showed greater forgetting of related, previously studied words from that same category. It is also important to note that the set-up of our effort-induced forgetting analysis is not a classic test of RIF, because there were no effort or generation scores that could be mapped onto the non-retrieval-practiced categories from Phase 1. Even so, because memory for the Generation items was significantly worse than memory for No Generation items on average, we interpret the linear mixed modeling results as likely reflecting the magnitude of RIF relative to baseline.

There are several alternative accounts to the inhibitory theory that warrant consideration. The competition-based account of RIF states that the act of retrieving some set of items strengthens those items and causes them to interfere with items in that category that were not retrieved (see Murayama et al., 2014 for a review of proposed RIF mechanisms). Although this type of competition may have contributed to some of the memory effects in the present study, it cannot fully account for forgetting. In our experiments, failed generation attempts predicted greater forgetting on the final memory test, suggesting that RIF can occur even without interference from successfully generated words. Thus, this result is better accounted for by an inhibitory mechanism than a competition-based one.

It is also possible that forgetting was driven by more passive forms of retroactive interference rather than an active inhibitory mechanism. Because we had individuals perform a semantic generation task rather than classic retrieval practice of the same exemplars, the retrieval of new items may have disrupted the consolidation of recently studied words (Alves and Bueno, 2017; Wixted, 2004). This possibility is supported by work showing that mental effort can induce retroactive interference (Wixted, 2004). The results of Experiment 2 appear to align with this account, showing that high subjective effort for generating new words predicted greater retrieval-induced forgetting. Critically, however, the present results and other semantic RIF literature suggest that rather than affecting memory generally, this interference can be targeted and constrained to specific categories of information. Thus, mental effort did not disrupt memory consolidation indiscriminately (Anderson and Hulbert, 2021; Wixted, 2004), an effect that is more consistent with retrieval-related inhibitory accounts of forgetting.

For both Experiments 1 and 2, we found that memory for items that were either negative themselves or encoded in negative contexts was impaired. Importantly, however, RIF still occurred for negative stimuli above and beyond this general impairment, suggesting that some aspects of negative episodic memories are not entirely resistant to RIF even if they are weaker to begin with. While this general impairment of negative memories was unexpected, it is possible that associative encoding strategies were more challenging to implement between the neutral categories and intrinsic (Experiment 1) or extrinsic (Experiment 2) emotional features. For example, it may be easier to associate/integrate neutral words with their neutral category cues (Murray and Kensinger, 2012); this effective unitization could, in turn, lead to better memory overall for neutral information in Experiment 1. Additionally, in Experiment 2, participants may have employed emotion regulation strategies to diminish their negative reactions to the highly aversive images. For example, if a participant tried to distract themselves from a gruesome image, this could impair their attention or processing of subsequent neutral information (Ortner et al., 2013). Critically, these encoding differences could explain the general memory advantage for neutral material over negative material across experiments. However, this would purely be a Study Phase effect and cannot fully explain why subsequent word generation elicits similar levels of RIF across categories.

Another possible explanation for why we did not observe a standard emotional memory enhancement effect is that arousal was not sufficiently high enough to enhance processing of negative items (e.g., Kensinger and Corkin, 2004). In Experiment 1, while the subjective arousal of word stimuli was rated as being significantly different between neutral and negative valences, ratings for negative words were not as high as the normative ratings indicated. This discrepancy suggests that differences in memory for negative and neutral words were more likely driven by the valence rather than arousal of the negative words. Prior work revealed that when using the typical RIF paradigm, forgetting of negative stimuli may decline as the emotional intensity of the stimuli increases (Kuhbandner et al., 2009). More research is warranted to investigate if semantic RIF is similarly reduced for highly arousing items.

In Experiment 2, we introduced emotional valence to neutral words by pairing lists with either neutral or highly arousing negative images. These visual stimuli may serve as better analogs of traumatic memories associated with PTSD, which are largely characterized by flashbacks of highly vivid visual information (Ehlers and Clark, 2000). Importantly, due to methodological differences, the experiments explore two similar yet distinct memory phenomena: Experiment 1 investigated the effects of semantic retrieval on forgetting of items that were intrinsically emotional or neutral; in contrast, Experiment 2 examined the effects of semantic retrieval on purely neutral items that were encoded under surrounding negative or neutral contexts. Interestingly, words that had been paired with negative images were more likely to be forgotten than those paired with neutral images. This is consistent with research suggesting that emotion can impair processing of neutral details. For example, Mather and Knight (2008) found that participants performed worse at remembering locations of neutral faces that had been previously associated with negative emotional images. Other work suggests that high emotional arousal impairs processing of neutral items when those items appear close by in time or space (Knight and Mather, 2009). Such impairments may be driven by the lower priority, or lower salience, of neutral stimuli compared to emotional stimuli (i.e., arousal-biased competition; see Mather and Sutherland, 2011 for a review; Cunningham et al., 2021).

