

Journal of Experimental Psychology: General

On the Search for a Selective and Retroactive Strengthening of Memory: Is There Evidence for Category-Specific Behavioral Tagging?

Felix Kalbe and Lars Schwabe

Online First Publication, July 15, 2021. <http://dx.doi.org/10.1037/xge0001075>

CITATION

Kalbe, F., & Schwabe, L. (2021, July 15). On the Search for a Selective and Retroactive Strengthening of Memory: Is There Evidence for Category-Specific Behavioral Tagging?. *Journal of Experimental Psychology: General*. Advance online publication. <http://dx.doi.org/10.1037/xge0001075>

On the Search for a Selective and Retroactive Strengthening of Memory: Is There Evidence for Category-Specific Behavioral Tagging?

Felix Kalbe and Lars Schwabe

Department of Cognitive Psychology, Institute of Psychology, Universität Hamburg

Storing motivationally salient experiences preferentially in long-term memory is generally adaptive. Although such relevant experiences are often immediately obvious, a problem arises when the relevance of initially ambiguous events becomes evident sometime after encoding. Is there a mechanism that enables the retroactive enhancement of specific memories? Recent evidence suggests the existence of such a mechanism that selectively strengthens weak memories for neutral stimuli from one category when their respective category gains motivational significance later. Although such a selective retroactive memory enhancement has considerable implications for adaptive memory, evidence for this phenomenon is based on only few studies. Here, we report data from four attempts to replicate category-specific retroactive memory enhancements for neutral stimuli from a category that was later predictive of aversive electric shocks. Although our data showed enhanced memory for the arousing stimuli themselves as well as related subsequent stimuli, none of our experiments provided any evidence for category-specific retroactive memory enhancement when strictly replicating the analysis strategy from the original study. In an additional analysis focusing on high confidence memory only, one of four experiments indicated a significant retroactive memory effect but only in corrected recognition and not in d' based on signal detection theory. In an analysis pooled across all experiments, we found a small but significant retroactive memory effect again solely for high-confidence corrected recognition, although the corresponding Bayesian analysis indicated even substantial evidence for the null hypothesis. Overall, our data cast doubt on the reliability and generalizability of the proposed selective retroactive enhancement of initially weak memory.


Keywords: adaptive memory, behavioral tagging, episodic memory, reproducibility

Supplemental materials: <https://doi.org/10.1037/xge0001075.supp>

Our memories provide not only a window into the past but may also guide our future behavior. In particular, detailed memories of past experiences allow predicting future events as well as the potential

consequences of actions and can therefore serve as a basis for optimized choices in complex environments (Gershman & Daw, 2017; Murty et al., 2016). However, of the numerous experiences that we make every day, only few are of significant value for future decisions. According to the theory of adaptive memory, these motivationally significant experiences should be preferentially stored in episodic memory (Nairne et al., 2007; Nairne & Pandeirada, 2008; Shohamy & Adcock, 2010). Phylogenetically, such an adaptive memory might have been critical to survival by allowing the identification and subsequent avoidance of potentially threatening situations, thereby improving fitness (Nairne & Pandeirada, 2008). The preferential memory processing is relevant because limited memory resources during both encoding and retrieval should optimally be reserved for motivationally relevant experiences.

Such motivationally salient experiences are usually immediately obvious to an individual. Exciting or stressful experiences elicit physiological arousal during encoding, a well-known factor that promotes episodic memory formation (Cahill & McGaugh, 1998; LaBar & Cabeza, 2006; McGaugh, 2018; Schwabe et al., 2012; Vogel & Schwabe, 2016). However, other events appear initially neutral or mundane and their link to important consequences is only later revealed. Consider a bank customer entering her local branch as usual, when another presumed customer is leaving in a hurry. She barely

Lars Schwabe  <https://orcid.org/0000-0003-4429-4373>

Part of the data of Experiments 1 and 2 were analyzed to address a research question unrelated to the present article and has been reported in Kalbe, F., & Schwabe, L. (2020). Beyond arousal: Prediction error related to aversive events promotes episodic memory formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 234–246.

We gratefully acknowledge the assistance of Friederike Baier, Pia D'Agostino, Leandra Feldhusen, Hülya Keskin, Manuel Krohn, Vincent Kühn, Moana Lamm, Fabian Schacht, Felix Schiborn, Celine Schneller, Anne-Sophie Siegel, Elizabeth Sievert, Seher Teymuroglu, and Till Thelosen during data collection. We further thank Joseph Dunsmoor for providing the stimulus materials used in Experiments 2, 3, and 4 and for helpful advice on instructions and experimental procedure. The behavioral data analyzed in this series of experiments can be found at <https://osf.io/qpm3t/>.

Correspondence concerning this article should be addressed to Lars Schwabe, Department of Cognitive Psychology, Institute of Psychology, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. Email: Lars.Schwabe@uni-hamburg.de

notices his face. Soon it is revealed to her that the person she saw leaving had just robbed the bank and she realizes that she will soon be asked by the police to give a detailed description of the alleged robber's unmasked face. Because the initial encounter with the alleged robber was not particularly remarkable and therefore not paralleled by significant arousal, his face was not encoded preferentially. Is there still a mechanism to make such initially weak memories last? Adaptive memory would call for a mechanism that temporarily and nonselectively stores recent experiences and transfers them to long-term memory when a motivationally significant event follows within a certain time window. Indeed, such a mechanism was first discovered at the synaptic level and inspired the *tag and capture hypothesis* (Frey & Morris, 1997, 1998; Martin & Kosik, 2002; Rogerson et al., 2014). According to this hypothesis, at least two distinguishable steps are necessary to achieve long-term potentiation for initially weak experiences, the dominant neurophysiological model of long-term memory (Bliss & Collingridge, 1993; Malenka & Nicoll, 1999). First, weak stimulation of a neuron creates a local transient tag at a synapse, which decays within hours and by itself is insufficient to create long-lasting memories. To produce long-term potentiation, additional plasticity-related proteins are required that result from a stronger stimulation of the neuron. These proteins bind to the synaptic tag set earlier (i.e., the capturing step) inducing long-term physiological changes in synaptic signaling. Critically, plasticity related proteins evoked through strong stimulation of the neuron can bind to synaptic tags set earlier through unrelated weak stimulation and therefore create lasting memories for events that would by themselves be too weak to produce long-term potentiation (Moncada & Viola, 2007; Redondo & Morris, 2011). This mechanism therefore provides a neurophysiological basis for the retroactive memory enhancement of events with an initially unclear motivational significance (Moncada et al., 2015).

Evidence that the synaptic tag and capture hypothesis can be translated to behavior has been found in both rodents (Almaguer-Melian et al., 2012; Ballarini et al., 2009; de Carvalho Myskiw et al., 2013; Moncada & Viola, 2007; Wang et al., 2010) and more recently in humans (Ballarini et al., 2013). In paradigms demonstrating a *behavioral tagging* mechanism, participants first superficially encode stimuli. This encoding session is then followed by either a significant event (e.g., an aversive or novel experience) or a nonsignificant control event. Subsequent memory tests typically show that the significant event—compared with a neutral control event—retroactively enhanced memory for the previously encoded stimuli. In these paradigms, retroactive memory enhancements are usually unspecific in the sense that an event enhances memory for any stimuli encoded within a certain time window before the significant event, even if these are not directly linked to the latter. In the case of the bank robbery, such unrelated details might include the color of the tie the bank clerk was wearing at the time of the robbery. From an adaptive memory perspective, promoting memory for such irrelevant details might be regarded as suboptimal when they lack any predictive value for the memory-promoting event.

A recent study suggests that there is—in addition to rather broad and unspecific behavioral tagging—a retroactive memory enhancement that is highly specific (Dunsmoor et al., 2015). This study combined an incidental encoding task with a fear learning procedure. In a preconditioning phase, participants first encoded neutral pictures of animals and tools and were asked to indicate to which of the two categories a picture

belonged. Following this weak encoding session, in a Pavlovian fear conditioning phase, additional, previously unseen pictures from the same two categories were presented. Pictures from one of the two categories (i.e., either animals or tools; CS⁺) were followed by an aversive electric shock in two thirds of all trials, while pictures from the remaining category (CS⁻) were never followed by a shock. Whether shocks followed pictures of animals or tools was counterbalanced across participants. A postconditioning phase with an identical procedure as the preconditioning phase but novel stimuli followed the fear-conditioning phase. To test participants' memory for stimuli from the three encoding phases, a surprise recognition test followed either immediately, 6 hr, or 24 hr later (manipulated between subjects). In this recognition test, participants saw all previously presented pictures of animals and tools together with the same number of previously unseen (new) pictures from both categories and classified each picture as either *old* or *new*. Results showed an enhanced recognition performance for CS⁺ pictures encoded during fear-conditioning compared with CS⁻ pictures encoded in the same phase in all three delay groups. In the 24-hr delay group, this CS⁺ memory carried over to pictures presented after the fear-conditioning phase, although these items were never paired with a shock themselves. Most importantly, however, the authors found category-specific retroactive memory enhancements in both the 6-hr and 24-hr delay groups, as indicated by better recognition of CS⁺ pictures encoded before the fear-conditioning compared with CS⁻ pictures encoded in the same phase. This finding is particularly remarkable because participants had no information about shock contingencies being linked to one of the two categories when these pictures were encoded. When the recognition test followed immediately after the encoding, no category-specific retroactive memory enhancement was observable, suggesting the critical involvement of consolidation processes. Interestingly, there also was a negative linear relationship between the size of the retroactive memory effect and the temporal proximity of preconditioning items to the fear-conditioning procedure, suggesting that pictures from the CS⁺ category that were encoded first (i.e., furthest from the following fear-conditioning) received the strongest memory enhancement. Furthermore, another group of participants encoded stimuli from the preconditioning phase more strongly through repeated presentation of each picture. These participants showed no signs of category-specific retroactive memory enhancement after 24 hr, indicating that only initially weak memories are susceptible to this effect, a finding that is congruent with the literature on synaptic tagging (Frey & Morris, 1997, 1998; Martin & Kosik, 2002; Rogerson et al., 2014).

Another study from the same group of authors showed that selective, category-specific retroactive memory enhancements cannot only be triggered through aversive events, but also through reward (Patil et al., 2017). Following a similar design as the study by Dunsmoor et al. (2015), the authors showed that memory for initially neutral pictures of animals and tools could be enhanced for the category that was later associated with high compared with low reward opportunities in a delayed matching-to-sample task. Notably, in this task, participants were rewarded for correct responses, whereas shocks were

independent of participants' actions in the study by Dunsmoor et al. (2015). In contrast to these findings, another study from an independent lab using a similar classical conditioning procedure as Dunsmoor et al. (2015) featuring monetary reward instead of aversive shocks obtained no evidence for category-specific retroactive memory enhancement (Oyarzún et al., 2016). To our knowledge, no other studies so far have investigated category-specific retroactive memory enhancement, neither in the aversive, nor in the appetitive domain.

The findings showing a selective, retroactive memory enhancement are exciting; they provide novel insights into how our memory works and may have considerable practical implications for clinical or legal settings. A selective behavioral tagging mechanism may also inspire new tools for boosting memory retrospectively. Given the far-reaching implications of selective, retroactive memory enhancements, we initially aimed to shed light on the cognitive mechanisms underlying this effect. However, what started as an attempt to unravel the fundamental mechanisms underlying selective behavioral tagging, turned out to be a search for the phenomenon itself. We present here evidence from four experiments aimed to replicate findings of category-specific retroactive memory enhancement through aversive electric shocks (Dunsmoor et al., 2015).

Experiment 1: Testing the Fear-Related Category-Specific Retroactive Memory Enhancement

Experiment 1 was designed to replicate findings of category-specific retroactive memory enhancement in the context of an aversive learning task (Dunsmoor et al., 2015). Because the experiment of Dunsmoor et al. (2015) showed that observed retroactive memory effects were most pronounced in a recognition test 24 hr after encoding, we used here a 24-hr interval between encoding and recognition test. Instead of the original stimulus set, we used pictures that were conceptually very similar to those used by Dunsmoor et al. (2015); that is, also pictures from the categories 'animals' and 'tools'. Procedural differences included the placing of the shock electrode on the lower leg (rather than on the wrist as in the original study) and employing a two-stage recognition test (rather than a single-stage as in the original study) that first asked participants to indicate whether an item was old or new, followed by their certainty with this decision. As further discussed below, we implemented a different CS-UCS timing compared with Dunsmoor et al. (2015). Finally, we did not control for stimulus typicality across encoding phases because this aspect was not mentioned in Dunsmoor et al. (2015). Instead, it was only revealed during later stages of the peer-review process for this article that Dunsmoor et al. (2015) controlled for stimulus typicality. This aspect is later explicitly addressed in Experiment 4.

Method

Participants

Forty-four healthy participants (30 women) between 19 and 33 years of age took part in this experiment ($M = 25.05$, $SD = 3.75$). This sample size was based on an a priori sample size calculation with G*Power 3 (Faul et al., 2007). Dunsmoor et al. (2015) reported retroactive memory improvements from a paired t -test for items conceptually related to the CS⁺ compared with items related

to the CS⁻ in the 24-hr retrieval group with weak encoding ($n = 30$) and obtained a t value of 2.48 with an effect size of $d_{av} = .41$. Based on this information, Cohen's d_z , another measure of effect size in within-subject designs used by G*Power, can be calculated using the following formula (Lakens, 2013):

$$\text{Cohen's } d_z = \frac{t}{\sqrt{n}}$$

Using the values reported by Dunsmoor et al. (2015) yielded the following estimate for Cohen's d_z :

$$\text{Cohen's } d_z = \frac{2.48}{\sqrt{30}} = 0.45$$

We treated this effect size as a point estimate for the category-specific retroactive memory effect in our power analysis. This indicated that, using a two-tailed paired t -test with $\alpha = .05$, at least 41 participants would be required to detect such an effect with 80% certainty. This target sample size also represents an approximately 40% increase compared with the 24-hr group in the original study ($n = 30$). Exclusion criteria for participation in this experiment comprised any current or past physical or mental illness, electric medical devices such as pacemakers, and pregnancy in women. Participants gave written informed consent prior to testing and received a monetary compensation of 20€ after completing the experiment. The ethics committee of the Faculty of Psychology and Human Movement Sciences of the Universität Hamburg approved the study protocol.

