Supplemental Material for: *Aversive learning retroactively prioritizes neutral episodic memories structured by prior knowledge of predictive sequences*

Blazej M. Baczkowski[1,2,3,4] & co-authors

[1] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[2] IMPRS NeuroCom, Leipzig, Germany

[3] Institute of Psychology, University of Leipzig, Leipzig, Germany

[4] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

Author Note

This manuscript is a preprint and has not been peer reviewed. Blazej M. Baczkowski is now at the Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany. Correspondence concerning this article should be addressed to Blazej M. Baczkowski, blazej.baczkowski@uni-tuebingen.de, University of Tübingen.

Contents

Deviations from the pre-registration

Deviations from the pre-registration are listed below in Table S1.

Table S1
Preregistration deviations.

No.	Details	Original wording	Deviation description	Extent of deviation	Judgment of impact
1	Study: 1 Type: Analyses Reason: New knowledge Timing: After results known	The analysis plan of the semantic judgment task included the test for the effect of `position` and the effect of `sequence`.	Main text reports an analysis using nested contrasts (effect of `position` nested within `sequence`)	Minor	The planned analysis of the main effects is reported in the supplemental tables of the models testing the interaction effect. The deviation has a positive impact, i.e., the analysis is more suitable to the nature of the design.
2	Study: 1 Type: Analyses Reason: New knowledge Timing: After results known	The analysis plan for the recognition task included percent correct as the main outcome measure, alongside corrected recognition score (P[hit] − P[FA]).	Main text reports linear-model analyses with d-prime (a common signal-detection measure) instead of corrected recognition score. We excluded corrected recognition from the main analyses because it is the difference of two binomial proportions; when analyzed with a Gaussian model, this can produce heteroskedastic residuals and biased p-values due to distorted standard errors.	Minor	The analyses with corrected recognition as the outcome measure are reported in the supplemental results. The conclusions remain consistent across various analysis strategies.
3	Study: 1 Type: Analyses Reason: New knowledge Timing: After data access	The analysis plan of the recognition task did not include the nuisance predictor of `semantic category`.	Main text reports analyses of linear models including `semantic category` as a nuisance predictor.	Minor	The analyses without the nuisance predictor of `semantic category` are reported in the supplemental results. The conclusions are the same.

Supplemental Methods

Participants

The study was approved by the local ethical review board (University of Leipzig, Germany) and carried out in accordance with the Declaration of Helsinki. We recruited healthy right-handed volunteers between 18 and 35 years old with normal or corrected-to-normal vision from the general population in the city of Leipzig, Germany using MPI-CBS participant database. Eligibility criteria excluded pregnancy, current or a history of a neurological, psychiatric or drug-abuse disorder. Further eligibility criteria excluded use of medication (except for oral contraceptives and pain killers) and excessive alcohol consumption 24 hours before the study.

Forty-five volunteers provided their written informed consent prior to the start of the study. One participant did not show up for the second day. One discontinued the experiment following the shock intensity calibration on day 1. One could not concentrate on learning the order of categories during day 1 and stopped the experiment. One exhibited weak compliance to the staircase procedure of shock calibration and reported after conditioning that the shock was not aversive at all, and therefore was not invited for the second day.

To ensure sufficient data quality, we specified two pre-registered exclusion criteria. First, to ensure that participants learned the category level association between its exemplars and unconditioned stimulus (US) through experience during conditioning, participants who did not verbalize explicit awareness of the contingency between the conditioned category and the US were not invited for the second day ($n = 2$). Second, to ensure successful acquisition of temporal sequence knowledge, participants were excluded if they did not demonstrate explicit knowledge of the temporal arrangement of six semantic categories on day 1 ($n = 6$). Four of these participants were not invited back for day 2, including one who also failed to show awareness of the CS-US contingency. Sequence knowledge was assessed with a drag-and-drop test (see below) administered just before conditioning, in which participants were required to correctly arrange all three categories for each sequence at least once.

Therefore, the final sample of participants who completed the experiment and met our pre-registered inclusion criteria resulted in 34 participants (18 female) between 19 to 35 years old (mode = 24), as specified in the pre-registered sampling plan.

Behavioral tasks

Semantic judgment task. Six semantic categories were organized into two temporal sequences, each comprising three serial positions, which predicted the occurrence of category exemplars in a semantic judgment task (cf., Hsieh, Gruber, Jenkins, & Ranganath, 2014). In this task, participants were instructed to make a semantic decision in every trial (e.g., "Is it a furniture or a tool?"). Each trial started with a presentation of two category names at the bottom of a screen (2 s), followed by a fixation cross (0.5 s) and the presentation of an image (2.5 s). Participants could respond during the image presentation as long as the two options were on the screen. The maximum response duration was set to 1 s for all participants during training. During incidental encoding, this duration varied across participants to maintain a similar response pace across the block; it was set individually to the participant's mean training reaction time plus 1 standard deviation. Participants received visual feedback to their button press which vanished at the image offset, but no information about the accuracy of their response. Trials were spaced by 0.5 s interval and organized in a set of three, forming a triplet. Each triplet followed an order of categories given their serial positions in a sequence. For example, for a given triplet, participants always saw a sequence of three trials such that an image of a music instrument was followed by an image of a tool, followed by an image of a vegetable. To separate one sequence from another, triplets were spaced by 6 ± 2 s interval filled with a blue fixation dot. The order of triplets was pseudo-randomized so that no more than three triplets of the same sequence could occur in a row.

To learn the two sequences of categories, participants started with a training where the semantic judgment task consisted of three blocks spaced by the drag and drop test. In each block, they saw one new exemplar of each category that was presented four times per sequence (8 triplets in total). The 18 unique category exemplars were not included later in

either the threat conditioning task or the surprise recognition test. In the drag and drop test, all six category names were simultaneously presented on a screen and participants had to reconstruct the order of categories in which their images were shown before, followed by a feedback on their performance. In case of an incorrect response, they were asked to reconstruct the sequence again. After the training, participants continued with the task to incidentally encode 150 trial-unique exemplars (25 per category) in one block (50 triplets in total). To test for the acquisition of relational knowledge among the categories, participants were subsequently asked to do the drag and drop test twice (spaced by a 4 s interval), but without feedback on their performance.