Retrieval-induced forgetting could be problematic in real-world contexts when the stakes of remembering certain information are high. For example, eyewitnesses to a crime often show superior memory for details of a weapon, the emotion-eliciting stimulus, but impaired memory for peripheral neutral details ('weapon focus effect'; Loftus et al., 1987). Our data suggest that retrieving certain information can *further* weaken memory for those details, which could occur based when eyewitnesses recount related neutral experiences or other features of the event itself. Indeed, previous research supports the idea that repeated questioning can lead to retrieval-induced forgetting of eyewitness information (Camp et al., 2012). Our findings suggest that forgetting driven by semantic retrieval is a wholesale effect, and it does not matter if the conceptually related memoranda are intrinsically emotional or simply encoded during emotional events.

There are situations where emotional RIF may be beneficial if forgetting those details is desirable. From a therapeutic perspective, weakening negative episodic memories without re-exposing individuals to distressing information could lead to better treatment outcomes and protect subjective wellbeing. Although we employed semantic RIF in a laboratory setting, these results may shed some light on the potential therapeutic benefits of this indirect forgetting technique to treat negative memory symptoms in emotion-related disorders. We nevertheless acknowledge that there are several characteristics of PTSD that may limit the efficacy of semantic retrieval as a treatment option. One defining characteristic of PTSD is the involuntary recollection of a traumatic event by intrusive thoughts or vivid flashbacks (Mcnally, 2006). Research suggests that PTSD sufferers may exhibit cognitive deficits regarding memory (Vasterling et al., 1998; Vasterling and Brailey, 2005), executive control, and inhibition (McNally et al., 1998). Due to these executive deficits, RIF may not be appropriate for some PTSD patients. For example, words related to the trauma endured by PTSD patients are difficult to suppress, which is attributed to enhanced accessibility of trauma-related words as well as failure of inhibition (McNally et al., 1998). Additionally, while non-trauma-exposed controls show RIF effects for emotional material, both individuals with PTSD and individuals exposed to trauma do not (Amir et al., 2009). Given the critical role that inhibition plays in the mechanism of RIF (Anderson, 2003), RIF may be disrupted in PTSD patients, though it is unclear if this disruption also extends to semantic generation techniques. Therefore, targeting positive memories to engage RIF mechanisms may be more appropriate than targeting neutral memories for PTSD patients. Supporting this possibility, recounting positive autobiographical memories during therapy has been demonstrated to be more effective than traditional cognitive behavioral therapy for treating negative symptoms in depressed patients (Craske et al., 2019).

Beyond PTSD, semantic retrieval methods of RIF may be appropriate for patients with other mental health conditions (e.g., depression, anxiety), as well as for healthy individuals coping with everyday negative incidents. Several studies indicate that RIF can be used to induce forgetting of neutral and emotional autobiographical memories, providing evidence that the RIF effects can translate from laboratory-based experiments to more naturalistic clinical settings (Barnier et al., 2004; Harris et al., 2010; Hauer and Wessel, 2006; Somos et al., 2023). Our results suggest that forgetting may be attainable through *neutral semantic* retrieval, which would be beneficial as the retrieval of any part of the original emotional memory, or other negative memories, would not be required. Future research should explore applications of neutral semantic generation to forgetting negative autobiographical memories. Ultimately, this indirect memory intervention may provide a simple and effective way to promote well-being in both clinical settings and everyday life.

Credit authorship contribution statement

Jamie Greer: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Amna Ali:** Methodology, Investigation, Data curation. **Camille Laksman:** Methodology, Investigation, Data curation. **Ringo Huang:** Writing – review & editing, Methodology. **Mason McClay:** Writing – review & editing, Methodology. **David Clewett:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare no competing interests.

Data and code availability statement

All behavioral data and programming code will be made publicly available on the first author's Open Science Framework account upon acceptance of this manuscript.

Acknowledgements

This project was funded by university faculty retentions start-up funds to David Clewett.