Materials

As in the original study, stimuli were 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. We acquired photographs from the Bank of Standardized Stimuli (Brodeur et al., 2010; Brodeur et al., 2014) and from publicly available Internet sources. All photographs were of neutral valence and selected to be unique exemplars of their respective category. For example, there were not two different photographs of dogs or two different photographs of hammers. From the total pool of 360 photographs, 180 (90 animals, 90 tools) were randomly selected per participant to serve as learning items, while the remaining 180 served as lures for the surprise recognition test on the second experimental day. The 180 learning items were then randomly allocated to the three different incidental encoding phases for the first experimental day, such that each phase featured 30 photographs of animals and 30 photographs of tools.

Procedure

The first experimental day featured an incidental encoding session with three phases: a preconditioning phase, a fear conditioning phase, and a postconditioning phase. Approximately 24 hr later, participants completed a surprise recognition test for photographs that had been presented in all three encoding phases on the previous day. Unlike in the study by Dunsmoor et al. (2015), we did not vary the interval between encoding and recognition test between subjects but kept it fixed at 24 hr, because the 24 hr group had previously shown the clearest evidence for both category-specific retroactive and prospective memory enhancement. Additionally, another study featuring a reward learning task also demonstrated category-specific

retroactive memory enhancement only after a 24-hr interval, but not in an immediate recognition test, suggesting a crucial role of a sufficiently long consolidation period before the recognition test (Patil et al., 2017).

Upon arrival on the first experimental day, participants gave written informed consent and received detailed written instructions about the following three learning phases. Importantly, they were not informed that the study investigated episodic memory, nor that a recognition test would follow on the second experimental day. In the *preconditioning phase*, participants saw 30 photographs of animals and 30 photographs of tools in a pseudorandomized order, such that no more than three photographs from the same category could appear in a row. Each stimulus was presented for 2.5 s, during which participants should indicate whether the photograph showed an animal or a tool by pressing the '1' or '2' button on the computer keyboard (see Figure 1). Each stimulus was followed by a black fixation cross on a white background for $6 \text{ s} \pm 2 \text{ s}$. The allocation of buttons to each of the two categories was counterbalanced across participants. The total duration of the preconditioning phase was approximately 8 min.

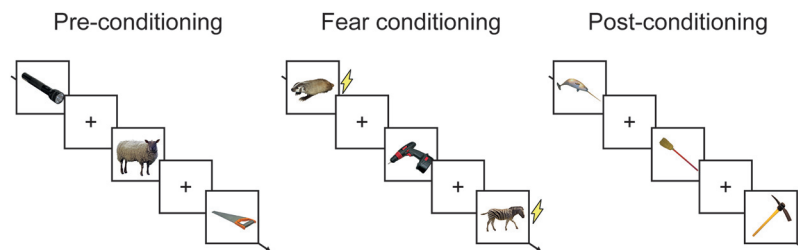
Before the *conditioning phase*, we attached electrodes on the distal phalanx of the second and third finger of the left hand to record skin conductance responses (SCRs). Skin conductance was measured using a MP-160 BIOPAC system (BIOPAC systems, Goleta, CA). An additional STM-200 module (BIOPAC systems, Goleta, CA) was connected to the MP-160 for electrical stimulation. The stimulation electrode was placed on the right lower leg, approximately 25 cm centrally above the heel. To determine the individual stimulation intensity, we used a standardized procedure consisting of twelve 200-ms single-pulse shocks with an initial intensity of 20 V. After each trial, participants rated the shock that they had just received as either painful or not painful in a forced-choice fashion. Whenever a shock was rated as not painful, its intensity was increased slightly in the following trial. Similarly, whenever participants rated a shock as painful, the intensity was decreased slightly. The goal was to select an intensity that participants perceived as

unpleasant, but not painful. In total, these steps following the preconditioning phase took approximately 10 min.

The following conditioning phase again consisted of 30 photographs of animals and 30 photographs of tools, none of which had been presented before. As in the preconditioning phase, stimuli were presented in a pseudorandomized order, so that no more than three photographs from the same category appeared in a row. Each photograph was presented centrally on the screen for 4.5 s, during which participants were instructed to make a binary prediction about the possible occurrence of a following shock using the '1' and '2' buttons on the keyboard, corresponding to *no shock* and *shock*, respectively. In 20 of the 60 trials in this phase, a 200-ms electric shock was presented immediately after the offset of the photograph. Note that in Dunsmoor et al. (2015), shocks coterminated with photograph presentation, leading to a 200ms relative offset of the shock in our replication attempt. This procedural difference was unintentional and addressed in a later experiment (Experiment 4).

Importantly, shock contingencies were linked to the item categories, such that one image category (e.g., tools) served as the CS^+ category, whereas the remaining category (e.g., animals) was never paired with a shock and thus served as the CS^- category. Whether photographs of animals or tools served as the CS^+ category was counterbalanced across participants. In CS^+ trials, the shock probability was two thirds, with a fixed number of 20 shocks occurring in the 30 CS^+ trials. In CS^- trials, on the other hand, none of the photographs was followed by a shock. Participants were not informed about category-shock contingencies but had to learn them by trial and error. To avoid that participants could misinterpret shocks as consequences of their actions, they were explicitly told that their choices had no effect on the probability that a shock would occur (Dunsmoor et al., 2015). Each trial was followed by a black fixation cross on a white background for $8 \pm 2 \text{ s}$, which enabled measuring the relatively slow SCRs elicited by electric shocks and allowed skin conductance levels to return to baseline before the next trial started. The total duration of the conditioning phase was approximately 12 min. After the

Figure 1
Procedure in Experiments 1–4



Note. In each phase, participants saw 60 unique photographs of animals and tools. During pre- and postconditioning, they were instructed to categorize each photograph as an animal or tool. During fear conditioning, photographs from one category (CS^+ ; animals in the example above) were followed by an electric shock in two-thirds of all trials, whereas photographs of the remaining category (CS^- ; tools in the example above) were never followed by a shock. Whether photographs of animals or tools served as the CS^+ category was counterbalanced across participants. For each photograph, participants were instructed to indicate whether they expected that a shock would follow. Note that in Experiment 3, the interval between preconditioning and Pavlovian fear conditioning was increased by 10 min, based on previous reports that this would lead to increased category-specific retroactive memory enhancement (Dunsmoor et al., 2015). Approximately 24h after encoding, participants completed a surprise recognition test in which they saw all previously presented photographs of animals and tools together with the same number of new photographs and indicated for each of them whether they thought it had been presented on the previous day. See the online article for the color version of this figure.

conditioning phase, we removed both the SCR- and the shock-electrodes. Participants then rated the shock intensity on a scale from 1 (*not unpleasant at all*) to 10 (*extremely unpleasant*).

The subsequent *postconditioning phase* consisted of 60 previously unseen photographs (30 animals and 30 tools) and otherwise followed an identical procedure as the preconditioning phase. Thus, the duration of the postconditioning phase was approximately 8 minutes again.

Participants returned for a *memory test* 22 hr to 26 hr after encoding on the first experimental day. They first completed a short questionnaire to assess whether they had already anticipated the following recognition test. To this end, after being informed about the following memory test, they rated how surprised they were about the upcoming memory test on a scale from 1 (*not surprised at all*) to 5 (*very surprised*). For later analyses, we inverted the scale of this measure so that larger values indicate less surprise as in Dunsmoor et al. (2015). Next, they received written instructions explaining details of the recognition test. In the recognition test, they were presented all 180 photographs from the three encoding phases of the previous day intermixed with an equal number of “new” photographs (i.e., photographs that had not been presented previously). Half of these lures were photographs of animals and half were photographs of tools. Stimuli were presented one by one centrally on a white background. For each of these photographs, participants first decided whether it was “old” or “new” in a forced-choice fashion. Then, participants had to indicate how confident they were that this decision was correct by pressing buttons corresponding to *very unsure* (German: *sehr unsicher*), *rather unsure* (*eher unsicher*), *rather sure* (*eher sicher*) and *very sure* (*sehr sicher*). If in any of the two stages no response was given within 5 s, the rest of the trial was skipped. Between trials, a black fixation cross on a white background was presented centrally for $1.5 \text{ s} \pm .5 \text{ s}$.

Data Analysis

Confirmatory statistical analyses were kept as close as possible to the analyses described in the original study by Dunsmoor et al. (2015). Specifically, these memory analyses were performed on corrected recognition scores to account for different response criteria between subjects. These were derived by subtracting the individual per image category false alarm rate from the per image category and per phase hit rate. Responses were collapsed across confidence, that is, only the forced-choice decision between “old” and “new” items was considered for memory performance. Besides *t*-tests on corrected recognition scores as reported by Dunsmoor et al. (2015), we also report *t*-tests on sensitivity scores (d') based on signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002). Before computing their *z* scores from the standard normal distribution, hit- and false-alarm-rates were restricted to the range of 1% to 99%. All *t*-tests were two-tailed.

Further, Bayes factors were calculated using the *ttestBF* R-function from the *BayesFactor* package to directly compare the adequacy of the null hypothesis H_0 that the true effect is equal to zero against the one-sided alternative hypothesis H_1 that the effect is greater than zero. We applied a Cauchy prior distribution with a default scale parameter of $r = .707$ (Morey et al., 2018; Rouder et al., 2009). The resulting BF_{10} metric indicates relative evidence for the H_1 versus the H_0 such that values greater than 1 favor the

alternative hypothesis H_1 and values smaller than 1 favor the null hypothesis H_0 . We interpret values greater than 3 as substantial evidence for the H_1 , while values smaller than 1/3 are interpreted as substantial evidence for the H_0 (Jarosz & Wiley, 2014).

As a manipulation check for successful fear conditioning, we analyzed skin conductance data obtained during the second encoding phase using both (a) a continuous decomposition analysis (CDA) and (b) a more classic through-to-peak (TTP) analysis, which was more similar to the SCR analysis in Dunsmoor et al. (2015), using Ledalab Version 3.4.9 (Benedek & Kaernbach, 2010). First, the skin conductance signal was downsampled to a resolution of 50 Hz and optimized using four sets of initial values. The minimum amplitude threshold was set to $.01 \mu\text{S}$. For each trial during the conditioning phase, we derived anticipatory SCRs as the average phasic driver within a response window of .5 s to 4.5 s after each stimulus onset to obtain CDA-estimates. Like Dunsmoor et al. (2015), we also obtained more classic through-to-peak results, expressed as the sum of significant SCR-amplitudes within the specified response window. Importantly, as shocks always appeared exactly 4.5 s after stimulus onset and therefore outside the response window, the resulting estimates could not have been biased by the UCS.

Results and Discussion

Successful Fear Conditioning

An analysis of skin conductance responses confirmed that our procedure successfully induced conditioned fear for items from the CS^+ category. During the conditioning phase, participants showed significantly higher anticipatory SCRs to CS^+ items compared with CS^- items (TTP: $t[43] = 4.20$, $p < .001$, $d_{av} = .51$; CDA: $t[43] = 4.79$, $p < .001$, $d_{av} = .52$; Figure 2).

Anticipation of the Recognition Test

On the second experimental day, participants were first informed about the following recognition test for photographs from the previous day and then rated how surprised they were by this task on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). Responses from six participants were missing. The average response in the remaining sample was 3.08 ($SD = .97$), showing that, on average, participants were moderately surprised. Four participants indicated that they were *not surprised at all*. Exclusion of these four participants had no effect on the pattern of results. Therefore, these participants were still included in the following analysis.

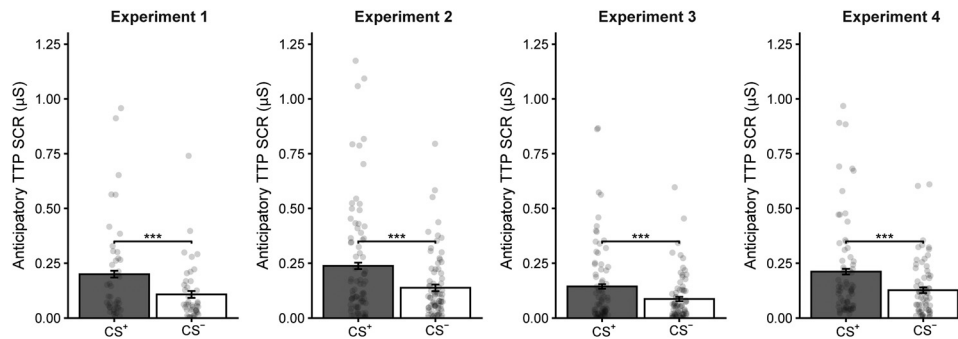
Overall Memory Performance

Overall, participants performed well in the recognition task (see Table 1), as reflected in a markedly higher average hit rate for items from all three encoding phases (i.e., the rate of correctly classifying previously seen photographs as old) of 69.6% ($SD = .11$) than the false alarm rate (i.e., the rate of incorrectly classifying previously unseen photographs as old) of 24.4% ($SD = .09$).

No Evidence for Category-Specific Retroactive Memory Enhancement

To address our main research question, we investigated how recognition performance for the photographs presented on the first experimental day was affected by the encoding phase (before, during or after the fear conditioning) and the conditioning category

Figure 2
Successful Fear Conditioning as Indicated by Average Anticipatory Skin Conductance Responses (SCRs) in Experiments 1–4 That Were Estimated Using a Through-to-Peak Analysis Similar to Dunsmoor et al. (2015)



Note. In all four experiments, participants showed significantly greater anticipatory skin-conductance responses during fear conditioning to CS⁺ items compared with CS⁻ items, confirming a successful fear induction. TTP = through-to-peak. Error bars represent ± 1 SEM.