The allocation of categories to serial positions was pseudo-randomized across participants. First, to maximize a counterbalanced allocation of categories to the pre-conditioned positions that constituted our main hypothesis (see also statistical analyses), four categories were selected (i.e., "furniture", "tools", "vehicles", and "music instruments") and randomly assigned to the first and second positions in two sequences. The four categories were selected to minimize the influence of any intrinsic differences in 24-hour recognition memory across pre-conditioned categories that was tested in a pilot study (see Supplemental Experiment 2). Second, to fully counterbalance categories across the CS+/CS- conditions in the conditioning, two other categories (i.e., "fruits & vegetables" and "clothes") were always assigned to the third position.

Skin conductance acquisition and pre-processing

Skin conductance was collected at sampling rate of 500 Hz using BIOPAC MP35 (Biopac Systems Inc., CA, USA) with disposable snap Ag/AgCl electrodes attached to the non-dominant (left) hand on the intermediate phalanges of the index and middle fingers. Continuous raw skin conductance time series were analyzed in the window starting from 10 s prior to the onset of the first stimulus and terminating 16 s after the offset of the last stimulus. The raw time series was filtered with a 1st order butterworth high-pass filter at 0.05 Hz and smoothed with a low-pass moving average filter at 5 Hz. To avoid any remaining high-frequency fluctuations, the time series was down-sampled at 10 Hz and

interpolated using linear interpolation.

Procedure

Two experimental sessions took place on two consecutive days in the same test room and around the same time of day. Stimulus presentations were generated with the Psychophysics Toolbox (Brainard, 1997) for MATLAB 2016a (the MathWorks). Participants viewed the stimuli from a distance of approx. 60 cm on a 22" display (Ilyama Vision Master Pro 514, resolution 1280 x 1024, aspect ratio 5:4, refresh rate 80 Hz) in a dimly lit room.

The first session (~90 min) consisted of a pre-conditioning phase (i.e., when the threat of the US was absent) followed by a conditioning phase. In the pre-conditioning phase, participants performed a cover task, i.e., a semantic judgment task, to first learn that six semantic categories were assigned to two temporal sequences of three serial positions each and next incidentally encode their trial-unique exemplars (i.e., images). In the semantic judgment task, participants viewed neutral images of everyday objects of the six categories (i.e., "tools", "fruits & vegetables", "furniture", "vehicles", "clothes", "music instruments") and were instructed to make semantic decisions about each image presented (e.g., "Is it a furniture or a tool?") while their response accuracy and reaction times were recorded (cf. Hsieh et al., 2014). In the learning part, participants were explicitly instructed that the six categories were split into two sequences, i.e., categories appeared in a specific temporal order to which they should have paid attention and have learned so that it could help them respond faster during the task. In the incidental encoding part, participants were told that the task would continue in the same way with the exception that images would always be new. There was no instruction to remember the images.

After the semantic judgment task, the session proceeded with the attachment of the electrodes to administer electric shocks (~10 min). Shock electrodes were attached to ring and pinky fingers of the dominant (right) hand since the other fingers were used for providing responses. The shock was delivered via a constant current stimulator (Digitimer DS7A, Digitimer Ltd., the UK) and consisted of a 200-ms train of 40 square pulses with

individual pulse width of 0.2 ms at frequency of 200 Hz. The shock intensity level was calibrated using an ascending staircase procedure starting with a low voltage (near a perceptible threshold) to reach a level deemed "maximally uncomfortable without being painful" by the participant. To this end, shock intensity was scored by the participant on a modified pain assessment scale from 0 (no sensation) to 10 (very high intensity), and ranged from 6 to 9 (mode = 8) in the current sample.

To form Pavlovian threat memory, participants were subsequently exposed to the category-based differential delay threat conditioning procedure (e.g., de Voogd, Fernández, & Hermans, 2016; Dunsmoor, Martin, & LaBar, 2012) while concurrent measurements of their skin conductance were being collected. Participants were told that they would see images of two out of the six categories they knew from the previous part. They were told that when there was an image on the screen, they may receive a shock and there was a relationship between the shocks and the categories. Participants were not explicitly instructed the CS-US contingency but they were instructed to figure out the contingency through experience.

After the conditioning task, the shock electrodes were removed and the measurements of skin conductance were stopped. At the conclusion of the session, participants were asked to rate the intensity of the shock felt during the study on a scale from 0 (not at all unpleasant) to 10 (extremely unpleasant), and how much fear they felt during the experiment from 0 (not at all afraid) to 10 (extremely afraid). They were also asked to estimate how many shocks they received throughout the whole study (not counting shocks received during the calibration procedure) and report their understanding of what the contingency between the categories and shocks was, that was then rated on a 3-point scale by an experimenter (0-not aware; 1-sort of aware; 2-aware).

In the second session (~60 min), participants were first screened whether they consumed alcohol or used medication in the past 24 hours. Next, participants started with a surprise recognition memory test. To assess whether participants expected the surprise memory test, they were asked just prior to the memory test whether they had any expectations for the experiment ("Do you have any expectations of what the next task

might be about: yes or no?"). Participants were then told that there would be a test of their memory for the pictures they had seen earlier, and were asked to indicate on a 5-point scale how surprised they were by a memory test, from 1 ("I did not expect a memory test at all") to 5 ["Yes, I knew there would be a memory test"; Dunsmoor and Paz (2015); Dunsmoor et al. (2018)]. In the current sample, the surprise was rated from 1 to 4 (mode = 3).

To test for the relational memory among the categories acquired on the previous day, at the end of the recognition memory test, participants were asked to order the six category names into two sequences in the way they remembered experiencing them on the previous day (see the drag and drop test).

At the conclusion of the session, participants were asked to fill out four self-report questionnaires: (a) the trait inventory of the State-Trait Anxiety Inventory-STAI (Laux, Glanzmann, Schaffner, & Spielberger, 1981; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983); (b) Childhood Trauma Questionnaire-CTQ (Bernstein et al., 1994, 2003; Klinitzke, Romppel, Häuser, Brähler, & Glaesmer, 2012); (c) Intolerance of Uncertainty Scale-IUS (Buhr & Dugas, 2002; Gerlach, Andor, & Patzelt, 2008); and (d) Berkman-Syme Social Network Index-SNI (Berkman & Syme, 1979).

The study ended with the debriefing – the rationale, motivation, and expected contribution of the study was briefly described. Finally, participants were compensated for their participation.