References

Alves, M. V. C., & Bueno, O. F. A. (2017). Retroactive interference: Forgetting as an interruption of memory consolidation. *Temas Em Psicologia, 25*(3), 1055–1067. https://doi.org/10.9788/TP2017.3-07En

Amir, N., Badour, C. L., & Freese, B. (2009). The effect of retrieval on recall of information in individuals with posttraumatic stress disorder. *Journal of Anxiety Disorders, 23*(4), 535–540. https://doi.org/10.1016/j.janxdis.2008.10.012

Amir, N., Coles, M. E., Brigidi, B., & Foa, E. B. (2001). The effect of practice on recall of emotional information in individuals with generalized social phobia. *Journal of Abnormal Psychology, 110*(1), 76–82. https://doi.org/10.1037/0021-843X.110.1.76

Anderson, M. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*(4), 415–445. https://doi.org/10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

Anderson, M. C., & Hulbert, J. C. (2021). Active forgetting: Adaptation of memory by prefrontal control. *Annual Review of Psychology, 72*(1), 1–36. https://doi.org/10.1146/annurev-psych-072720-094140

Arnsten, A. F. T. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews Neuroscience, 10*(6), 410–422. https://doi.org/10.1038/nrn2648

Barber, S. J., & Mather, M. (2012). Forgetting in context: The effects of age, emotion, and social factors on retrieval-induced forgetting. *Memory & Cognition, 40*(6), 874–888. https://doi.org/10.3758/s13421-012-0202-8

Barnier, A., Hung, L., & Conway, M. (2004). Retrieval-induced forgetting of emotional and unemotional autobiographical memories. *Cognition & Emotion, 18*(4), 457–477. https://doi.org/10.1080/0269993034000392

Bartoń, K. (2023). MuMIn: Multi-Model Inference. R package version 1.47.5. Available at http://cran.r-project.org/package=MuMIn.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Bäuml, K.-H. (2002). Semantic generation can cause episodic forgetting. *Psychological Science, 13*(4), 356–360. https://doi.org/10.1111/1467-9280.00464

Bäuml, K.-H., & Kuhbandner, C. (2007). Remembering can cause forgetting—But not in negative moods. *Psychological Science, 18*(2), 111–115. https://doi.org/10.1111/j.1467-9280.2007.01857.x

Beck, A. T., Steer, R. A., & Brown, G. (2011). *Beck depression inventory–II [dataset]*. American Psychological Association. https://doi.org/10.1037/t00742-000

Blix, I., & Brennen, T. (2012). Retrieval-induced forgetting after trauma: A study with victims of sexual assault. *Cognition & Emotion, 26*(2), 321–331. https://doi.org/10.1080/02699931.2011.570312

Camp, G., Wesstein, H., & Bruin, A. B. H. (2012). Can questioning induce forgetting? Retrieval-induced forgetting of eyewitness information: Can questioning induce forgetting? *Applied Cognitive Psychology, 26*(3), 431–435. https://doi.org/10.1002/acp.2815

Campbell, J. I. D., & Thompson, V. A. (2012). Retrieval-induced forgetting of arithmetic facts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(1), 118–129. https://doi.org/10.1037/a0025056

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*(2), 153–170. https://doi.org/10.1016/j.jml.2009.04.004

Craske, M. G., Meuret, A. E., Ritz, T., Treanor, M., Dour, H., & Rosenfield, D. (2019). Positive affect treatment for depression and anxiety: A randomized clinical trial for a core feature of anhedonia. *Journal of Consulting and Clinical Psychology, 87*(5), 457–471. https://doi.org/10.1037/ccp0000396

Cunningham, T. J., Mattingly, S. M., Tlatenchi, A., Wirth, M. M., Alger, S. E., Kensinger, E. A., & Payne, J. D. (2021). Higher post-encoding cortisol benefits the selective consolidation of emotional aspects of memory. *Neurobiology of Learning and Memory, 180*, Article 107411. https://doi.org/10.1016/j.nlm.2021.107411

Dehli, L., & Brennen, T. (2009). Does retrieval-induced forgetting occur for emotional stimuli? *Cognition and Emotion, 23*(6), 1056–1068. https://doi.org/10.1080/02699930802285221

Dunn, E. W., & Spellman, B. A. (2003). Forgetting by remembering: Stereotype inhibition through rehearsal of alternative aspects of identity. *Journal of Experimental Social Psychology, 39*(5), 420–433. https://doi.org/10.1016/S0022-1031(03)00032-5

Ehlers, A., & Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behaviour Research and Therapy, 38*(4), 319–345. https://doi.org/10.1016/S0005-7967(99)00123-0

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2024). Rstanarm: Bayesian applied regression modeling via Stan. R package version 2.32.1 https://mc-stan.org/rstanarm/.