*** $p < .001$.

(CS⁺ or CS⁻) an item belonged to through a repeated-measures ANOVA on corrected recognition scores. For the factor phase, Mauchly's test indicated that the sphericity assumption was violated, $W = .85$, $p = .030$. Hence, results for the factor phase are reported after applying a Greenhouse-Geisser correction. Overall, corrected recognition scores differed according to the phase an item was encoded in, $F(1.69, 72.73) = 17.1$, $p < .001$, $\eta^2_G = .05$. Whether an item belonged to the conditioned category, on the other hand, had no significant overall effect on recognition performance, although a trend was visible, $F(1, 43) = 3.92$, $p = .054$, $\eta^2_G = .01$. There was no significant interaction between the encoding phase and the conditioning category an item belonged to, $F(2, 86) = 1.65$, $p = .20$, $\eta^2_G = .004$. We further performed paired t -tests comparing the corrected recognition for items from the CS⁺ category versus items from the CS⁻ category separately per phase. These confirmed previous findings of an enhanced memory formation for CS⁺ items versus CS⁻ items in the conditioning phase, $t(43) = 2.31$, $p = .025$, $d_{av} = .35$ (Figure 3, upper left panel; Dunsmoor et al. 2015). At trend level, there was evidence that this memory benefit persisted for CS⁺ items over CS⁻ items in the postconditioning phase, even though these photographs were never directly paired with the UCS, $t(43) = 1.82$, $p = .076$, $d_{av} = .29$. Critically, for items that were encoded during the preconditioning phase, there was no evidence for a category-specific retroactive memory enhancement for CS⁺ items over CS⁻ items, $t(43) = .36$, $p = .72$, $d_{av} = .06$.

Finally, we also tested for preconditioning items the previously reported positive linear relationship between the temporal distance to the conditioning phase and the size of category-specific retroactive memory enhancements (Dunsmoor et al., 2015). To this end, CS⁺ and CS⁻ preconditioning items were each binned in tertiles corresponding to trials 0–10, 11–20, and 21–30 relative to the conditioning phase. A repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ as the dependent variable and the time bin as a within-subject factor showed no significant effect of time bins, $F(2, 86) = .20$, $p = .82$, $\eta^2_G = .002$. In contrast with previous reports (Dunsmoor et al., 2015), this finding indicates that the relative time of encoding of an item within the preconditioning phase had no effect on a putative category-specific retroactive memory enhancement.

Complementary Analyses

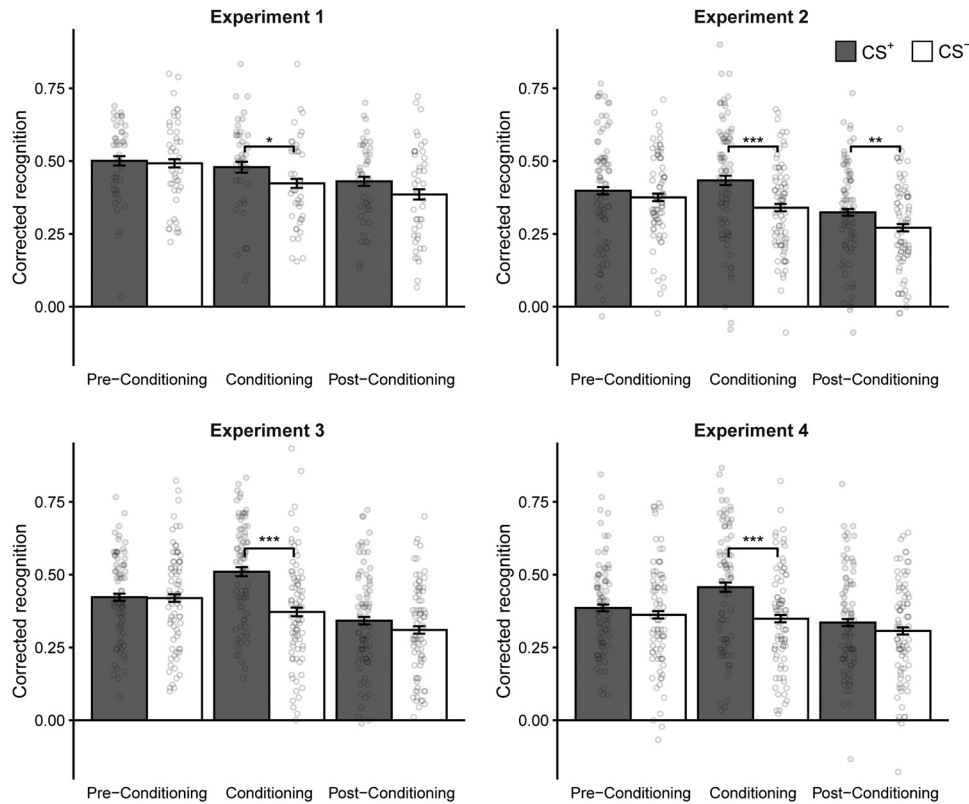
Although previous analyses showed no evidence for any category-specific retroactive memory enhancement, these relied on classic frequentist statistics and can therefore only indicate evidence *against*, but not *in support of* the null hypothesis. To this end, we reanalyzed previously reported classic paired t -tests with their Bayesian counterparts (see the Method section). For items encoded during the conditioning phase, these provided substantial support for the alternative hypothesis of a positive memory effect for CS⁺ compared with CS⁻ item from the same phase, $BF_{10} = 3.53$. Similarly, for items encoded after fear conditioning, results also favored the alternative hypothesis of a memory

Table 1
Retrieval Memory Results in Experiment 1, Mean Proportion of Responses by Certainty

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.603	0.148	0.129	0.120	0.571	0.162	0.126	0.141
Conditioning	0.580	0.148	0.134	0.138	0.511	0.154	0.154	0.181
Postconditioning	0.521	0.158	0.152	0.170	0.482	0.145	0.168	0.205
New	0.121	0.126	0.236	0.518	0.121	0.119	0.227	0.533

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

Figure 3
Recognition Performance Expressed as Hit Rate Minus False Alarm Rate in Experiments 1–4 by Encoding Phase and Conditioning Category



Note. In all four experiments, recognition was improved for items from the CS⁺ category that were encoded during Pavlovian fear conditioning. Only in Experiment 2 was this effect significantly carried over to items encoded after the end of the fear conditioning. Most importantly, none of the four experiments provided any evidence for category-specific retroactive memory enhancement. Error bars represent ± 1 SEM.

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

advantage for CS⁺ items relative to CS⁻ items, although evidence was only anecdotal, $BF_{10} = 1.41$. Most importantly, for preconditioning items, Bayesian analysis further provided substantial support for the null hypothesis rejecting any category-specific retroactive memory enhancement, $BF_{10} = .22$.

Dunsmoor et al. (2015) performed all memory analyses on corrected recognition scores, defined as hit rates minus false alarm rates. Here, we repeated their main analyses using sensitivity scores (d') based on signal detection theory, another common measure of recognition performance in the memory literature (Macmillan & Creelman, 2005; Wickens, 2002). These parallel analyses showed no significant differences between CS⁺ and CS⁻ items in any of the three encoding phases, all t s < 1.60 , all p s $> .12$. In the online supplemental materials, we further present results of parallel analyses using generalized linear mixed-effect models, showing an identical pattern of results as in the analysis based on memory sensitivity.

To identify possible factors hindering us from replicating the category-specific retroactive memory effect, we performed additional analyses beyond merely replicating the analysis strategy reported by Dunsmoor et al. (2015). Notably, participants were slightly less surprised by the recognition test than in the original study. However, there was no significant correlation between

memory test anticipation and recognition performance, Spearman's $r_s = .22$, $p = .18$. Further, overall memory performance per participant did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS⁺ minus mean SCRs to CS⁻), TTP: Spearman's $r_s = -.01$, $p = .93$, CDA: Spearman's $r_s = -.01$, $p = .95$. Only in very few trials of the recognition test, participants failed to respond quickly enough rejecting the notion that this might have biased our results. The mean number of missed trials per participant was .39 ($SD = .75$) of 360.

Experiment 2: Testing the Fear-Related Category-Specific Retroactive Memory Enhancement With Increased Statistical Power and the Original Stimulus Set

Despite successful fear conditioning and a replication of the procedure and analysis strategy from Dunsmoor et al. (2015), we found no evidence for category-specific retroactive memory enhancements in Experiment 1. It should be noted that, although conceptually very similar, Experiment 1 did not use the original stimulus set. Additionally, there were subtle differences in the procedure. For example, in our recognition test, "old" versus "new" decisions and memory confidence were tested separately for each

item, while this was a single step in the original study (Dunsmoor et al., 2015). Although there is no theoretical justification how these small deviations from the original study should prevent the detection of the proposed category-specific retroactive memory effect, we aimed to investigate whether we could replicate the findings when using the original stimulus set and sticking closer to the original procedure. Therefore, we contacted the lead author of the original study and asked him to provide us the stimulus materials, including all experimental instructions, which were used in the original study (Dunsmoor et al., 2015). We received these materials and used them for a direct replication of the original study by Dunsmoor et al. (2015). Additionally, we substantially increased the sample size compared with Experiment 1 to minimize the chance of null findings attributable to insufficient statistical power. Experiment 2 used the same variation in the CS-UCS timing as Experiment 1. Further, we did not control for stimulus typicality in this experiment as this aspect was not mentioned in Dunsmoor et al. (2015) and brought forward to us at a later stage.

Method

Participants

Eighty-four healthy participants (60 women) between 19 and 35 years of age took part in Experiment 2 ($M = 25.23$, $SD = 4.08$). Four participants had to be excluded from the analysis, either because they did not return for the memory test on the second experimental day or because of technical and experimenter errors. Because these exclusions, there was a slight imbalance regarding the between-subjects factor conditioned image category, such that for 41 participants tools served as CS⁺ category, whereas animals served as CS⁺ category for only 39 participants. We examined whether this slight imbalance affected our results by randomly excluding two participants with tools as the CS⁺ category from our analysis (10 permutations). Because the pattern of results remained unchanged, the following analyses were performed for the full sample of 80 participants. Again, the target sample size was determined using an a priori power analysis in G*Power 3 (Faul et al., 2007) with the aim to considerably increase the statistical power compared with Experiment 1 and thereby minimize the chance of null findings due to an insufficient sample size. As in Experiment 1, we assumed $d_z = .45$ as a point-estimate for the previously reported category-specific retroactive memory effect (Dunsmoor et al., 2015). A two-tailed paired t -test with $\alpha = .05$ required at least 82 participants to achieve a statistical power of .98. Exclusion criteria were identical to Experiment 1. None of the participants from Experiment 1 participated in this experiment. As before, participants received a monetary compensation of 20€. The study protocol was approved by the ethics committee of the Faculty of Psychology and Human Movement Science at the Universität Hamburg.

Materials

For this experiment, we used the stimulus set from the study by Dunsmoor et al. (2015), consisting of 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. As in Experiment 1, photographs were of neutral valence and each tool and animal represented a unique exemplar of its respective category. The stimulus set was randomly divided

into learning items and lures per participant and learning items were allocated to the three encoding phases in the same manner as in Experiment 1.

Procedure

The procedure in this experiment was largely identical to Experiment 1, except for some minor changes to achieve consistency with the original study (Dunsmoor et al., 2015). More precisely, we changed the location of the stimulation electrode from the right lower leg to the right wrist. Because this area tends to be more sensitive to electric stimulation, we also reduced the initial intensity in the procedure for determining the pain threshold from 20V in Experiment 1 to 10V in this experiment. Furthermore, we replaced the two-step forced-choice decision in the surprise recognition test on the second experimental day with a task assessing both “old” versus “new” decisions and certainty in a single step as reported in the original study (Dunsmoor et al., 2015). More precisely, for each stimulus in the recognition test, participants performed only a single button press with the four possible options that the currently presented item was either *definitely old* (German: *sicher alt*), *maybe old* (*eher alt*), *maybe new* (*eher neu*), or *definitely new* (*sicher neu*) by pressing the “1”, “2”, “3”, or “4” button on the keyboard, respectively. Additionally, the 5-s time limit per response that was used in Experiment 1 was removed. In sum, Experiment 2 used the same experimental procedure and stimuli as the study by Dunsmoor et al. (2015) but a significantly larger sample size.

Data Analysis

The statistical analysis was identical to Experiment 1 and the statistical analysis of Dunsmoor et al. (2015), complemented by analyses based on signal detection theory parameters and a Bayesian analysis.

Results and Discussion

Successful Fear Conditioning

An analysis of skin conductance data indicated that our procedure successfully induced conditioned fear for CS⁺ items. Specifically, during Pavlovian conditioning participants showed greater anticipatory SCRs to items from the CS⁺ category compared with items from the CS⁻ category (TTP: $t[79] = 4.75$, $p < .001$, $d_{av} = .48$; CDA: $t[79] = 4.32$, $p < .001$, $d_{av} = .35$; Figure 2).

Anticipation of the Recognition Test

As in Experiment 1, participants rated how surprised they were by the recognition test on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). On average, they indicated that they were moderately surprised ($M = 3.11$, $SD = 1.12$). Nine participants chose the *not surprised at all* option. Because excluding these participants did not affect the pattern of results, they were still included in the following analyses.

Overall Memory Performance

Participants performed overall very well in the surprise recognition test (see Table 2) with an average hit rate for items from all

Table 2*Retrieval Memory Results in Experiment 2, Mean Proportion of Responses by Certainty*

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.421	0.224	0.216	0.138	0.388	0.245	0.245	0.123
Conditioning	0.454	0.227	0.205	0.114	0.331	0.265	0.267	0.137
Postconditioning	0.352	0.219	0.258	0.171	0.298	0.230	0.307	0.165
New	0.078	0.169	0.368	0.386	0.083	0.174	0.381	0.362

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

three encoding phases of 60.9% ($SD = .14$) and an average false alarm rate of 25.2% ($SD = .10$).

No Evidence for Category-Specific Retroactive Memory Enhancement

We performed a repeated-measures ANOVA on corrected recognition scores to identify factors affecting memory formation over the task. As in Experiment 1, the recognition performance generally differed between phases, $F(2, 158) = 26.8, p < .001, \eta^2_G = .06$. The recognition performance was also generally different between CS⁺ and CS⁻ items, $F(1, 79) = 17.0, p < .001, \eta^2_G = .03$. Finally, there was a significant interaction between the encoding phase and the item category, indicating that the effect of membership of an item to the CS⁺ versus CS⁻ category differed between the encoding phases, $F(2, 158) = 6.66, p = .002, \eta^2_G = .007$. To further qualify these results, we performed paired t -tests comparing corrected recognition scores for items belonging to the CS⁺ versus CS⁻ category separately per phase. For photographs presented during Pavlovian conditioning, we obtained an enhanced memory for CS⁺ items compared with CS⁻ items, $t(79) = 4.89, p < .001, d_{av} = .53$ (Figure 3, upper right panel). This memory benefit for items belonging to the CS⁺ category carried over to the postconditioning phase, even though shock leads were removed beforehand, as indicated by improved corrected recognition scores, $t(79) = 3.32, p = .001, d_{av} = .33$. Most importantly, despite the high statistical power in this replication study, corrected recognition scores provided no evidence for a retroactive memory enhancing effect for items from the CS⁺ category over items from the CS⁻ category presented before the Pavlovian conditioning phase, $t(79) = 1.28, p = .20, d_{av} = .14$. As in Experiment 1, we also tested for preconditioning items the previously reported linear relationship between their temporal distance and the size of the category-specific retroactive memory effect using the same procedure as in Experiment 1. Contrary to this hypothesis, a repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ items as the dependent variable and the time bin as a within-subject factor showed no significant effect of time bins, $F(2, 158) = 1.08, p = .34, \eta^2_G = .007$. This finding shows that the relative time of encoding of an item within the preconditioning phase had no effect on the proposed category-specific retroactive memory enhancement.