Statistical analyses

To test whether participants utilized the relational knowledge among semantic categories in the semantic judgment task during the incidental encoding, we specified statistical models that quantified the fixed effect of trial position (3-level factor) within a sequence on accuracy judgment and its reaction time (RT). We expected the performance (accuracy and RT) to improve across the positions regardless of the sequence. To this end, we used sum contrast to test successive differences between positions (i.e., 2-1 and 3-2) nested within a sequence. To additionally test whether the effect of position depended on a

sequence, we specified a separate linear model including an interaction term. For the accuracy, we fitted a GLMM with a *logit* link function and binomial error structure to the number of correct responses out of total number of valid trials, i.e., trials in which a response was given within the pre-specified time limit. We additionally re-fitted the model to the number of correct responses out of total number of trials including omissions, i.e, trials in which a response was given after the pre-specified time limit. In this model, omissions were treated as valid but incorrect responses. Responses that occurred before the image onset were treated as invalid, i.e., they were treated as NAs. To account for the varying number of valid trials across participants, we included weights in both models that corresponded to the number of valid trials per condition. To verify the results due to potential under- or over-dispersion, we additionally fitted GLMMs with a *logit* link function and binomial error structure to binary responses in a single trial. For the RT, we fitted a linear mixed model (LMM) with an identity link function and a Gaussian error structure to investigate the effect of trial position on the average RT to correct trials.

To test whether participants acquired Pavlovian conditioned threat responses, we specified statistical models that quantified the effect of condition (CS+ vs. CS-) on shock expectancy ratings and square-root transformed skin conductance response (SCR) magnitude. We expected higher responses in the CS+ than CS- condition. To test the fixed effect of the two-level factor of condition on the number of responses indicating "yes" (i.e., 1) out of total number of trials per condition (i.e., 25), we fitted a GLMM with a *logit* link function and binomial error structure including an additional control predictor of category identity. To verify the results of the model which exhibited a large under-dispersion parameter, we fitted an additional GLMM with a *logit* link function and binomial error structure to binary responses in a single trial. To test the fixed effect of a two-level factor of condition on square-root transformed SCR magnitude, we fitted a LMM with an identity link function and a Gaussian error structure including an additional control predictor of category identity.

To analyze data of the recognition memory task, we aggregated the number of correct

responses to the total number of responses per category, which is expressed as proportion of correct responses, i.e., when an old item is identified as "old" and a new item is identified as "new" for the respective semantic category. To keep the number of responses equal across participants, the "old"/"new" responses were collapsed across the two confidence levels (i.e., "maybe" and "definitely) in line with prior reports (de Voogd et al., 2016; Dunsmoor, Murty, Davachi, & Phelps, 2015; Kalbe & Schwabe, 2020). Given that the number of old items was equal to the number of new items, the proportion of correct responses was equivalent to the corrected recognition score, expressed as P(hit) - P(false alarm) – the correlation of the two indices is equal to 1 (see Supplemental Experiment 3). To verify the results of proportion correct, we additionally expressed the recognition memory as d-prime index – measure based on signal detection theory. Hit rate was calculated as total number of correct "definitely old" and "maybe old" responses divided by the the total number of old items of the respective category. Likewise, false alarm rate was calculated as total number of incorrect "definitely old" and "maybe old" responses divided by the total number of new items of the respective category. Next, the inverse of the standard normal cumulative distribution function was evaluated at the probability values of P(hit) and P(false alarm) that were then subtracted from one another. If hit or false alarm rates were equal to one or zero, the scores were replaced with 1–1/(2 * number of responses) and 1/(2 * number of responses), respectively, where responses refer to the total number of old and new items from a respective category.

To explain the recognition memory variance allocated exclusively to the pre-conditioning phase as our hypothesis states, we split the recognition data into pre-conditioned categories (allocated to positions 1 and 2) and conditioned categories (allocated to position 3). This way, the analysis of the pre-conditioned categories was treated as the main test of our hypothesis. To verify whether the data split affected the results, we subsequently fitted a corresponding model comprising the hypothesized effect of pre-conditioned categories to the whole data set.

To test whether Pavlovian conditioning affected the profile of recognition memory of pre-conditioned categories, we fitted a linear model to the data of the four categories

presented only in the pre-conditioning phase (i.e., “furniture”, “tools”, “vehicles”, and “music instruments”). We used sum contrast and included a two-way interaction between two-level factor of position (1 vs 2) and two-level factor of sequence (CS+ vs CS-) as a test predictor together with its lower terms (i.e., main effect of position and sequence). To test whether Pavlovian conditioning prioritized memory for the CS+ category compared with the CS- category, we fitted a linear model to the data of the two categories presented in both the pre-conditioning and conditioning phase (i.e., ”fruits & vegetables” and “clothes”). The effect of category was coded using a sum contrast. In all analyses, we included category identity as a control predictor. Proportion of correct responses was analyzed using a GLMM with a *logit* link function and binomial error structure while d-prime was analyzed using a corresponding LMM with an identity link function and Gaussian error structure. To verify the results of the GLMM, we also fitted the same model to single trial data where response was coded as either 1 or 0.

Supplemental Results

Fitting an alternative model to the whole dataset

To verify whether the data split affected the results of the interaction effect for pre-conditioned categories, we ran a binomial model with the *logit* link function fitted to the whole dataset, i.e., comprising data from the pre-conditioning and conditioning phase. We used a coding scheme of sum contrast, and extended the original model by adding a 2-level predictor of *phase* (pre-conditioning vs. conditioning), a 2-level predictor of *CS* (CS+ vs. CS- category), and corresponding predictors to account for the differences related to category identity. The GLMM revealed a significant interaction effect ($\chi^2(1) = 5.17$, $p = .023$, marginal-$R^2 = 0.20$, conditional-$R^2 = 0.41$, $\phi = 0.79$).

Controlling for the performance in the semantic judgment task

To verify whether the interaction effect in the pre-conditioned categories was independent from the performance in the semantic judgment task, we ran three additional models that included an extra control predictor: proportion of correct responses to valid

trials, proportion of correct responses to valid trials together with omissions, or average reaction time to correct trials, respectively. In all three GLMMs interaction effect remained significant ($\chi^2(1) = 5.05$, $p = .025$; $\chi^2(1) = 4.62$, $p = .032$; $\chi^2(1) = 4.37$, $p = .037$ for each model respectively).