Harris, C. B., Sharman, S. J., Barnier, A. J., & Moulds, M. L. (2010). Mood and retrieval-induced forgetting of positive and negative autobiographical memories. *Applied Cognitive Psychology, 24*(3), 399–413. https://doi.org/10.1002/acp.1685

Hauer, B. J. A., & Wessel, I. (2006). Retrieval-induced forgetting of autobiographical memory details. *Cognition and Emotion, 20*(3–4), 430–447. https://doi.org/10.1080/02699930500342464

Kensinger, E. A., & Corkin, S. (2004). Two routes to emotional memory: Distinct neural processes for valence and arousal. *Proceedings of the National Academy of Sciences, 101*(9), 3310–3315. https://doi.org/10.1073/pnas.0306408101

Kensinger, E. A., Garoff-Eaton, R. J., & Schacter, D. L. (2006). Memory for specific visual details can be enhanced by negative arousing content. *Journal of Memory and Language, 54*(1), 99–112. https://doi.org/10.1016/j.jml.2005.05.005

Knight, M., & Mather, M. (2009). Reconciling findings of emotion-induced memory enhancement and impairment of preceding items. *Emotion, 9*(6), 763–781. https://doi.org/10.1037/a0017281

Kobayashi, M., & Tanno, Y. (2013). Retrieval-induced forgetting of words with negative emotionality. *Memory, 21*(3), 315–323. https://doi.org/10.1080/09658211.2012.725741

Kobayashi, M., & Tanno, Y. (2015). Remembering episodic memories is not necessary for forgetting of negative words: Semantic retrieval can cause forgetting of negative words. *Psychonomic Bulletin & Review, 22*(3), 766–771. https://doi.org/10.3758/s13423-014-0719-x

Koessler, S., Engler, H., Riether, C., & Kissler, J. (2009). No retrieval-induced forgetting under stress. *Psychological Science, 20*(11), 1356–1363. https://doi.org/10.1111/j.1467-9280.2009.02450.x

Kuhbandner, C., Bäuml, K.-H., & Stiedl, F. C. (2009). Retrieval-induced forgetting of negative stimuli: The role of emotional intensity. *Cognition & Emotion, 23*(4), 817–830. https://doi.org/10.1080/02699930802204768

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about "weapon focus". *Law and Human Behavior, 11*(1), 55–62. https://doi.org/10.1007/BF01044839

Lüdecke, D. (2023). sjPlot: Data visualization for statistics in social science. R package version 2.8.12. Available at https://cran.r-project.org/web/packages/sjPlot/index.html.

MacLeod, C. M., & Kampe, K. E. (1996). Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 132–142. https://doi.org/10.1037/0278-7393.22.1.132

Magson, N. R., Rapee, R. M., Fardouly, J., Forbes, M. K., Richardson, C. E., Johnco, C. J., & Oar, E. L. (2019). Measuring repetitive negative thinking: Development and validation of the Persistent and Intrusive Negative Thoughts Scale (PINTS). *Psychological Assessment, 31*(11), 1329–1339. https://doi.org/10.1037/pas0000755

Marchewka, A., Żurawski, Ł., Jednoróg, K., & Grabowska, A. (2014). The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods, 46*(2), 596–610. https://doi.org/10.3758/s13428-013-0379-1

Mather, M., & Knight, M. (2008). The emotional harbinger effect: Poor context memory for cues that previously predicted something arousing. *Emotion, 8*(6), 850–860. https://doi.org/10.1037/a0014087

Mather, M., & Sutherland, M. R. (2011). Arousal-biased competition in perception and memory. *Perspectives on Psychological Science, 6*(2), 114–133. https://doi.org/10.1177/1745691611400234

Matsumoto, N., Mochizuki, S., Marsh, L., & Kawaguchi, J. (2021). Repeated retrieval of generalized memories can impair specific autobiographical recall: A retrieval induced forgetting account. *Journal of Experimental Psychology: General, 150*(9), 1825–1836. https://doi.org/10.1037/xge0001028

Mcnally, R. (2006). Cognitive abnormalities in post-traumatic stress disorder. *Trends in Cognitive Sciences, 10*(6), 271–277. https://doi.org/10.1016/j.tics.2006.04.007

McNally, R. J., Metzger, L. J., Lasko, N. B., Clancy, S. A., & Pitman, R. K. (1998). Directed forgetting of trauma cues in adult survivors of childhood sexual abuse with and without posttraumatic stress disorder. *Journal of Abnormal Psychology, 107*(4), 596–601. https://doi.org/10.1037/0021-843X.107.4.596

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. https://doi.org/10.48550/ARXIV.1301.3781