Complementary Analyses

As in Experiment 1, we performed additional Bayesian paired t -tests to quantify relative evidence for the null versus the alternative hypotheses regarding effects of fear conditioning on memory

formation in the different encoding phases. These confirmed previous findings by showing substantial evidence for the alternative hypothesis of an enhanced memory for CS⁺ versus CS⁻ items that were encoded during Pavlovian conditioning, $BF_{10} = 6223$. Similarly, a Bayesian analysis indicating substantial support for the hypothesis of a memory advantage for CS⁺ over CS⁻ items that were encoded after fear conditioning, $BF_{10} = 36.38$. Most critically, for the comparison of CS⁺ and CS⁻ items encoded before fear conditioning, results from the Bayesian analysis spoke against category-specific retroactive memory enhancement, although unlike in Experiment 1, evidence for the null hypothesis was only anecdotal, $BF_{10} = .48$.

Although Dunsmoor et al. (2015) based their critical analyses on corrected recognition scores, we also aimed to replicate their findings using memory sensitivity scores (d'). As expected, results were very similar to those based on corrected recognition scores. Specifically, we found improved memory for CS⁺ items encoded during fear conditioning compared with CS⁻ items from the same phase, $t(79) = 4.31, p < .001, d_{av} = .49$. This improved memory sensitivity for CS⁺ items also carried over to the postconditioning phase, $t(79) = 2.58, p = .012, d_{av} = .27$. As for corrected recognition scores, there was no evidence for category-specific retroactive memory enhancement in memory sensitivity scores (d'), $t(79) = 1.28, p = .20, d_{av} = .14$. Again, analyses based on generalized linear mixed-effect models showing the same pattern of results are included in the online supplemental materials.

To identify possible factors contributing to the lack of category-specific retroactive memory enhancements in this experiment, we again performed additional analysis beyond those reported by Dunsmoor et al. (2015). Although participants were slightly less surprised by the recognition test than in Dunsmoor et al. (2015), there again was no significant correlation between the anticipation of the memory test and recognition performance, Spearman's $r_s = .15, p = .19$. Further, individual memory performance did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS⁺ minus mean SCRs to CS⁻), TTP: Spearman's $r_s = -.08, p = .50$; CDA: Spearman's $r_s = -.01, p = .95$.

Experiment 3: Testing the Effect of an Increased Interval Between Preconditioning and Fear-Conditioning on Category-Specific Retroactive Memory Enhancement

Thus far, we were unable to find any evidence for category-specific retroactive memory enhancements for weakly encoded stimuli belonging to a category that was later associated with the occurrence of shocks in a fear conditioning paradigm. Experiment

2 showed the absence of the category-specific retroactive memory effect could not be attributed to the stimulus set, nor to small deviations in the procedure. Additionally, as Experiment 2 had high statistical power, it is highly unlikely that the absence of the category-specific retroactive memory effect was due to an insufficient sample size.

This third replication attempt was designed to investigate one aspect that moderated the size of the category-specific retroactive memory effect in the original study, namely the interval between the encoding during the preconditioning phase and the subsequent fear-conditioning (Dunsmoor et al., 2015). It had been shown that items from the preconditioning phase that were presented the longest before fear-conditioning showed the strongest category-specific retroactive memory enhancement, whereas this effect seemed to diminish the closer items were encoded relative to the conditioning phase (Dunsmoor et al., 2015). This finding was in line with previous work on nonspecific behavioral tagging in rodents, which suggests that there might be a minimal interval between the weak encoding (setting the tag) and associated arousing event to enable retroactive memory enhancement (de Carvalho Myskiw et al., 2013; Moncada et al., 2015). However, it is important to note that, in these experiments investigating unspecific behavioral tagging, the interval between initial weak encoding and the subsequent memory promoting event was relatively long, typically more than one hour. Dunsmoor et al. (2015), on the other hand, observed positive effects of the temporal distance of a preconditioning item to conditioning procedure at a much shorter time scale, that is, only minutes. Here, we built the encoding-conditioning interval on the finding from Dunsmoor et al. (2015) to maximize the chances of detecting a category-specific retroactive memory enhancement. Therefore, in this third experiment, we increased the interval between the preconditioning and the fear conditioning phase by 10 minutes to investigate whether this change could produce the category-specific retroactive memory enhancement that was not detectable in Experiments 1 and 2. Apart from this single aspect, we retained both the procedure and the high statistical power from Experiment 2. Therefore, the 200-ms variation in the CS-UCS timing compared with Dunsmoor et al. (2015) was also retained in this experiment. Again, we did not control for stimulus typicality in this experiment as this aspect was not mentioned in Dunsmoor et al. (2015) and brought forward to us only at a later stage.

Method

Participants

Eighty-four healthy volunteers (59 women) between 18 and 33 years of age participated in this experiment ($M = 25.11$, $SD = 3.57$). Six participants had to be excluded from the analysis, either because they did not return for the memory test on the second experimental day or because of technical and experimenter errors. As in Experiment 2, these exclusions led to a slight imbalance regarding the between-subjects factor conditioned image category, such that for 40 participants tools served as CS⁺ category, whereas animals served as CS⁺ category for only 38 participants. Again, we examined whether this imbalance affected our results by randomly excluding two participants with tools as the CS⁺ category from our analysis (10 permutations). Because the pattern of results

remained unchanged, the following analyses were performed for the full sample of 78 participants. The target sample size was calculated using an a priori power analysis with identical parameters as in Experiment 2. Exclusion criteria were identical as in Experiments 1 and 2. None of the participants had previously participated in Experiment 1 nor in Experiment 2. Again, participants received a monetary compensation of 20€. The study protocol was approved by the ethics committee of the Faculty of Human Movement Science at the Universität Hamburg.

Materials

We used the same stimulus set as in Experiment 2, corresponding to the material used by Dunsmoor et al. (2015) and consisting of 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. As in Experiment 1 and Experiment 2, per participant, half of the stimuli from each category were randomly selected as learning items, whereas the remaining half served as lures. The learning items were allocated to each of the three encoding phases in the same manner as in Experiment 1 and Experiment 2. Furthermore, the assignment of photographs of tools and animals as CS⁺ and CS⁻, respectively, was counterbalanced across participants.

Procedure

The only difference compared with the procedure in Experiment 2 was the extension of the interval between the preconditioning phase and the subsequent fear conditioning phase. This change was based on the finding that the category-specific retroactive memory effect was positively correlated with the temporal distance between the encoding of an item and the following fear conditioning procedure in the original study (Dunsmoor et al., 2015) as well as evidence from studies in rodents (de Carvalho Myskiw et al., 2013; Moncada et al., 2015). In this experiment, when participants finished the preconditioning phase—unlike in Experiment 1 and 2—we did not immediately attach the electrodes. Instead, participants were first presented the following series of questionnaires: The State-Trait Anxiety Inventory (Spielberger, 1983), a multidimensional mood questionnaire (Steyer et al., 1997), a chronic stress questionnaire (Schulz et al., 2004), the Beck Depression Inventory (Beck et al., 1996), the Social Interaction Anxiety Scale (Mattick & Clarke, 1998), and the Positive and Negative Affect Schedule (Watson et al., 1988). After participants had worked on these questionnaires for exactly 10 minutes, they were interrupted and told that the remaining questions could be finished at a later stage. In fact, questionnaires were only added to keep participants occupied during the prolonged interval before the fear-conditioning. For this reason, we also chose a greater number of questionnaires than could usually be completed within 10 minutes, so that no participant would finish them earlier. Afterward, the experiment continued in the same manner as described for Experiment 2, by first attaching electrodes and determining the pain threshold (taking an additional approximately 10 min), before the start of the fear-conditioning phase, followed by the postconditioning phase and the 24-hr-delayed recognition test.

Data Analysis

The statistical analysis was identical to Experiments 1 and 2.

Results and Discussion

Successful Fear Conditioning

As in both previous experiments, an analysis of skin conductance data confirmed that our procedure was successful in inducing conditioned fear for CS⁺ items. SCR data for one additional participant were missing due to experimenter error. For the remaining sample of 77 participants, during Pavlovian fear conditioning anticipatory SCRs to items from the CS⁺ category were significantly higher compared with items from the CS⁻ category (TTP: $t[76] = 3.97, p < .001, d_{av} = .39$; CDA: $t[76] = 4.10, p < .001, d_{av} = .30$; Figure 2).

Anticipation of the Recognition Test

As in the previous experiments, participants rated how surprised they were by the recognition test on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). Data from one participant were missing as a result of experimenter error. On average, the remaining 77 participants indicated moderate levels of surprise ($M = 2.95, SD = .97$). Three participants chose the *not surprised at all* option. Because excluding these participants did not affect the pattern of results, they were still included in the following analyses.

Overall Memory Performance

As in Experiments 1 and 2, participants performed overall very well in the surprise recognition test (see Table 3) with an average hit rate for items from all three encoding phases of 62.7% ($SD = .16$) and an average false alarm rate of 23.1% ($SD = .10$).

No Evidence for Category-Specific Retroactive Memory Enhancement

To analyze factors affecting memory performance for the different phases of the task, we ran a repeated-measures ANOVA on corrected recognition scores. As in the two previous experiments, corrected recognition scores generally differed between phases, $F(2, 154) = 35.72, p < .001, \eta^2_G = .08$. Corrected recognition scores were also generally different between items from the CS⁺ and CS⁻ categories, $F(1, 77) = 17.8, p < .001, \eta^2_G = .03$. Finally, this effect of item category membership differed between the encoding phases, as indicated by a significant interaction between the encoding phase and the item category, $F(2, 154) = 22.12, p < .001, \eta^2_G = .03$. To further qualify these results, we performed paired t -tests to compare corrected recognition scores for items belonging to the CS⁺ versus CS⁻ category separately per

encoding phase. As in Experiments 1 and 2, these showed an enhanced memory performance for CS⁺ items encoded during Pavlovian conditioning compared with CS⁻ items encoded in the same phase, $t(77) = 6.60, p < .001, d_{av} = .77$ (Figure 3, lower left panel). As in Experiment 1, there also was a trend toward improved recognition memory for CS⁺ items encoded after Pavlovian conditioning compared with CS⁻ items encoded in the same phase, although unlike in Experiment 2, this trend was not statistically significant, $t(77) = 1.90, p = .061, d_{av} = .19$. Above all, despite the increase in the interval between encoding and Pavlovian conditioning, we obtained no evidence for a retroactive enhancement of memory for CS⁺ items compared with CS⁻ items encoded before Pavlovian conditioning in corrected recognition scores, $t(77) = .17, p = .86, d_{av} = .02$. Notably, even at descriptive level, the memory difference between CS⁺ and CS⁻ items encoded before fear conditioning was negligible. We again tested the possibility of a previously suggested linear trend between preconditioning items' temporal distance to the conditioning phase and the size of retroactive memory enhancement. As in Experiment 1 and 2, a repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ items as the dependent variable and the time bin as a within-subject factor showed no significant effect of time bins, $F(2, 154) = .17, p = .85, \eta^2_G = .001$. This indicates that the relative time of encoding of an item within the preconditioning phase had no effect on putative category-specific retroactive memory enhancement.

Complementary Analyses

As for both previous experiments, to quantify relative evidence for the null versus the alternative hypothesis of memory enhancements through fear learning in each of the three encoding phases, we conducted complementary Bayesian paired t -test. As before, these indicated substantial evidence for enhanced memory formation of CS⁺ relative to CS⁻ items that were encoding during fear conditioning, $BF_{10} = 4727037$. A corresponding Bayesian analysis for the postconditioning phase also favored the alternative hypothesis of enhanced memory for CS⁺ items, although evidence was only anecdotal, $BF_{10} = 1.34$. As in both previous experiments, the Bayesian analysis favored the null hypothesis rejecting the notion of category-specific retroactive memory enhancements and as in Experiment 1, evidence for the null hypothesis was substantial, $BF_{10} = .14$.

Whereas previous analyses focused on corrected recognition scores to closely replicate Dunsmoor et al. (2015), we also performed parallel analyses on memory sensitivity (d'). These yielded

Table 3

Retrieval Memory Results in Experiment 3, Mean Proportion of Responses by Certainty

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.407	0.242	0.225	0.126	0.420	0.235	0.221	0.125
Conditioning	0.481	0.256	0.166	0.097	0.352	0.255	0.261	0.132
Postconditioning	0.341	0.227	0.256	0.175	0.310	0.235	0.272	0.183
New	0.070	0.156	0.362	0.412	0.066	0.169	0.372	0.393

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

the same pattern of results as analyses based on corrected recognition. Specifically, the analysis based on d' confirmed previous findings of enhanced memory for CS⁺ items encoded during fear conditioning, $t(77) = 5.95$, $p < .001$, $d_{av} = .69$. For items encoded after conditioning, a similar, but nonsignificant trend was obtained, $t(77) = 1.72$, $p = .089$, $d_{av} = .20$. Above all, memory sensitivity scores (d') indicated no evidence for any category-specific retroactive memory enhancement for CS⁺ items from the preconditioning phase, $t(77) = .45$, $p = .65$, $d_{av} = .05$. Again, analyses based on generalized linear mixed-effect models showing the same pattern of results are presented in the online supplemental materials.