Exploring dependence on individual differences and conditioning-specific physiological arousal

To explore whether the interaction effect in the pre-conditioned categories was dependent on individual differences in trait anxiety [STAI-T; Laux et al. (1981); Spielberger et al. (1983)], intolerance for uncertainty [IUS; Buhr and Dugas (2002); Gerlach et al. (2008)], or childhood traumatic experience [CTQ; Bernstein et al. (1994); Bernstein et al. (2003); Klinitzke et al. (2012)], as well as the physiological arousal (SCR) elicited during conditioning, we ran four additional LMMs. To simplify the model, we computed a difference score between the first and second position for each sequence based on the d-prime index of confidence-collapsed responses. Next, we included a two-level control predictor of a sequence (i.e., CS+ vs. CS- condition) and a numerical test predictor corresponding to the respective variable of interest. None of the models revealed a significant dependence of recognition memory on these measures ($\chi^2(1) = 2.84$, $p = .092$; $\chi^2(1) = 0.20$, $p = .653$; $\chi^2(1) = 0.920$, $p = .338$; $\chi^2(1) = 2.97$, $p = .085$; for STAI, IUS, CTQ, and SCR, respectively).

Considering only high confidence responses

To explore whether the interaction effect in the pre-conditioned categories could be obtained from only "high confidence" responses, d-prime index was recalculated and the LMM model was re-run (Figure S1). Hit rate was calculated as a number of correct "definitely old" responses divided by the the total number of "definitely old" or "definitely new" responses to old exemplars of the respective category. False alarm rate was calculated as a number of incorrect "definitely old" responses divided by the the total number of "definitely old" or "definitely new" responses to new exemplars of the respective category. Hit and false alarm rates were z-transformed and subtracted from one another leading to a new d-prime index. "High confidence" responses constituted about 50% of all responses –

responses ranged greatly from 2 to 25 (out of 25 possible) and the mode across the four categories ranged from 12 to 16. We did not observe the effect of interaction but the effect of position only ($\chi^2(1) = 5.48$, $p = .019$, marginal-$R^2 = 0.05$, conditional-$R^2 = 0.53$). Numerical comparison revealed that d-prime was increased for the items of the categories allocated to the first than second position (CS+ sequence: P1 = 2.12 [1.88, 2.34]; P2 = 1.79 [1.56, 2.02]; CS- sequence: P1 = 2.04 [1.82, 2.25]; P2 = 1.88 [1.62, 2.14]). Yet, the effect of the position appeared to be stronger in the CS+ sequence than in the CS- sequence (Cohen's d: 0.16 and 0.08, respectively), suggesting that perhaps the lack of the interaction effect could be potentially caused by lower number of responses used to estimate the d-prime index compared to the results based on all responses. In line with the results based on all responses, we did not observe any differences in the conditioned categories ($\chi^2(1) = 1.27$, $p = .259$, marginal-$R^2 = 0.14$, conditional-$R^2 = 0.53$).

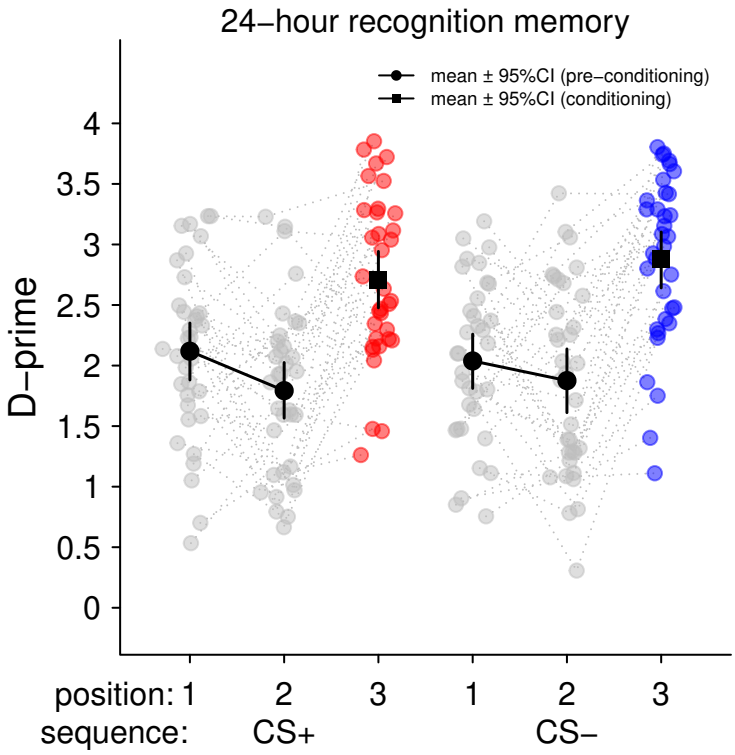

Figure S1. 24-hour recognition memory based on "high confidence" responses only. While the d-prime of conditioned categories (CS+ vs. CS-) was similar ($\chi^2(1) = 1.27$, $p = .259$, marginal-$R^2 = 0.14$, conditional-$R^2 = 0.53$, $\sigma = 0.41$), it differed in the pre-conditioned categories between positions regardless of the sequence ($\chi^2(1) = 5.48$, $p = .019$, marginal-$R^2 = 0.05$, conditional-$R^2 = 0.53$, $\sigma = 0.42$).

Dropping the control predictor of category identity. Since our pre-registered analysis plan did not specify explicitly to use category identity as a control predictor in the linear models, we report the analyses where the control predictor was dropped. Both

analyses, using binomial (proportion correct) and gaussian (d-prime) models, showed that the interaction effect of memory recognition in the pre-conditioned categories remained significant (GLMM: $\chi^2(1) = 4.59$, $p = .032$, marginal-$R^2 = 0.07$, conditional-$R^2 = 0.52$, $\phi = 0.78$; LMM: $\chi^2(1) = 4.72$, $p = .030$, marginal-$R^2 = 0.06$, conditional-$R^2 = 0.20$, $\sigma = 0.48$).

Additional outcome measure of recognition memory. The analysis plan treated percent of correct response as the main outcome measure of recognition memory which was reported in the main text. To verify these results with a more common approach, we ran additional analyses based on d-prime score (Neath & Surprenant, 2003; Yonelinas & Parks, 2007). We did not include the d-prime score explicitly in the pre-registration. Instead, we included corrected recognition score: hit rate minus false-alarm rate, based on which the d-prime score is based. We decided to use the d-prime rather than corrected recognition score because the latter is the difference of two binomial proportions; analyzing it with a Gaussian model inflates noise, ignores heteroskedasticity, and can distort p-values (Dixon, 2008; Jaeger, 2008). The d-prime instead is a continous metric and therefore can be analyzed with a gaussian linear model. For transparency reasons, however, we report the analysis using corrected recognition score. These data revealed a similar profile of responses as the d-prime index. Responses to the conditioned categories were similar (CS+ = 0.64 [0.59, 0.69], CS- = 0.65 [0.60, 0.69]), and the LMM did not reveal a significant effect of the condition ($\chi^2(1) = 0.02$, $p = .884$, marginal-$R^2 = 0.20$, conditional-$R^2 = 0.41$). In line with the results of the d-prime, numerical comparisons of the pre-conditioned categories revealed that while memory recognition for the items of categories embedded in the CS+ sequence was better for the first position (0.46 [0.40, 0.51], compared with the second position (0.35 [0.29, 0.42]), it was similar for the items of both categories embedded in the CS- sequence (P1: 0.44 [0.40, 0.49]; P2: 0.44 [0.38, 0.49]). This interaction was significant in the LMM that included a control predictor of category identity ($\chi^2(1) = 3.90$, $p = .048$, marginal-$R^2 = 0.07$, conditional-$R^2 = 0.25$).