Moulds, M. L., & Kandris, E. (2006). The effect of practice on recall of negative material in dysphoria. *Journal of Affective Disorders, 91*(2–3), 269–272. https://doi.org/10.1016/j.jad.2006.01.003

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin, 140*(5), 1383–1409. https://doi.org/10.1037/a0037505

Murray, B. D., & Kensinger, E. A. (2012). The effects of emotion and encoding strategy on associative memory. *Memory & Cognition, 40*(7), 1056–1069. https://doi.org/10.3758/s13421-012-0215-3

Ortner, C. N. M., Zelazo, P. D., & Anderson, A. K. (2013). Effects of emotion regulation on concurrent attentional performance. *Motivation and Emotion, 37*(2), 346–354. https://doi.org/10.1007/s11031-012-9310-9

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Penolazzi, B., Stramaccia, D. F., Braga, M., Mondini, S., & Galfano, G. (2014). Human memory retrieval and inhibitory control in the brain: Beyond correlational evidence. *The Journal of Neuroscience, 34*(19), 6606–6610. https://doi.org/10.1523/JNEUROSCI.0349-14.2014

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., … Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications, 9*(1), 963. https://doi.org/10.1038/s41467-018-03068-4

Rupprecht, J., & Bäuml, K.-H. T. (2016). Retrieval-induced forgetting in item recognition: Retrieval specificity revisited. *Journal of Memory and Language, 86*, 97–118. https://doi.org/10.1016/j.jml.2015.09.003

Shaw, J. S., Bjork, R. A., & Handal, A. (1995). Retrieval-induced forgetting in an eyewitness-memory paradigm. *Psychonomic Bulletin & Review, 2*(2), 249–253. https://doi.org/10.3758/BF03210965
Somos, E., Mazzoni, G., Gatti, D., & Jellema, T. (2023). "Be careful what you recall": Retrieval-induced forgetting of genuine real-life autobiographical memories. *Quarterly Journal of Experimental Psychology, 76*(1), 84–92. https://doi.org/10.1177/17470218221078499
Spielberger, C. (1983). *State-trait anxiety inventory for adults: Manual, instrument, and scoring guide*. Menlo Park, CA: MindGarden.
Storm, B. C., Bjork, E. L., Bjork, R. A., & Nestojko, J. F. (2006). Is retrieval success a necessary condition for retrieval-induced forgetting? *Psychonomic Bulletin & Review, 13*(6), 1023–1027. https://doi.org/10.3758/BF03213919
Storm, B. C., & Jobe, T. A. (2012). Retrieval-induced forgetting predicts failure to recall negative autobiographical memories. *Psychological Science, 23*(11), 1356–1363. https://doi.org/10.1177/0956797612443837
Storm, B. C., & Nestojko, J. F. (2010). Successful inhibition, unsuccessful retrieval: Manipulating time and success during retrieval practice. *Memory, 18*(2), 99–114. https://doi.org/10.1080/09658210903107853
Talmi, D., Lohnas, L. J., & Daw, N. D. (2019). A retrieved context model of the emotional modulation of memory. *Psychological Review, 126*(4), 455–485. https://doi.org/10.1037/rev0000132
Tempel, T., & Frings, C. (2013). Resolving interference between body movements: Retrieval-induced forgetting of motor sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1152–1161. https://doi.org/10.1037/a0030336
Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003
Vasterling, J. J., & Brailey, K. (2005). Neuropsychological findings in adults with PTSD. In *Neuropsychology of PTSD: Biological, cognitive, and clinical perspectives* (pp. 178–207). The Guilford Press.
Vasterling, J. J., Brailey, K., Constans, J. I., & Sutker, P. B. (1998). Attention and memory dysfunction in posttraumatic stress disorder. *Neuropsychology, 12*(1), 125–133. https://doi.org/10.1037/0894-4105.12.1.125
Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x
Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., & Dunnington, D. (2023). Ggplot2: Create elegant data Visualisations using the grammar of graphics. R package version 3.4.3. Available at https://cran.r-project.org/web/packages/ggplot2/index.html.
Wimber, M., Rutschmann, R. M., Greenlee, M. W., & Bäuml, K.-H. (2009). Retrieval from episodic memory: Neural mechanisms of interference resolution. *Journal of Cognitive Neuroscience, 21*(3), 538–549. https://doi.org/10.1162/jocn.2009.21043
Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*(1), 235–269. https://doi.org/10.1146/annurev.psych.55.090902.141555
Yonelinas, A. P., & Ritchey, M. (2015). The slow forgetting of emotional episodic memories: An emotional binding account. *Trends in Cognitive Sciences, 19*(5), 259–267. https://doi.org/10.1016/j.tics.2015.02.009