As in all previous experiments, there was no significant correlation between the anticipation of the memory test and recognition performance, Spearman's $r_s = -.04$, $p = .74$. Again, individual memory performance did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS⁺ minus mean SCRs to CS⁻), TTP: Spearman's $r_s = -.04$, $p = .76$; CDA: Spearman's $r_s = .11$, $p = .36$.

Experiment 4: Replicating Category-Specific Retroactive Memory Enhancements After Adopting Original UCS Timings and Balanced Stimulus Typicality Across Phases

The three previous experiments aimed to replicate findings of category-specific retroactive memory enhancements for stimuli from a category that was later associated with shock occurrences in a fear conditioning procedure. Compared with Experiment 1, Experiments 2 and 3 adopted additional details from the original procedure, namely the original stimulus set and the same format for the recognition tests. Based on comments from authors of the original study and reviewers, two additional deviations from Dunsmoor et al. (2015) were identified that applied to Experiments 1 to 3. First, in the original study, shocks coterminated with the stimulus presentation during fear-conditioning, whereas in Experiments 1–3 shock onsets were administered exactly at the point of stimulus offsets. Although this only leads to a 200-ms relative difference between studies (i.e., one shock length), it implies that stimuli were still present when shocks occurred in Dunsmoor et al. (2015), whereas in our Experiments 1 to 3 shocks followed immediately after stimulus offset. We address this issue here by using exactly the same shock timings that were used in Dunsmoor et al. (2015).

Second, Dunsmoor et al. (2015) controlled typicality and superordinate categories of stimuli, such that these were balanced across each of the three encoding phases and the recognition test. Unfortunately, they did not report on this in their study and we only learned about this aspect through the peer review process for this article. This contrasts with our procedure in Experiments 1–3, in which the set of stimuli was randomly distributed to each encoding phase. Therefore, our allocation of stimuli was unique per participant. We aimed to investigate whether this procedural difference might explain the lack of category-specific retroactive memory enhancements in our previous experiments. This fourth replication attempt had been preregistered and prereviewed before the beginning of data collection. The preregistration can be found at <https://osf.io/9hzmk>.

Method

Participants

Eighty-four healthy men and women between 18 and 34 years of age participated in this experiment ($M = 25.17$, $SD = 4.26$). Data from 13 participants had to be excluded because of an error in an early version of the experimental software that would in some trials incorrectly administer shocks to CS⁻ items. Because these exclusions might negatively affect the statistical power, we decided to recruit replacements for these 13 participants. One additional participant had to be excluded due to technical problems on the first experimental day. Therefore, the final sample included in the memory analysis consisted of 83 participants.

As in Experiment 2, there was a slight imbalance regarding the between-subjects factor conditioned image category, such that for 42 participants tools served as CS⁺ category, whereas animals served as CS⁺ category for 41 participants. Again, we examined whether this imbalance affected our results by randomly excluding two participants with tools as the CS⁺ category from our analysis (10 permutations). Because the pattern of results remained unchanged, the following analyses were performed for the full sample of 83 participants. The target sample size was calculated using an a priori power analysis with identical parameters as in Experiment 2 and 3. Exclusion criteria were identical as in Experiment 1, 2, and 3. None of the participants had previously participated in any of the other Experiments. Participants received a monetary compensation of 30€. The study protocol was approved by the ethics committee of the Faculty of Psychology and Human Movement Science at the Universität Hamburg.

Materials

We used the same stimulus set as in Experiments 2 and 3, corresponding to the material used by Dunsmoor et al. (2015) and consisting of 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. Unlike in Experiments 1–3, stimuli were not randomly allocated as learning items or distractors. Instead, we received the fixed stimulus allocation table that was used in Dunsmoor et al. (2015; Joseph E. Dunsmoor, personal communication, August 13, 2018) which was intended to match each of the encoding phases in terms of stimulus typicality and superordinate categories. In an online pilot-study, we recruited an additional independent sample of 41 participants (31 women, 10 men; aged 19–42 years; $M = 26.55$, $SD = 6.08$) who rated the typicality of all 360 stimuli. In random succession, they saw all 360 photographs (180 animals and 180 tools) and rated how typical each photograph was for its respective category on a scale from 1 (*very untypical*) to 10 (*very typical*). Ratings were self-paced (i.e., there was no time limit per photograph).

Results showed that simply adopting the allocation table from Dunsmoor et al. (2015) would lead to significant differences in typicality between encoding phases. After swapping four photographs of tools and two photographs of animals between sets, we obtained even typicality per category across sets (Figure 7 in the online supplemental materials). This procedure ensured that we had comparable typicality ratings per category across sets on the one hand, while sticking as closely as possible to the stimulus allocation used in Dunsmoor et al. (2015). The resulting stimulus sets consisted of three encoding sets with 30

photographs of animals and 30 photographs of tools each and a fourth set consisting of 90 photographs of animals and 90 photographs of tools that were used as lures in the recognition test. For each participant, the allocation of encoding sets to encoding phases was randomized. Further, as in all previous experiments, the assignment of photographs of tools and animals as CS⁺ and CS⁻, respectively, was counterbalanced across participants.

Procedure

The procedure in this experiment was identical to Experiment 2, except that we changed the timing of the shock (i.e., the UCS) during fear conditioning to be identical to Dunsmoor et al. (2015). During fear conditioning, a 200-ms shock occurred (under the same contingencies as in the previous two experiments), presented 4.3 s after stimulus onset and thus coterminated with the stimulus.

Data Analysis

The statistical analysis was identical to Experiment 1 and 3, and the statistical analysis of Dunsmoor et al. (2015), complemented by additional exploratory and a Bayesian analysis.

Results and Discussion

Successful Fear Conditioning

We again performed an analysis of skin conductance data to confirm the success of our fear-conditioning procedure. SCR data for twelve participants were not usable because of equipment misconfiguration. For the remaining sample of 71 participants, the TTP analysis (i.e., a more traditional approach of analyzing SCR data also utilized by Dunsmoor et al., 2015) indicated successful Pavlovian fear conditioning as expressed in increased anticipatory SCRs to CS⁺ items compared with CS⁻ items, $t(70) = 4.59, p < .001, d_{av} = .48$; for the CDA there was no significant effect, $t(70) = 1.24, p = .22, d_{av} = .08$ (see Figure 2).

Anticipation of the Recognition test

Again, participants rated how surprised they were by the recognition test on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). On average, they indicated that they were moderately and slightly more surprised than in Experiments 1–3 ($M = 2.82, SD = 1.14$). Eight participants reported being *not surprised at all*. Because excluding these participants did not affect the pattern of results, they were still included in the following analyses.

Overall Memory Performance

As in all three previous experiments, participants performed overall well in the recognition test (see Table 4). The average hit rate for items from all three encoding phases was 62.8% ($SD = .14$), with an average false alarm rate of 26.2% ($SD = .10$).

No Evidence for Category-Specific Retroactive Memory Enhancement

As in all three previous experiments, we ran a repeated-measures ANOVA on corrected recognition scores to analyze factors affecting memory performance for the different phases of the task. For the factor phase, Mauchly's test indicated that the sphericity assumption was violated, $W = .89, p = .008$. Hence, results for the factor phase are reported after applying a Greenhouse-Geisser correction. Corrected recognition scores generally differed between phases, $F(1.80, 147.30) = 18.19, p < .001, \eta^2_G = .036$. They were also generally different between items from the CS⁺ and CS⁻ categories, $F(1, 82) = 17.61, p < .001, \eta^2_G = .022$. Finally, this effect of item category membership differed between the encoding phases, as indicated by a significant interaction between the encoding phase and the item category, $F(2, 164) = 10.19, p < .001, \eta^2_G = .012$. These results were further qualified by paired *t*-tests comparing corrected recognition scores for items belonging to the CS⁺ versus CS⁻ category separately per encoding phase. As in all three previous experiments, these showed an enhanced memory performance for CS⁺ items encoded during Pavlovian conditioning compared with CS⁻ items encoded in the same phase, $t(82) = 5.75, p < .001, d_{av} = .58$ (Figure 3, lower right panel). As in Experiments 1 and 3, there also was a (nonsignificant) trend toward improved recognition memory for CS⁺ items encoded after Pavlovian conditioning compared with CS⁻ items encoded in the same phase, $t(82) = 1.71, p = .091, d_{av} = .14$. Most importantly, even after additionally adopting the exact UCS-CS timings from Dunsmoor et al. (2015) and controlling for stimulus typicality across phases, we obtained no evidence for a category-specific retroactive enhancement of memory for CS⁺ items compared with CS⁻ items encoded before Pavlovian conditioning in corrected recognition scores, $t(82) = 1.37, p = .18, d_{av} = .14$. We again tested the possibility of a previously suggested linear trend between pre-conditioning items' temporal distance to the conditioning phase and the size of retroactive memory enhancement. As in all three previous experiments, a repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ items as the dependent variable and the time bin as a within-subject

Table 4

Retrieval Memory Results in Experiment 4, Mean Proportion of Responses by Certainty

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.412	0.243	0.226	0.120	0.375	0.243	0.243	0.139
Conditioning	0.497	0.229	0.186	0.088	0.348	0.257	0.251	0.144
Postconditioning	0.364	0.240	0.250	0.146	0.332	0.231	0.275	0.162
New	0.090	0.179	0.393	0.339	0.084	0.171	0.377	0.367

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

factor showed no significant effect of time bins, $F(2, 164) = .68, p = .51, \eta^2_G = .004$. Thus, we could not find any effect of the relative time of encoding of an item within the preconditioning phase on the size of the putative category-specific retroactive memory enhancement.

Complementary Analyses

We conducted complementary Bayesian paired t -tests to quantify relative evidence for the null versus the alternative hypothesis of memory enhancements through fear learning in each of the three encoding phases. As before, these indicated substantial evidence for enhanced memory formation of CS^+ relative to CS^- items that were encoded during fear conditioning, $BF_{10} = 176943$. A corresponding Bayesian analysis for the postconditioning phase slightly favored the null hypothesis of no memory advantage for CS^+ items, although evidence was only anecdotal, $BF_{10} = .93$. As in the three previous experiments, the Bayesian analysis favored the null hypothesis rejecting the notion of category-specific retroactive memory enhancements, although evidence for the null hypothesis was only anecdotal, $BF_{10} = .54$.

In addition to analyses focusing on corrected recognition scores to closely replicate Dunsmoor et al. (2015), we also performed parallel analysis on memory sensitivity (d'). These yielded the same pattern of results as analyses based on corrected recognition. Specifically, analysis based on d' confirmed previous findings of enhanced memory for CS^+ items encoded during fear conditioning, $t(82) = 5.37, p < .001, d_{av} = .57$. For items encoded after conditioning, there were no significant differences in d' between CS^+ and CS^- items, $t(82) = 1.36, p = .18, d_{av} = .14$. Most importantly, memory sensitivity scores (d') indicated no evidence for any category-specific retroactive memory enhancement for CS^+ items from the preconditioning phase, $t(82) = 1.12, p = .27, d_{av} = .11$. Analyses based on generalized linear mixed-effect models again showed the same pattern of results and are included in the online supplemental materials.

As in all three previous experiments, we found no significant correlation between the anticipation of the memory test and recognition performance, Spearman's $r_s = .13, p = .25$. Again, individual memory performance did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS^+ minus mean SCRs to CS^-), TTP: Spearman's $r_s = .03, p = .78$; CDA: Spearman's $r_s = .01, p = .91$.

Analyses Focusing on High Confidence Hits

Although Dunsmoor et al. (2015) collapsed responses from the surprise recognition test across confidence, we also reanalyzed our data by focusing only high confidence hits using the same paired t -tests on corrected recognition scores as reported in the article, complemented by their Bayesian counterparts to quantify the relative evidence for the null versus the alternative hypothesis of category-specific retroactive memory enhancement. For this analysis, we used a definition of high confidence hits that treated any *rather old* responses like *new* responses (Dunsmoor et al., 2012; Keller & Dunsmoor, 2020). Therefore, only *definitely old* responses could result in either a hit or a false alarm, whereas *rather old* responses were always scored as either misses or correct rejections depending on the actual status of the item. Note that focusing on high confidence hits therefore implies a different scoring of existing

responses, while no trials were omitted from the recognition analysis.

For Experiment 1, these analyses focusing on high confidence hits showed no evidence for category-specific retroactive memory enhancement on corrected recognition scores, $t(43) = 1.05, p = .30, d_{av} = .18, BF_{10} = .46$, nor on memory sensitivity (d'), $t(43) = .58, p = .56, d_{av} = .11, BF_{10} = .27$. In contrast to previously reported results after collapsing across memory confidence, for Experiment 2 an analysis focusing on high confidence hits showed the proposed category-specific retroactive enhancement on corrected recognition scores, $t(79) = 2.31, p = .024, d_{av} = .22, BF_{10} = 2.95$, but not on memory sensitivity (d'), $t(79) = 1.30, p = .20, d_{av} = .14, BF_{10} = .50$. For Experiment 3, results were again consistent with those obtained from the analysis of recognition collapsed over confidence and showed no evidence for category-specific retroactive memory enhancement in neither corrected recognition, $t(77) = .98, p = .33, d_{av} = .10, BF_{10} = .07$, nor in memory sensitivity (d'), $t(77) = .67, p = .50, d_{av} = .08, BF_{10} = .08$. Likewise, analyses on high confidence memory for Experiment 4 provided again neither evidence for the category-specific retroactive memory effect on corrected recognition, $t(82) = 1.72, p = .088, d_{av} = .19, BF_{10} = .95$, nor on memory sensitivity (d'), $t(82) = .20, p = .84, d_{av} = .08, BF_{10} = .14$.