Supplemental Tables

Table S2
Semantic judgment task (valid trials): nested effect (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			3.90e+00	3.56e+00	4.24e+00
s2-s1			-7.74e-01	-2.05e+00	5.01e-01
s1p2-p1*			1.42e+00	6.84e-01	2.17e+00
s1p3-p2*			-2.28e-01	-1.06e+00	6.05e-01
s2p2-p1*			7.18e-01	2.15e-01	1.22e+00
s2p3-p2*			1.23e+00	4.24e-01	2.03e+00
Random effects					
Subject	(34 levels)				
	intercept				
intercept	4.68e-01				
Subject	(34 levels)				
	s1p2-p1				
s1p2-p1	2.39e-01				
Subject	(34 levels)				
	s2p2-p1				
s2p2-p1	3.3e-07				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	5	4.73e+02			
Full	9	4.28e+02	4.52e+01	4	3.57e-09

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S3
Semantic judgment task (valid trials): nested effect (single trial data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			3.89e+00	3.56e+00	4.22e+00
s2-s1			-7.06e-01	-1.92e+00	5.11e-01
s1p2-p1*			1.45e+00	7.57e-01	2.14e+00
s1p3-p2*			-2.28e-01	-1.04e+00	5.88e-01
s2p2-p1*			7.19e-01	2.26e-01	1.21e+00
s2p3-p2*			1.23e+00	4.39e-01	2.02e+00
Random effects					
Subject	(34 levels)				
	intercept				
intercept	4.66e-01				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	3	1.23e+03			
Full	7	1.17e+03	6.14e+01	4	1.48e-12

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S4
Semantic judgment task (valid trials): interaction effect (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			3.90e+00	3.54e+00	4.25e+00
s2-s1			-7.10e-01	-1.95e+00	5.33e-01
p2-p1			2.19e+00	1.31e+00	3.06e+00
p3-p2			8.03e-01	-7.70e-01	2.38e+00
s:p2-p1*			-7.40e-01	-1.61e+00	1.30e-01
s:p3-p2*			1.46e+00	3.05e-01	2.62e+00
Random effects					
Subject	(34 levels)				
	intercept	p3-p2			
intercept	4.07e-01	-3.2e-01			
p3-p2	-3.2e-01	1.24e+00			
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	7	4.33e+02			
Full	9	4.27e+02	6.58e+00	2	3.73e-02

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S5
Semantic judgment task (valid trials): interaction effect (single trial data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			3.90e+00	3.56e+00	4.23e+00
s2-s1			-7.07e-01	-1.92e+00	5.06e-01
p2-p1			2.15e+00	1.29e+00	3.00e+00
p3-p2			1.00e+00	-1.31e-01	2.13e+00
s:p2-p1*			-7.21e-01	-1.60e+00	1.56e-01
s:p3-p2*			1.46e+00	3.24e-01	2.59e+00
Random effects					
Subject	(34 levels)				
	intercept				
intercept	4.66e-01				
Subject	(34 levels)				
	s:p2-p1				
s:p2-p1	3.23e-01				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	6	1.18e+03			
Full	8	1.17e+03	6.52e+00	2	3.84e-02

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S6
Semantic judgment task (including omissions): nested effect (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			2.42e+00	2.23e+00	2.62e+00
s2-s1			-1.99e-01	-8.48e-01	4.50e-01
s1p2-p1*			8.95e-01	5.63e-01	1.23e+00
s1p3-p2*			6.90e-01	2.77e-01	1.10e+00
s2p2-p1*			7.21e-01	3.14e-01	1.13e+00
s2p3-p2*			1.08e+00	6.70e-01	1.49e+00
Random effects					
Subject	(34 levels)				
	intercept				
intercept	2.06e-01				
Subject	(34 levels)				
	s1p2-p1				
s1p2-p1	1.41e-01				
Subject	(34 levels)				
	s2p2-p1				
s2p2-p1	7.3e-01				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	5	8.43e+02			
Full	9	7.44e+02	9.93e+01	4	1.36e-20

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S7
Semantic judgment task (including omissions): nested effect (single trial data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			2.43e+00	2.24e+00	2.62e+00
s2-s1			-1.75e-01	-8.83e-01	5.33e-01
s1p2-p1*			8.94e-01	5.65e-01	1.22e+00
s1p3-p2*			6.90e-01	2.78e-01	1.10e+00
s2p2-p1*			7.29e-01	3.20e-01	1.14e+00
s2p3-p2*			1.08e+00	6.72e-01	1.49e+00
Random effects					
Subject	(34 levels)				
	intercept				
intercept	2.06e-01				
Subject	(34 levels)				
	s2-s1				
s2-s1	6.14e-01				
Subject	(34 levels)				
	s1p2-p1				
s1p2-p1	1.29e-01				
Subject	(34 levels)				
	s2p2-p1				
s2p2-p1	7.28e-01				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	6	3.29e+03			
Full	10	3.19e+03	9.98e+01	4	1.1e-20

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S8
Semantic judgment task (including omissions): interaction effect (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			2.43e+00	2.23e+00	2.63e+00
s2-s1			-3.25e-01	-9.61e-01	3.12e-01
p2-p1			1.66e+00	1.07e+00	2.25e+00
p3-p2			1.77e+00	1.19e+00	2.36e+00
s:p2-p1*			-1.55e-01	-6.08e-01	2.98e-01
s:p3-p2*			3.87e-01	-1.96e-01	9.70e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	2.17e-01				
Subject	(34 levels)				
	p2-p1				
p2-p1	1.44e+00				
Subject	(34 levels)				
	s:p2-p1				
s:p2-p1	2.78e-01				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	7	7.46e+02			
Full	9	7.44e+02	1.68e+00	2	4.31e-01

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S9
Semantic judgment task (including omissions): interaction effect (single trial data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			2.43e+00	2.24e+00	2.63e+00
s2-s1			-3.17e-01	-1.03e+00	3.94e-01
p2-p1			1.68e+00	1.09e+00	2.27e+00
p3-p2			1.78e+00	1.19e+00	2.36e+00
s:p2-p1*			-1.41e-01	-5.57e-01	2.75e-01
s:p3-p2*			3.88e-01	-1.95e-01	9.70e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	2.16e-01				
Subject	(34 levels)				
	s2-s1				
s2-s1	8.24e-01				
Subject	(34 levels)				
	p2-p1				
p2-p1	1.46e+00				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	7	3.19e+03			
Full	9	3.19e+03	1.67e+00	2	4.34e-01

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S10
Semantic judgment task (RT): nested effect (aggregated data).

Family:	gaussian	(identity)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			3.68e-01	3.45e-01	3.92e-01
s2-s1			2.82e-03	-1.82e-02	2.38e-02
s1p2-p1*			-1.36e-01	-1.58e-01	-1.14e-01
s1p3-p2*			-3.11e-02	-4.32e-02	-1.90e-02
s2p2-p1*			-1.34e-01	-1.58e-01	-1.10e-01
s2p3-p2*			-2.06e-02	-3.27e-02	-8.45e-03
Random effects					
Subject	(34 levels)				
	intercept				
intercept	4.64e-03				
Subject	(34 levels)				
	s1p2-p1				
s1p2-p1	2.96e-03				
Subject	(34 levels)				
	s2p2-p1				
s2p2-p1	3.76e-03				
Residual var					2.53e-02
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	6	-5.36e+02			
Full	10	-6.9e+02	1.54e+02	4	2.71e-32

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S11
Semantic judgment task (RT): interaction effect (aggregated data).

Family:	gaussian	(identity)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			3.68e-01	3.45e-01	3.92e-01
s2-s1			2.82e-03	-1.92e-02	2.48e-02
p2-p1			-1.36e-01	-3.13e-01	-2.26e-01
p3-p2			-3.11e-02	-7.04e-02	-3.30e-02
s:p2-p1*			-1.34e-01	-1.15e-02	1.59e-02
s:p3-p2*			-2.06e-02	-3.19e-03	2.42e-02
Random effects					
Subject	(34 levels)				
	intercept				
intercept	4.64e-03				
Subject	(34 levels)				
	s2-s1				
s2-s1	2.96e-03				
Subject	(34 levels)				
	p2-p1				
p2-p1	3.76e-03				
Subject	(34 levels)				
	p3-p2				
p3-p2	3.76e-03				
Residual var					2.53e-02
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	9	-7.36e+02			
Full	11	-7.4e+02	3.72e+00	2	1.55e-01

Note. s- sequence; p- position; * test predictor; random effects present var-cov matrix.

Table S12
Threat conditioning (expectancy rating): effect of condition (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			-1.39e+00	-1.74e+00	-1.04e+00
CS*			5.45e+00	4.67e+00	6.24e+00
category			1.16e-01	-5.32e-01	7.64e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	3.56e-01				
Subject	(34 levels)				
	CS				
CS	2.16e+00				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	4	3.37e+02			
Full	5	2.57e+02	8e+01	1	3.67e-19

Note. CS- condition; * test predictor; random effects present var-cov matrix.

Table S13
Threat conditioning (expectancy rating): effect of condition (single trial data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			-1.79e+00	-2.35e+00	-1.24e+00
CS*			6.13e+00	4.95e+00	7.30e+00
category			-1.30e-01	-6.32e-01	3.72e-01
Random effects					
Subject	(34 levels)				
	intercept	CS			
intercept	6.76e-01	-1.38e+00			
CS	-1.38e+00	3.96e+00			
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	5	1.15e+03			
Full	6	1.09e+03	6.32e+01	1	1.86e-15

Note. CS- condition; * test predictor; random effects present var-cov matrix.

Table S14
Threat conditioning (skin conductance): effect of condition (aggregated data).

Family:	gaussian	(identity)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			3.24e-01	2.89e-01	3.58e-01
CS*			9.40e-02	6.37e-02	1.24e-01
category			-1.30e-02	-4.34e-02	1.74e-02
Random effects					
Subject	(34 levels)				
	intercept				
intercept	8.23e-03				
Residual var					6.26e-02
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	4	-1.02e+02			
Full	5	-1.28e+02	2.56e+01	1	4.13e-07

Note. CS- condition; * test predictor; random effects present var-cov matrix.

Table S15
Recognition memory (pre-conditioned categories): interaction effect (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			9.17e-01	8.32e-01	1.00e+00
position			2.64e-01	-7.77e-04	5.29e-01
sequence			-1.63e-01	-4.20e-01	9.46e-02
position:sequence*			2.49e-01	3.52e-02	4.64e-01
cat1-cat2			-3.30e-02	-2.51e-01	1.85e-01
cat1-cat3			5.66e-02	-1.61e-01	2.74e-01
cat1-cat4			1.24e-01	-8.94e-02	3.37e-01
Random effects					
Subject	(34 levels)				
	intercept	position	sequence		
intercept	3.98e-02	-8.5e-03	3.16e-02		
position	-8.5e-03	2.1e-01	2.77e-02		
sequence	3.16e-02	2.77e-02	1.76e-01		
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	12	7.62e+02			
Full	13	7.56e+02	5.2e+00	1	2.26e-02

Note. cat- category; * test predictor; random effects present var-cov matrix.

Table S16
Recognition memory (pre-conditioned categories): interaction effect (single trial data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			9.46e-01	8.41e-01	1.05e+00
position			2.75e-01	5.95e-02	4.91e-01
sequence			-1.81e-01	-3.97e-01	3.40e-02
position:sequence*			2.50e-01	3.51e-02	4.65e-01
cat1-cat2			7.48e-03	-2.60e-01	2.75e-01
cat1-cat3			2.20e-02	-2.46e-01	2.90e-01
cat1-cat4			9.77e-02	-1.71e-01	3.66e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	4.290012e-02				
Item	(300 levels)				
	intercept				
intercept	1.6e-01				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	8	8.08e+03			
Full	9	8.07e+03	5.17e+00	1	2.3e-02

Note. cat- category; * test predictor; random effects present var-cov matrix.

Table S17
Recognition memory (pre-conditioned categories): interaction effect (d-prime).