Response Bias Analysis

A possible explanation for the inconsistent findings between corrected recognition and d' regarding high-confidence memory in Experiment 2 (and at trend level in Experiment 4) could be that findings appearing to show category-specific retroactive memory enhancement in corrected recognition for high confidence hits instead reflect a response bias toward more liberal *old* responses for items from the CS^+ category without any actual difference in memory sensitivity between items from the CS^+ versus CS^- category that were encoded during preconditioning (Dougal & Rotello, 2007; Rotello et al., 2008). We investigated this possibility by calculating response bias scores c based on signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002) and comparing them for items from the CS^+ versus CS^- category separately for each experiment and encoding phase. Detailed results from this analysis are provided in the online supplemental materials (Tables 1–4). In short, we found that participants overall showed a bias to classify items from the CS^+ category (over item from the CS^- category) as *old* when these were encoded during fear-conditioning. For the critical influence of response biases on findings of category-specific retroactive memory enhancement, Experiments 2 and 4 were the most interesting, because these were the only two experiments in which an analysis of high confidence corrected recognition provided some evidence for this effect (although only at trend level in Experiment 4). In Experiment 2, participants descriptively, but nonsignificantly, showed a slightly increased response bias in the high-confidence hit rate toward items from the CS^+ category that were encoded before fear-conditioning, $t(79) = 1.25, p = .21, d_{av} = .14$. For Experiment 4, this effect was significant, indicating that participants more liberally classified preconditioning items from the CS^+ category (compared with items from the CS^- category) as 'old', regardless of their actual status, $t(82) = 2.06, p = .042, d_{av} = .22$. For Experiments 1 and 3, there was no

significant difference in response bias for high confidence memory of items from the preconditioning phase (both $ps > .36$).

Pooled Analysis Across All Experiments

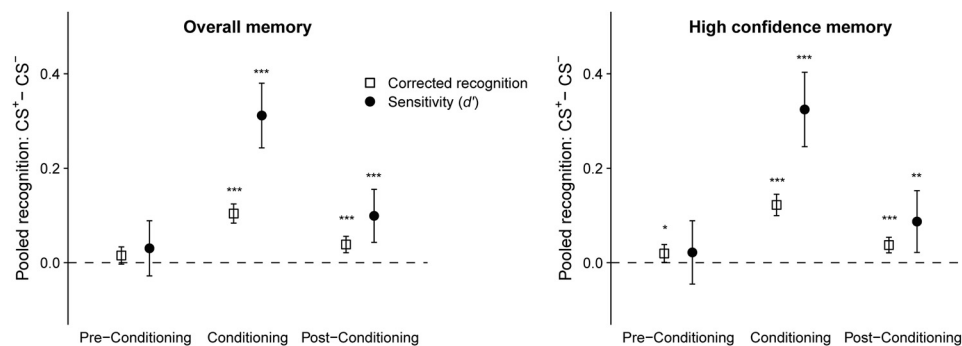
Experiments 1–4 were designed to replicate the previously reported finding of a category-specific retroactive memory enhancement through subsequent electric shocks (Dunsmoor et al., 2015). Although we varied certain aspects regarding stimuli and the procedure between these experiments, the procedure of Experiments 1–4 was conceptually very similar. To summarize findings from the four studies in a combined statistical model, we pooled data across experiments and fit separate linear mixed-effects models for both corrected recognition and memory sensitivity (d') using the *R* library *lme4* (Bates et al., 2015) for each of the three conditioning phases. In each of these three models, the conditioned category an item belonged to (binary coding: 0 for CS^- ; 1 for CS^+) was treated as a fixed effect (Hedges & Vevea, 1998) to explain corrected recognition scores. Additionally, random intercepts were fitted both per subject and per experiment, to account for differences in memory performance between participants and procedural differences between experiments, respectively. To further qualify the results reported in the previous section, we also ran separate models for memory collapsed over confidence levels and high confidence memory only, respectively.

Even after pooling data collapsed across confidence (parallel to Dunsmoor et al., 2015) from Experiments 1–4 with a total of 285 unique participants, there was no evidence for category-specific retroactive memory enhancement, as indicated by neither a significant effect of conditioning category membership of preconditioning items on corrected recognition scores, $\beta = .015$, 95% CI $[-.003, .033]$, $t(284) = 1.66$, $p = .097$, $BF_{10} = .17$, nor a significant effect of

conditioning category membership of preconditioning items on d' , $\beta = .03$, 95% CI $[-.028, .090]$, $t(284) = 1.03$, $p = .30$, $BF_{10} = .12$ (see Figure 4). Bayes factors obtained in both cases indicated substantial evidence for the null hypothesis speaking against category-specific retroactive memory enhancement. Although there was no evidence for a selective retroactive memory effect, the pooled analysis revealed that CS^+ photographs encoded during fear-conditioning were significantly better recognized than CS^- photographs, as reflected in both corrected recognition, $\beta = .10$, 95% CI $[.084, .124]$, $t(284) = 10.07$, $p < .001$, $BF_{10} = 544211770$, and in d' , $\beta = .31$, 95% CI $[.24, .38]$, $t(284) = 8.93$, $p < .001$, $BF_{10} = 13621099$. Finally, the pooled analysis collapsed over confidence confirmed results from Experiment 2 that this memory enhancement for CS^+ photographs relative to CS^- photographs carried over to the postconditioning phase, even though these items were never directly paired with the UCS. This was reflected in both corrected recognition, $\beta = .039$, 95% CI $[.021, .056]$, $t(284) = 4.37$, $p < .001$, $BF_{10} = 3.75$, and in d' , $\beta = .099$, 95% CI $[.043, .155]$, $t(284) = 3.47$, $p < .001$, $BF_{10} = 1.30$.

Next, we fitted the same pooled models for the exploratory analyses on high confidence recognition memory. This pooled analysis showed a significant category-specific retroactive memory enhancement in corrected recognition scores, although a Bayesian analysis of this model indicated substantial evidence for the null hypothesis, thus speaking against a category-specific retroactive memory enhancement, $\beta = .20$, 95% CI $[.0008, .039]$, $t(284) = 2.05$, $p = .042$, $BF_{10} = .22$. The same analysis on d' indicated no significant category-specific retroactive memory enhancement and substantial evidence for the null hypothesis, $\beta = .02$, 95% CI $[-.045, .089]$, $t(284) = .64$, $p = .52$, $BF_{10} = .10$. As for the analysis collapsed across confidence levels, the analysis focusing on high confidence recognition memory revealed clear evidence that CS^+ photographs encoded during fear-conditioning were significantly better recognized than CS^- photographs in both corrected

Figure 4
Results of a Pooled Analysis Across Data From Experiments 1–4 Using Linear-Mixed Effect Models



Note. The left panel shows the advantage in CS^+ over CS^- recognition performance after collapsing across memory confidence, whereas the right panel shows parallel results for high confidence memory only. In both analyses, items from the CS^+ category encoded during both Pavlovian fear conditioning and after fear conditioning were recognized significantly better compared with items from the CS^- category encoded within their respective phase, as reflected in both corrected recognition scores and d' from signal detection theory. For items encoded before the conditioning phase, there was a retroactive enhancement for items from the CS^+ vs. CS^- category when only high confidence memory was analyzed, but not when overall memory performance was analyzed. Moreover, this effect for high confidence memory was only present for corrected recognition, but not for d' . Error bars represent 95% confidence intervals.

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

recognition scores, $\beta = .12$, 95% CI [.10, .14], $t(284) = 10.60$, $p < .001$, $BF_{10} = 1.26 \times 10^{11}$; as well as d' , $\beta = .32$, 95% CI [.25, .40], $t(284) = 8.09$, $p < .001$, $BF_{10} = 2938815$. Finally, the pooled analysis for high confidence recognition memory confirmed that this memory advantage also carried over to the postconditioning phase, as reflected in significantly increased CS^+ over CS^- scores in both corrected recognition, $\beta = .037$, 95% CI [.021, .054], $t(284) = 4.47$, $p < .001$, $BF_{10} = 3.62$, as well as in d' , $\beta = .087$, 95% CI [.022, .152], $t(284) = 2.62$, $p = .009$, $BF_{10} = .49$.

General Discussion

Adaptive episodic memory has been theorized to preferentially store motivationally significant experiences that can be useful to guide future behavior (Nairne et al., 2007; Nairne & Pandeirada, 2008; Shohamy & Adcock, 2010). How can such an adaptive prioritization be achieved for stimuli that appear neutral during encoding, but are subsequently revealed to relate to important consequences? Recently, a possible mechanism has been suggested that retroactively and selectively promotes memory for initially neutral items when their respective category is later predictive of either aversive or appetitive events (Dunsmoor et al., 2015; Patil et al., 2017). These findings have challenged existing models of episodic memory formation by demonstrating for the first time that postencoding processes can selectively enhance memory for a specific group of stimuli, but not others, based on the (categorical) relatedness of the stimuli to the emotional event. In this framework, memories can exist in a weak, transient form ('tagged') that relies on a subsequent event ('capture') to store them permanently. This tag and capture framework had previously been developed at the level of individual neurons and was referred to as synaptic tagging (Frey & Morris, 1997, 1998). Studies in rodents and more recently in humans have successfully translated this framework to the behavioral level (Ballarini et al., 2009, 2013; de Carvalho Myskiw et al., 2013; Moncada et al., 2015). Importantly, these studies investigated general, nonselective retroactive effects on memory, irrespective of a semantical link between tagged stimuli and the subsequent memory-promoting event. Only recently, it has been reported that retroactive enhancements may selectively promote memory for one category of stimuli that has been linked to a subsequent arousing event, while leaving irrelevant stimuli unaffected (Dunsmoor et al., 2015; Patil et al., 2017). In addition to this category-specific retroactive (backward) effect, selective category-specific memory enhancements were also observed when appetitive or aversive stimuli were present during encoding (online) and for items from the relevant category that were encoded after these salient stimuli were present (i.e., a forward effect). Together, these findings of highly selective backward and forward memory effects are in contrast to more traditional models of memory formation, which have focused on effects that are driven through the allocation of attention during online encoding (Mulligan, 1998; Uncapher & Rugg, 2005) and general offline effects of physiological arousal that enhance consolidation in a nonselective fashion (McGaugh, 2015), irrespective of the semantic or conceptual relatedness of stimuli. In particular, the finding of a category-specific retroactive memory enhancement is incompatible with previous attentional models, as unlike online and forward enhancements, this backward effect cannot be explained by increased attention to stimuli from the category that had been

linked with salient outcomes, since this associative link was only established after the encoding of these items. Therefore, this highly selective retroactive memory enhancement is at the heart of this new framework.

In a series of four experiments, we aimed to replicate findings of the first published study showing category-specific retroactive memory enhancement for initially neutral stimuli through a following Pavlovian fear-conditioning procedure that linked aversive electric shocks to only one category of stimuli (Dunsmoor et al., 2015). Based on recent reports (Dunsmoor et al., 2015; Patil et al., 2017), we expected that memory for the initially neutral items would retroactively be enhanced when these are later revealed to belong to a relevant category. In sharp contrast to our hypotheses, analyses of overall recognition memory performance (as in Dunsmoor et al., 2015) failed to produce any evidence for a category-specific retroactive memory enhancement through aversive learning in all four experiments. Parallel Bayesian analyses provided substantial evidence for the null hypothesis speaking against a category-specific retroactive memory effect in Experiments 1 and 3 and anecdotal evidence for the null hypothesis in Experiments 2 and 4.

In a pooled analysis across all four experiments, we observed a similar pattern of results: When recognition memory was collapsed over confidence, evidence for category-specific retroactive memory enhancement was found neither in corrected recognition scores, nor in memory sensitivity (d'). In both cases, Bayes factors indicated substantial evidence for the null hypothesis. Only when additional analyses focused on high confidence memory and corrected recognition was there some evidence for the predicted category-specific retroactive memory effect, which was however only significant in one out of four experiments and was not paralleled by a significant improvement in memory sensitivity (d'), nor was it supported by a Bayesian analysis.

How can the inconsistencies between the previous reports of category-specific retroactive memory enhancements and the current findings be explained? Although close replications can be challenging (Stroebe & Strack, 2014) and seemingly small deviations from the original procedure can dramatically affect the replicability of a finding (Noah et al., 2018; Wagenmakers et al., 2016), Experiments 1, 2, and 4 were designed to match the procedure of the previous studies regarding various aspects such as timing, instructions, and stimuli, while substantially increasing the sample size. We focused only on the group of participants in which there was a 24-hr interval between encoding and recognition test, as these participants had shown the most robust evidence for category-specific retroactive memory enhancement (Dunsmoor et al., 2015). Other groups featured in the original study, such as an immediate retrieval or a strong encoding 24-hr retrieval group had not shown evidence for category-specific retroactive memory enhancement. Importantly, these group differences only become meaningful once the existence of the phenomenon is demonstrated in the first place. Achieved statistical powers were generally greater than 95% (except for Experiment 1, which used a sample size comparable to previous reports suggesting a selective behavioral tagging effect). Thus, a lack of statistical power is very unlikely.

Two further aspects that could have potentially affected the replicability of category-specific retroactive memory enhancements in Experiments 1–3 were (a) deviations in the relative timing of the

CS to the UCS compared with Dunsmoor et al. (2015) and (b) the random allocation of stimuli to each learning phase instead of controlling for typicality across their respective categories. In Dunsmoor et al. (2015), each 200-ms shock coterminated with the end of the stimulus presentation, whereas in our Experiments 1–3, the 200-ms shocks started with the end of the stimulus presentation. This resulting a 200-ms deviation in CS-UCS timing compared with Dunsmoor et al. (2015) was unintentional. Potentially, this issue could be relevant as the differential timing between CS and UCS can be used to differentiate between trace and delay conditioning, which involve different processes (Kochli et al., 2015; McLaughlin et al., 2002; Weike et al., 2007). However, trace conditioning would only be present if an additional pause were implemented between stimulus offset and the following UCS. Because this was not the case in Experiments 1–3, our procedure may still be considered a delay conditioning procedure like the one used in Dunsmoor et al. (2015). Moreover, we obtained significantly higher anticipatory SCRs for CS⁺ compared with CS⁻ items, indicating that our fear conditioning manipulation was successful. Furthermore, we could replicate the memory benefit for CS⁺ items that were presented during fear conditioning and partly the prospective memory effect for items that were presented after fear-conditioning, indicating that, despite the deviation in CS-UCS timing, the UCS was still able to modulate memory in Experiments 1 to 3. To our knowledge, there is no theoretical justification why only the retroactive effect, but not the online, nor the prospective effect should be affected by this difference in timing. Finally, after explicitly addressing the issue of CS-UCS timing in Experiment 4, we obtained a similar pattern of results as in Experiments 1–3 that most prominently did not show any signs of category-specific retroactive memory enhancement.

Regarding stimulus typicality, we unfortunately only learned during the peer review process that Dunsmoor et al. (2015) kept stimulus typicality constant across learning phases as this aspect was not mentioned at all in their original article. Even after balancing stimulus typicality across encoding phases in Experiment 4, we still found no evidence for any category-specific retroactive memory enhancement.