Family:	gaussian	(identity)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			1.28e+00	1.17e+00	1.38e+00
position			3.24e-01	-5.52e-03	6.54e-01
sequence			-2.35e-01	-5.64e-01	9.48e-02
position:sequence*			3.73e-01	4.39e-02	7.02e-01
cat1-cat2			-3.15e-02	-3.16e-01	2.53e-01
cat1-cat3			1.17e-02	-2.74e-01	2.97e-01
cat1-cat4			1.16e-01	-1.69e-01	4.00e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	4.02e-02				
Residual var					4.82e-01
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	8	2.1e+02			
Full	9	2.05e+02	4.91e+00	1	2.67e-02

Note. cat- category; * test predictor; random effects present var-cov matrix.

Table S18
Recognition memory (conditioned categories): effect of condition (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			1.57e+00	1.45e+00	1.69e+00
CS*			1.27e-02	-1.68e-01	1.93e-01
category			4.39e-01	2.58e-01	6.21e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	5.34e-02				
Subject	(34 levels)				
	CS				
CS	6.84e-03				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	4	3.44e+02			
Full	5	3.44e+02	1.9e-02	1	8.9e-01

Note. CS- condition; * test predictor; random effects present var-cov matrix.

Table S19
Recognition memory (conditioned categories): effect of condition (single trial data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			1.63e+00	1.48e+00	1.79e+00
CS*			1.34e-02	-1.72e-01	1.98e-01
category			4.55e-01	1.99e-01	7.12e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	5.86e-02				
Subject	(34 levels)				
	CS				
CS	1.98e-01				
Item	(300 levels)				
	intercept				
intercept	5.86e-02				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	5	3.12e+03			
Full	6	3.12e+03	1.98e-02	1	8.88e-01

Note. CS- condition; * test predictor; random effects present var-cov matrix.

Table S20
Recognition memory (conditioned categories): effect of condition (d-prime).

Family:	gaussian	(identity)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			2.09e+00	1.95e+00	2.22e+00
CS*			3.38e-02	-1.79e-01	2.47e-01
category			4.67e-01	2.54e-01	6.80e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	6.37e-02				
Residual var					4.39e-01
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	4	9.84e+01			
Full	5	9.83e+01	1e-01	1	7.52e-01

Note. CS- condition; * test predictor; random effects present var-cov matrix.

Table S21
Recognition memory (full dataset): interaction effect (aggregated data).

Family:	binomial	(logit)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			1.13e+00	1.05e+00	1.21e+00
cs			5.07e-03	-1.73e-01	1.83e-01
phase			1.28e+00	6.06e-01	1.95e+00
position			2.65e-01	1.40e-03	5.28e-01
sequence			-1.75e-01	-4.31e-01	8.01e-02
position:sequence*			2.47e-01	3.42e-02	4.60e-01
cat1-cat2			-2.40e-01	-4.75e-01	-5.43e-03
cat1-cat4			-1.63e-01	-3.94e-01	6.80e-02
cat1-cat5			-1.01e-01	-3.31e-01	1.30e-01
cat3-cat6			8.77e-01	5.19e-01	1.23e+00
Random effects					
Subject	(34 levels)				
	intercept				
intercept	3.91e-02				
Subject	(34 levels)				
	phase				
phase	3.6e-01				
Subject	(34 levels)				
	position				
position	2.07e-01				
Subject	(34 levels)				
	sequence				
sequence	1.72e-01				
Subject	(34 levels)				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	13	1.1e+03			
Full	14	1.1e+03	5.17e+00	1	2.3e-02

Note. cs- condition; cat- category; * test predictor; random effects present var-cov matrix.

Table S22
Recognition memory (high-confidence): effect of position (d-prime).

Family:	gaussian	(identity)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			1.96e+00	1.79e+00	2.13e+00
position*			5.44e-01	1.11e-01	9.77e-01
sequence			-5.46e-02	-4.41e-01	3.32e-01
position:sequence			1.38e-01	-2.48e-01	5.25e-01
cat1-cat2			2.17e-02	-3.19e-01	3.62e-01
cat1-cat3			4.10e-04	-3.48e-01	3.48e-01
cat1-cat4			-2.08e-02	-3.68e-01	3.26e-01
Random effects					
Subject	(34 levels)				
	intercept	position			
intercept	1.62e-01	-6.21e-02			
position	-6.21e-02	3.04e-01			
Residual var					4.2e-01
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	8	2.86e+02			
Full	9	2.8e+02	5.48e+00	1	1.92e-02

Note. cs- condition; cat- category; * test predictor; random effects present var-cov matrix.

Table S23
Recognition memory (high-confidence): effect of condition (d-prime).

Family:	gaussian	(identity)			
Fixed effects					
Term			β	95% CI [LL]	95% CI [UL]
Intercept (grand mean)			2.82e+00	2.64e+00	3.00e+00
CS*			-1.44e-01	-3.98e-01	1.09e-01
category			4.15e-01	1.62e-01	6.69e-01
Random effects					
Subject	(34 levels)				
	intercept				
intercept	1.45e-01				
Residual var					4.14e-01
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	4	1.32e+02			
Full	5	1.31e+02	1.27e+00	1	2.59e-01

Note. cs- condition; * test predictor; random effects present var-cov matrix.

Supplemental Experiment 1: Selecting stimuli from the initial pool of images

The initial pool of stimuli consisted of 62 images per category (372 images in total). To obtain a final pool of images of six categories (i.e., "tools", "fruits & vegetables", "furniture", "vehicles", "clothes", and "music instruments") such that they are maximally similar with respect to their recognition memory, we ran a pilot study. To capture any category- and item-level bias in memory, we tested the recognition memory 15 min after an incidental encoding in one experimental session.

Methods

Participants. Sixteen participants (7 female) between 19 and 34 years old (mode=27) with normal or corrected-to-normal vision were recruited from the general population in the city of Leipzig, Germany using MPI-CBS participant database.

Behavioral tasks.

Semantic judgment task. Participants were presented with images on a gray background for 4 s with a 6 $\pm$ 2 s variable inter-trial interval (cf., Dunsmoor et al., 2015). During the stimulus presentation, names of the six categories were visible under an image and participants were instructed to make a semantic decision to indicate a category to which the presented exemplar in the image belongs best. Participants did not receive any feedback about the accuracy of their response. Presentation of images was pseudo-randomize such that there were no more than two images of the same category presented in a row.

Recognition memory test. The test consisted of 372 images in total – for each category there were 31 old images and 31 new category-related foils that were never shown before to control for false alarm rate. Old vs new assignment of images was counterbalanced across participants, such that half of the participants was exposed to the same set of old images, which were conversely new in the other half of participants. The images were presented one at a time and the task was self-paced. On each trial, the image appeared on the screen for 1.5 s. Then, a 6-point Likert scale ("very sure old", "sure old", "maybe old", "maybe new", "sure new", "very sure new") appeared additionally under the picture to prompt the participant to indicate his or her memory response. The trial order was

pseudo-randomized to ensure that participants did not encounter a long string of old or new images in a row.