Another factor that has been suggested to moderate the extent of category-specific memory enhancement is the interval between the encoding of initially neutral stimuli and the following significant (either aversive or appetitive) event. Specifically, Dunsmoor et al. (2015) reported a linear trend between the distance of learning items to the following significant event and the strength of category-specific retroactive memory enhancement. This linear trend is in line with previous work on nonspecific behavioral tagging in animals suggesting that a minimal interval between initial learning and the following event is necessary for such effects to unfold (de Carvalho Myskiw et al., 2013; Moncada et al., 2015). We specifically addressed this issue in Experiment 3 by extending the interval between preconditioning and subsequent Pavlovian conditioning. It is important to note that, for practical reasons, this interval had to be at least approximately 10 min even in Experiments 1, 2, and 4. This time was needed to attach electrodes and adjust shock intensities and should correspond with Dunsmoor et al. (2015). For Experiment 3, we effectively doubled this interval to 20 min, which did not lead to the expected increase of the putative category-specific retroactive memory effect. Furthermore, none of Experiments 1, 2, 3, or 4 provided any evidence for the previously reported linear trend between

the temporal distance of an item of the preconditioning phase to the conditioning procedure and the size of category-specific retroactive memory enhancement.

Retroactive memory effects have been further theorized to only strengthen initially weak memories, but to have no additional benefit for already strongly encoded stimuli (Dunsmoor et al., 2015; Moncada & Viola, 2007; Wang et al., 2010). Accordingly, it could be speculated that our sample of participants included better learners, which might have prevented category-specific retroactive memory enhancement due to strong initial encoding. However, recognition performance in the present experiments was, with the exception of Experiment 1, comparable with previous studies reporting selective retroactive memory enhancements (Dunsmoor et al., 2015; Patil et al., 2017). Because Experiment 1 featured both a slightly different set of stimuli (although from the same categories) as well as a different format for the recognition test, this might explain the slightly increased overall preconditioning performance in this experiment compared with Dunsmoor et al. (2015). Both aspects were addressed in Experiments 2, 3, and 4 such that these featured the same set of stimuli and the same recognition test procedure. In these three experiments, we obtained similar memory performances during preconditioning as in Dunsmoor et al. (2015): For instance, CS⁻ preconditioning items were correctly classified as definitely old in 42.6% of all cases for the 24-hr retrieval group in Dunsmoor et al. (2015) and the corresponding performance ranged from 37.5% to 42.0% in our Experiments 2–4 (Tables 2–4). This renders overly strong memories as explanation for the absence of a selective retroactive memory effects rather unlikely. Additionally, despite participants indicating that they were overall slightly less surprised by the recognition test compared with Dunsmoor et al. (2015) and it cannot be completely ruled out that such differences may have influenced our results, although this would clearly question the robustness of the suggested category-specific tagging effect, there is no clear theoretical rationale why such a subtle difference should abolish the tagging effects. None of our experiments revealed any correlation between levels of surprise and memory performance.

It might be argued that emotion has a higher impact on memory for items recognized with high confidence (Kim & Cabeza, 2009; Phelps & Sharot, 2008). We therefore ran additional analysis that focused on high confidence memory only. In one of the four experiments (Experiment 2), this exploratory recognition analysis based on corrected recognition scores and focusing on high confidence hits showed a significant category-specific retroactive memory effect, although a parallel Bayesian analysis indicated that evidence was nonsubstantial. For Experiment 4, there was a nonsignificant trend in the same direction ($p = .088$). Interestingly, this retroactive memory enhancement for high confidence hits in Experiment 2 and respective trend in Experiment 4 were only detectable in corrected recognition scores, but not in memory sensitivity (d') from signal detection theory. Further, in the remaining two experiments, there was no evidence for a category-specific retroactive effect for high confidence memory and a Bayesian analysis on high confidence corrected recognition contrarily favored the null hypothesis. A pooled analysis across all four experiments that focused on high confidence corrected recognition showed a small but significant category-specific retroactive memory effect. A parallel Bayesian analysis, however, showed even substantial evidence for the null hypothesis rejecting the notion of category-

specific retroactive memory enhancement. The same pooled analysis for memory sensitivity (d') was much clearer: Neither for recognition scores collapsed across confidence, nor for those focusing on high confidence hits was there any evidence for the category-specific retroactive memory effect, with a parallel Bayesian analysis indicating substantial evidence for the null hypothesis in both cases. Although it is to be acknowledged that the two experiments in which we obtained an effect or a similar trend for a retroactive effect in high confidence memory might be considered the closest replication attempts to Dunsmoor et al. (2015), even in these experiments the evidence was not robust across memory parameters.

In the face of the findings of a significant category-specific retroactive memory effect for high confidence corrected recognition scores in Experiment 2 and the pooled analysis, it must also be noted we run multiple analyses (overall memory analysis, high confidence memory analysis, linear mixed models) across multiple parameters (corrected recognition score and d') and multiple experiments. This wide array of tests comes with a significantly increased risk of false positives (i.e., an inflated alpha-error rate). Only in one of the four experiments, there was a significant result and only in corrected recognition scores, but not in d' from signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002). As we aimed for the maximum sensitivity regarding possible effects, we did not correct for the relatively high number of statistical tests. If any correction for multiple testing was performed, none of the effects or trends for high confidence memory would be even close to statistical significance. Therefore, additional caution against interpreting the findings on high confidence memory as a successful replication of category-specific retroactive memory enhancement is warranted.

The observed discrepancy in results between analyses based on d' versus corrected recognition scores is interesting because both methods of estimating discrimination performance rely on different models of recognition memory (Snodgrass & Corwin, 1988). Whereas d' is rooted in signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002) and assumes curvilinear receiver operating characteristics (ROCs), corrected recognition scores as calculated by Dunsmoor et al. (2015) stem from the two-high-threshold model of recognition (Bröder et al., 2013; Snodgrass & Corwin, 1988) and assume linear ROCs. The issue of selecting the correct model is particularly important as we also found that participants showed for items from the CS^- category a more conservative response bias c (from signal detection theory) than for items from the CS^+ category for high confidence responses at least in Experiment 4. Ideally, this response bias should not influence memory discrimination scores, as it does not reflect true memory but rather a response tendency. Indeed, when that the assumptions of signal detection theory are correct, memory sensitivity (d') and response bias (c) are theoretically independent from each other (Snodgrass & Corwin, 1988). Likewise, if the model underlying corrected recognition scores is correct (i.e., the two-high-threshold model), these scores should equally be independent of the response bias. Although there has been some debate regarding the question which of these two approaches is generally more appropriate in the memory context, most empirical findings favor the use of signal detection theory (and therefore d') over the two-high-threshold model (associated with corrected recognition) when analyzing recognition performance (Dube & Rotello, 2012; Pazzaglia et al., 2013; Slotnick & Dodson, 2005). Future research on

the category-specific retroactive memory effect should optimally report results from both measures, consider theoretical implications if such an effect was detectable in only one measure but not the other, and consider possible response biases.

It should be noted that, although none of our four experiments provided consistent evidence for the existence of category-specific retroactive ('backward') memory enhancement, there was some evidence for the selective online and forward memory enhancements. In line with category-specific online effects, in all four experiments we consistently found a memory advantage for items from the CS^+ category that were presented during Pavlovian fear conditioning compared with items from the CS^- category encoded in the same learning phase. This finding corroborates previous studies showing enhanced memory for stimuli linked to arousing events (Dunsmoor et al., 2015; Dunsmoor & Kroes, 2019; Salehi et al., 2010; Vogel & Schwabe, 2016). In the context of adaptive memory, such a mechanism enables the preferential storage of stimuli that are associated with threat which may facilitate coping to similar situations in the future (Nairne et al., 2007; Nairne & Pandeirada, 2008). It is important to note that this memory enhancement for CS^+ items in Experiments 1 to 4 was evaluated by comparing them with CS^- items from the same category. Therefore, an alternative interpretation of these findings could be that CS^+ items encoded during fear conditioning did not experience a memory promotion per se, but instead that memory for CS^- items was diminished through fear conditioning. Modifying the task to test these two options is beyond the scope of our replication attempt.

Beyond selective backward and online memory enhancements, the proposed tag-and-capture framework predicts category-specific memory enhancement in a forward, prospective direction. Our results provided indeed evidence for a selective influence of emotionally arousing events on the encoding of subsequent related events. More specifically, the enhanced memory for stimuli paired with aversive shocks seemed to extend to subsequent stimuli belonging to the same category as the CS^+ . Although there was clear evidence for such a selective forward enhancement in the pooled analysis across all four experiments, it is to be noted that this effect was only significant in Experiment 2 and at trend level in Experiments 1, 3, and 4, suggesting a small to moderate effect.

Together, our results suggest a selective memory enhancement for aversive, threat-related stimuli, both online, while a threat is present (e.g., during the fear conditioning procedure), and in a forward direction for threat-related stimuli that are encoded after the threat (e.g., in the postconditioning phase). Both, the online and forward effects may be related to changes in stimulus saliency. During encoding stimuli predictive of motivationally relevant events will be more salient. Likewise, the previously learned association between stimuli and aversive events may increase the saliency of subsequently encoded stimuli that are conceptually linked to the threat-related stimuli. Such increases in saliency may help stimuli to directly exceed the threshold for long-term memory storage. The resulting selectivity in episodic memory has considerable impact on the architecture of our autobiographical memory and, although being generally adaptive, may propel dysfunctional memory in a variety of psychiatric disorders, such as anxiety disorders (Airaksinen et al., 2005; Coles et al., 2007; de Quervain et al., 2017), posttraumatic stress

disorder (Brown et al., 2014; Isaac et al., 2006), or depression (Airaksinen et al., 2007; Lemogne et al., 2006; McDermott & Ebmeier, 2009). In contrast to the online and forward memory enhancements, selective backward enhancements would require the retroactive enhancement of initially weakly encoded (tagged) stimuli to overcome the threshold for long-term storage. Most importantly, however, we obtained only very limited evidence for a selective retroactive (backward) memory enhancement. This raises the question how the brain adapts when certain stimuli only gain relevance after their initial encoding. One solution in line with previous studies is by nonselectively enhancing memory for events preceding an aversive (e.g., stressful) event, regardless of their relation to the relevant event (Cahill et al., 2003; Smeets et al., 2008). In fact, we cannot exclude that such a general, unspecific memory enhancement took place in our experiments. For example, memory for items from the preconditioning phase might have been promoted unspecifically through the following fear-conditioning procedure. Even in Dunsmoor et al. (2015), such a general effect might have played a role in addition to category-specific enhancements for CS⁺ items. However, because this task was specifically designed to investigate category-specific, rather than general retroactive memory enhancement through the within-subject comparison of CS⁺ and CS⁻ items, this question is beyond the scope of this replication attempt.

Another solution that has not been considered by the literature so far could be that in these cases, an even more specific retroactive enhancement takes place, which does not apply to a relatively wide array of stimuli of the same category but only strengthens the memory trace of a single stimulus. Relating back to the example of the bank customer's encounter with the bank robber, importance lies on the memory for only the specific face of the robber and not for other faces seen shortly before (e.g., that of all men). Therefore, adaptive memory would call for a memory promotion of only the specific face and not other faces from the same abstract category. Whether such a mechanism exists, however, is currently unknown and needs to be tested in future research.

In summary, the present series of experiments searched for category-specific, selective retroactive memory enhancement of initially neutral stimuli as suggested by two recent studies (Dunsmoor et al., 2015; Patil et al., 2017). Our data yielded only very limited evidence for a category-specific retroactive memory enhancement in line with Dunsmoor et al. (2015). We acknowledge that although we aimed to stick as closely as possible to the experimental procedure reported by Dunsmoor et al. (2015), subtle differences between studies (e.g., related to the specific sample) can hardly be ruled out. The fact that we did not obtain any evidence for a category-specific retroactive memory enhancement when strictly replicating the reported analysis across four separate experiments, with three of them being highly powered, suggests that this effect is not reliable. At least, the present data clearly question the generalizability of the suggested category-specific retroactive memory enhancement. Still, arousing events might promote episodic memory for recently encountered stimuli in a general, nonselective fashion, as previous evidence suggests (Christianson et al., 1991; McGaugh, 2018; McGaugh & Roozendaal, 2002). These findings of nonselective memory enhancement are in line with previous applications of the synaptic tag-and-

capture mechanism to the behavioral level, which demonstrated memory enhancement for weakly encoded stimuli through following arousing events even in absence of a semantical link between these two (Ballarini et al., 2009, 2013; de Carvalho Myskiw et al., 2013). From a theoretical point of view, this nonselective memory promotion might be regarded as a "safe" alternative to a category-specific retroactive memory promotion, since it does not require a model of events and their putative consequences, which is at risk to be incorrect and might therefore miss important predictors of significant outcomes. On the other hand, such non-specific memory promotion is not only inefficient as invalid cues are subjected to the same memory promotion as valid cues but might also contribute to psychopathology associated with errant memory functions such as posttraumatic stress disorder (Brown et al., 2014; Pitman, 1989). Elucidating how our memory balances the need for efficiency on the one hand and the need for an enhanced storage of experiences that preceded a significant event on the other hand remains a challenge for future research.

Context Paragraph

Our lab focusses on how emotion and stress can bias memory formation. Thus, we were intrigued by recent reports (Dunsmoor et al., 2015; Patil et al., 2017) suggesting a highly specific behavioral tagging mechanism according to which an emotionally arousing event could retroactively enhance memory selectively for preceding events that were conceptually relevant to the emotional event. The proposed mechanism would be highly adaptive in that it would enable our memory to retroactively enhance selectively the storage of material that turned out to be important later on. We aimed to elucidate the mechanisms underlying this selective, retroactive memory enhancement. However, when we tried to replicate the effect in the first place, we originally did not find evidence for a selective retroactive memory enhancement. Only in specific exploratory analyses proposed during the peer-review process, we obtained some limited evidence for the effect. Given the tremendous implications of the suggested retroactive and selective memory enhancement for understanding memory in general and for disorders such as PTSD, we believe that it is important to bring the findings of this series of experiments to the attention of our colleagues. Our hope is that these findings will inspire new theories and experimental paradigms to address the fundamental issue of how our memory can preferentially store events that are relevant for a subsequent emotional episode.

References

- Airaksinen, E., Larsson, M., & Forsell, Y. (2005). Neuropsychological functions in anxiety disorders in population-based samples: Evidence of episodic memory dysfunction. *Journal of Psychiatric Research*, *39*(2), 207–214. <https://doi.org/10.1016/j.jpsychires.2004.06.001>
- Airaksinen, E., Wahlin, Å., Forsell, Y., & Larsson, M. (2007). Low episodic memory performance as a premorbid marker of depression: Evidence from a 3-year follow-up. *Acta Psychiatrica Scandinavica*, *115*(6), 458–465. <https://doi.org/10.1111/j.1600-0447.2006.00932.x>
- Almaguer-Melian, W., Bergado-Rosado, J., Pavon-Fuentes, N., Alberti-Amador, E., Merceron-Martinez, D., & Frey, J. U. (2012). Novelty exposure overcomes foot shock-induced spatial-memory impairment by processes of synaptic-tagging in rats. *Proceedings of the National*

- Academy of Sciences of the United States of America*, 109(3), 953–958. <https://doi.org/10.1073/pnas.1114198109>
- Ballarini, F., Martínez, M. C., Díaz Perez, M., Moncada, D., & Viola, H. (2013). Memory in elementary school children is improved by an unrelated novel experience. *PLoS ONE*, 8(6), e66875. <https://doi.org/10.1371/journal.pone.0066875>
- Ballarini, F., Moncada, D., Martínez, M. C., Alen, N., & Viola, H. (2009). Behavioral tagging is a general mechanism of long-term memory formation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34), 14599–14604. <https://doi.org/10.1073/pnas.0907078106>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory II*. The Psychological Corporation.
- Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1), 80–91. <https://doi.org/10.1016/j.jneumeth.2010.04.028>
- Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature*, 361(6407), 31–39. <https://doi.org/10.1038/361031a0>
- Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, 21(8), 916–944. <https://doi.org/10.1080/09658211.2013.767348>
- Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE*, 5(5), e10773. <https://doi.org/10.1371/journal.pone.0010773>
- Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) Phase II: 930 new normative photos. *PLoS ONE*, 9(9), e106953. <https://doi.org/10.1371/journal.pone.0106953>
- Brown, A. D., Addis, D. R., Romano, T. A., Marmar, C. R., Bryant, R. A., Hirst, W., & Schacter, D. L. (2014). Episodic and semantic components of autobiographical memories and imagined future events in post-traumatic stress disorder. *Memory*, 22(6), 595–604. <https://doi.org/10.1080/09658211.2013.807842>
- Cahill, L., Gorski, L., & Le, K. (2003). Enhanced human memory consolidation with post-learning stress: Interaction with the degree of arousal at encoding. *Learning & Memory*, 10(4), 270–274. <https://doi.org/10.1101/lm.62403>
- Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, 21(7), 294–299. [https://doi.org/10.1016/S0166-2236\(97\)01214-9](https://doi.org/10.1016/S0166-2236(97)01214-9)
- Christianson, S. A., Loftus, E. F., Hoffman, H., & Loftus, G. R. (1991). Eye fixations and memory for emotional events. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(4), 693–701. <https://doi.org/10.1037/0278-7393.17.4.693>
- Coles, M. E., Turk, C. L., & Heimberg, R. G. (2007). Memory bias for threat in generalized anxiety disorder: The potential importance of stimulus relevance. *Cognitive Behaviour Therapy*, 36(2), 65–73. <https://doi.org/10.1080/16506070601070459>
- de Carvalho Myskiw, J., Benetti, F., & Izquierdo, I. (2013). Behavioral tagging of extinction learning. *Proceedings of the National Academy of Sciences of the United States of America*, 110(3), 1071–1076. <https://doi.org/10.1073/pnas.1220875110>
- de Quervain, D., Schwabe, L., & Roozendaal, B. (2017). Stress, glucocorticoids and memory: Implications for treating fear-related disorders. *Nature Reviews Neuroscience*, 18(1), 7–19. <https://doi.org/10.1038/nrn.2016.155>
- Dougal, S., & Rotello, C. M. (2007). Remembering” emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, 14(3), 423–429. <https://doi.org/10.3758/BF03194083>
- Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(1), 130–151. <https://doi.org/10.1037/a0024957>
- Dunsmoor, J. E., & Kroes, M. C. (2019). Episodic memory and Pavlovian conditioning: Ships passing in the night. *Current Opinion in Behavioral Sciences*, 26, 32–39. <https://doi.org/10.1016/j.cobeha.2018.09.019>
- Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology*, 89(2), 300–305. <https://doi.org/10.1016/j.biopsycho.2011.11.002>
- Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, 520(7547), 345–348. <https://doi.org/10.1038/nature14106>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Frey, U., & Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, 385(6616), 533–536. <https://doi.org/10.1038/385533a0>
- Frey, U., & Morris, R. G. M. (1998). Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiation. *Trends in Neurosciences*, 21(5), 181–188. [https://doi.org/10.1016/S0166-2236\(97\)01189-2](https://doi.org/10.1016/S0166-2236(97)01189-2)
- Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68(1), 101–128. <https://doi.org/10.1146/annurev-psych-122414-033625>
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486–504. <https://doi.org/10.1037/1082-989X.3.4.486>
- Isaac, C. L., Cushway, D., & Jones, G. V. (2006). Is posttraumatic stress disorder associated with specific deficits in episodic memory? *Clinical Psychology Review*, 26(8), 939–955. <https://doi.org/10.1016/j.cpr.2005.12.004>
- Jarosch, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, 7(1), 2. <https://doi.org/10.7771/1932-6246.1167>
- Keller, N. E., & Dunsmoor, J. E. (2020). The effects of aversive-to-appetitive counterconditioning on implicit and explicit fear memory. *Learning & Memory*, 27(1), 1212–1219. <https://doi.org/10.1101/lm.050740.119>
- Kim, H., & Cabeza, R. (2009). Common and specific brain regions in high- versus low-confidence recognition memory. *Brain Research*, 1282, 103–113. <https://doi.org/10.1016/j.brainres.2009.05.080>
- Kochli, D. E., Thompson, E. C., Fricke, E. A., Postle, A. F., & Quinn, J. J. (2015). The amygdala is critical for trace, delay, and contextual fear conditioning. *Learning & Memory*, 22(2), 92–100. <https://doi.org/10.1101/lm.034918.114>
- LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, 7(1), 54–64. <https://doi.org/10.1038/nrn1825>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lemogne, C., Piolino, P., Friszer, S., Claret, A., Girault, N., Jouvant, R., Allilaire, J. F., & Fossati, P. (2006). Episodic autobiographical memory in depression: Specificity, autoeocetic consciousness, and self-perspective.

- Consciousness and Cognition*, 15(2), 258–268. <https://doi.org/10.1016/j.concog.2005.07.005>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Erlbaum.
- Malenka, R. C., & Nicoll, R. A. (1999). Long-term potentiation—A decade of progress? *Science*, 285(5435), 1870–1874. <https://doi.org/10.1126/science.285.5435.1870>
- Martin, K. C., & Kosik, K. S. (2002). Synaptic tagging—Who's it? *Nature Reviews Neuroscience*, 3(10), 813–820. <https://doi.org/10.1038/nrn942>
- Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy*, 36(4), 455–470. [https://doi.org/10.1016/S0005-7967\(97\)10031-6](https://doi.org/10.1016/S0005-7967(97)10031-6)
- McDermott, L. M., & Ebmeier, K. P. (2009). A meta-analysis of depression severity and cognitive function. *Journal of Affective Disorders*, 119(1-3), 1–8. <https://doi.org/10.1016/j.jad.2009.04.022>
- McGaugh, J. L. (2015). Consolidating memories. *Annual Review of Psychology*, 66(1), 1–24. <https://doi.org/10.1146/annurev-psych-010814-014954>
- McGaugh, J. L. (2018). Emotional arousal regulation of memory consolidation. *Current Opinion in Behavioral Sciences*, 19, 55–60. <https://doi.org/10.1016/j.cobeha.2017.10.003>
- McGaugh, J. L., & Roozendaal, B. (2002). Role of adrenal stress hormones in forming lasting memories in the brain. *Current Opinion in Neurobiology*, 12(2), 205–210. [https://doi.org/10.1016/S0959-4388\(02\)00306-9](https://doi.org/10.1016/S0959-4388(02)00306-9)
- McLaughlin, J., Skaggs, H., Churchwell, J., & Powell, D. A. (2002). Medial prefrontal cortex and Pavlovian conditioning: Trace versus delay conditioning. *Behavioral Neuroscience*, 116(1), 37–47. <https://doi.org/10.1037/0735-7044.116.1.37>
- Moncada, D., Ballarini, F., & Viola, H. (2015). Behavioral tagging: A translation of the synaptic tagging and capture hypothesis. *Neural Plasticity*, 2015, 650780. <https://doi.org/10.1155/2015/650780>
- Moncada, D., & Viola, H. (2007). Induction of long-term memory by exposure to novelty requires protein synthesis: Evidence for a behavioral tagging. *The Journal of Neuroscience*, 27(28), 7476–7481. <https://doi.org/10.1523/JNEUROSCI.1083-07.2007>
- Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). Package “BayesFactor” 0.9.12-4.2. <https://cran.r-project.org/web/packages/BayesFactor/index.html>
- Mulligan, N. W. (1998). The role of attention during encoding in implicit and explicit memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(1), 27–47. <https://doi.org/10.1037/0278-7393.24.1.27>
- Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General*, 145(5), 548–558. <https://doi.org/10.1037/xge0000158>
- Nairne, J. S., & Pandeirada, J. N. S. (2008). Adaptive memory: Remembering with a stone-age brain. *Current Directions in Psychological Science*, 17(4), 239–243. <https://doi.org/10.1111/j.1467-8721.2008.00582.x>
- Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. S. (2007). Adaptive memory: Survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2), 263–273. <https://doi.org/10.1037/0278-7393.33.2.263>
- Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, 114(5), 657–664. <https://doi.org/10.1037/pspa0000121>
- Oyarzún, J. P., Packard, P. A., de Diego-Balaguer, R., & Fuentemilla, L. (2016). Motivated encoding selectively promotes memory for future inconsequential semantically-related events. *Neurobiology of Learning and Memory*, 133, 1–6. <https://doi.org/10.1016/j.nlm.2016.05.005>
- Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning & Memory*, 24(1), 6565–6569. <https://doi.org/10.1101/lm.042978.116>
- Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, 139(6), 1173–1203. <https://doi.org/10.1037/a0033044>
- Phelps, E. A., & Sharot, T. (2008). How (and why) emotion enhances the subjective sense of recollection. *Current Directions in Psychological Science*, 17(2), 147–152. <https://doi.org/10.1111/j.1467-8721.2008.00565.x>
- Pitman, R. K. (1989). Post-traumatic stress disorder, hormones, and memory. *Biological Psychiatry*, 26(3), 221–223. [https://doi.org/10.1016/0006-3223\(89\)90033-4](https://doi.org/10.1016/0006-3223(89)90033-4)
- Redondo, R. L., & Morris, R. G. M. (2011). Making memories last: The synaptic tagging and capture hypothesis. *Nature Reviews Neuroscience*, 12(1), 17–30. <https://doi.org/10.1038/nrn2963>
- Rogerson, T., Cai, D. J., Frank, A., Sano, Y., Shobe, J., Lopez-Aranda, M. F., & Silva, A. J. (2014). Synaptic tagging during memory allocation. *Nature Reviews Neuroscience*, 15(3), 157–169. <https://doi.org/10.1038/nrn3667>
- Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, 70(2), 389–401. <https://doi.org/10.3758/pp.70.2.389>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Salehi, B., Cordero, M. I., & Sandi, C. (2010). Learning under stress: The inverted-U-shape function revisited. *Learning & Memory*, 17(10), 522–530. <https://doi.org/10.1101/lm.1914110>
- Schulz, P., Schlotz, W., & Becker, P. (2004). *Trierer Inventar zum chronischen Stress: TICS* [Trier Inventory for Chronic Stress (TICS)]. Hogrefe.
- Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. *Neuroscience and Biobehavioral Reviews*, 36(7), 1740–1749. <https://doi.org/10.1016/j.neubiorev.2011.07.002>
- Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences*, 14(10), 464–472. <https://doi.org/10.1016/j.tics.2010.08.002>
- Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, 33(1), 151–170. <https://doi.org/10.3758/BF03195305>
- Smeets, T., Otgaar, H., Candel, I., & Wolf, O. T. (2008). True or false? Memory is differentially affected by stress-induced cortisol elevations and sympathetic activity at consolidation and retrieval. *Psychoneuroendocrinology*, 33(10), 1378–1386. <https://doi.org/10.1016/j.psyneuen.2008.07.009>
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117(1), 34–50. <https://doi.org/10.1037/0096-3445.117.1.34>
- Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Consulting Psychologists Press.
- Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen MDBF* [Multidimensional mood questionnaire]. Hogrefe.
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1), 59–71. <https://doi.org/10.1177/1745691613514450>
- Uncapher, M. R., & Rugg, M. D. (2005). Effects of divided attention on fMRI correlates of memory encoding. *Journal of Cognitive*

- Neuroscience*, 17(12), 1923–1935. <https://doi.org/10.1162/089892905775008616>
- Vogel, S., & Schwabe, L. (2016). Stress in the zoo: Tracking the impact of stress on memory formation over time. *Psychoneuroendocrinology*, 71, 64–72. <https://doi.org/10.1016/j.psyneuen.2016.04.027>
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Beek, T., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928. <https://doi.org/10.1177/1745691616674458>
- Wang, S.-H., Redondo, R. L., & Morris, R. G. M. (2010). Relevance of synaptic tagging and capture to the persistence of long-term potentiation and everyday spatial memory. *Proceedings of the National Academy of Sciences of the United States of America*, 107(45), 19537–19542. <https://doi.org/10.1073/pnas.1008638107>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.
- Weike, A. I., Schupp, H. T., and Hamm, A. O. (2007). Fear acquisition requires awareness in trace but not delay conditioning. *Psychophysiology*, 44, 170–180. <https://doi.org/10.1111/j.1469-8986.2006.00469.x>

Received December 18, 2019

Revision received March 5, 2021

Accepted March 8, 2021 ■