Procedure. The experimental session (~90 min) took place on one day and involved an incidental encoding and a surprise recognition memory test after a 15-min break. Stimulus presentations were generated with the Psychophysics Toolbox (Brainard, 1997) for MATLAB 2016a (the MathWorks). Participants viewed the stimuli from a distance of approx. 60 cm on a 22" display (Ilyama Vision Master Pro 514, resolution 1280 x 1024, aspect ratio 5:4, refresh rate 80Hz) in a dimly lit room. Having provided their written informed consent, participants were exposed to a semantic judgment task that served as a cover task for incidental encoding of 31 images per category (186 images in total). Next, participants watched a video (BBC Planet Earth) for about 12 min followed by a recognition memory test. At the end, participants were debriefed and asked to verbally indicate whether some images (old or new) might have been in their opinion problematic to categorize; for example, because they might be assigned to two categories or might appear being rather atypical for a given category.

Statistical analyses. To recover how easy an image was to be recognized, that is, an old image is identified as "old" and a new item is identified as "new", we used generalized linear mixed models (GLMMs) with a *logit* link function and binomial error structure. To keep the number of responses equal across participants, the "old"/"new" responses were collapsed across the three confidence levels.

First, to test for the effect of category on the correct responses, we fitted a GLMM to the full dataset including a test predictor of a category treated as a factor with six levels. The model included a random intercept nested within a subject (16 levels).

Second, to recover individual contribution of each and every category-exemplars to the overall memory performance for each category, we fitted a separate GLMM to each category including a fixed intercept and a random structure comprising an intercept nested within participants and an intercept nested within an item. The items were then sorted with respect to their individual intercept and excluded based on how much they added to the

overall category-level recognition score. Additionally, images presenting objects difficult to perceptually distinguish from one another, objects that may have been considered as belonging to more than one semantic category, or objects being rather atypical for a category were excluded first. For example, "cimbalom", "clavichord", and "celesta" were rather similar within the category of "music instruments". Likewise, "swim goggle" and "sunglasses" might have been considered as accessories rather than a piece of clothing.

Results

Overall, participants performed well in the task (~88%). The GLMM revealed an effect of category (likelihood ratio test: $\chi^2(5) = 37.47$, $p < .001$, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.05$). The category of "furniture" and "clothes" exhibited high memory scores (log-odds estimate $\pm$ SE $= 2.19 \pm 0.16$ and log-odds estimate $\pm$ SE $= 2.18 \pm 0.16$, respectively), while the category of "fruits & vegetables" and "music instruments" exhibited very low memory scores (log-odds estimate $\pm$ SE $= 1.69 \pm 0.15$ and log-odds estimate $\pm$ SE $= 1.56 \pm 0.15$, respectively).

12 images per category were excluded to arrive at the final sample of 50 images per category. The ratio of items excluded due to their intrinsic properties to items exhibiting high or low addition to the category-level mean based on GLMM was the following: (a) 8:4 for "furniture", (b) 8:4 for "vehicles", (c) 0:12 for "fruits & vegetables", (d) 3:9 for "tools", (e) 7:5 for "music instruments, and (f) 4:8 for "clothes".

Conclusion

By excluding images with atypical memorability or ambiguous category membership, we minimized intrinsic differences in recognition memory across categories. This selection procedure yielded a final stimulus set of 50 images per category (300 in total).

Supplemental Experiment 2: Selecting four categories for a pre-conditioning phase based on 24-hour recognition memory

To minimize the influence of intrinsic differences in 24-hour recognition memory across the four preconditioned categories, which could overshadow the predicted interaction between relational knowledge and Pavlovian threat conditioning, we selected four (out of six) categories with the most similar recognition memory. To this end, we ran a pilot study where exemplars (i.e., images) of all six categories were incidentally encoded in the same way as in the pre-conditioning phase of the main study.

Methods

Participants. Fourteen participants (8 female) between 19 and 29 years old (mode=27) with normal or corrected-to-normal vision were recruited from the general population in the city of Leipzig, Germany using MPI-CBS participant database. Eligibility criteria excluded pregnancy, current or a history of a neurological, psychiatric or drug-abuse disorder. Further eligibility criteria excluded use of medication (except for oral contraceptives and ibuprofen/ paracetamol) and excessive alcohol consumption 24 hours before the experiment.

Behavioral tasks.

Semantic judgment task. The task resembled the structure of the sequence learning task used in the main study with the exception that the presentation of category exemplars was random, i.e., a systematic order among semantic categories within a sequence was absent. As in the main study, images were shown in three-item sequences. Each stimulus presentation began with two category names (2 s), followed by a fixation cross (0.5 s) and then the image (2.5 s). Participants had 1 s to respond during the image presentation, which was terminated when the two options disappeared from the screen. Participants received visual feedback to their button press which vanished at the image offset, but no information about the accuracy of their response was presented. Image presentations were spaced by 0.5 s interval within a sequence and sequences were spaced by a 6 $\pm$ 2 s variable inter-trial interval.

Recognition memory test. The test consisted of 300 images in total – for each category there were 30 old images and 20 new category-related foils to control for false alarm rate. Old vs new assignment of images was randomized across participants. The images were presented one at a time and the task was self-paced. On each trial, the image appeared on the screen for 1.5 s. Then, "definitely old", "maybe old", "maybe new" and "definitely new" response options appeared additionally under the picture to prompt the participant to indicate his or her memory response. The trial order was pseudo-randomized to ensure that participants did not encounter a long string of old or new images in a row.

Procedure. Two experimental sessions (~35 min each) took place on two consecutive days in the same test room and around the same time of day. Stimulus presentations were generated with the Psychophysics Toolbox (Brainard, 1997) for MATLAB 2016a (the MathWorks). Participants viewed the stimuli from a distance of approx. 60 cm on a 22" display (Ilyama Vision Master Pro 514, resolution 1280 x 1024, aspect ratio 5:4, refresh rate 80Hz) in a dimly lit room.

In the first session (~30 min), having provided their written informed consent, participants were exposed to basic-level exemplars (i..e, images) of all six categories and instructed to do a semantic judgment task that served as a cover task for incidental encoding of 30 images per category (180 images in total). In the second session (~30 min), participants did a surprise recognition memory test that had the same structure as the test in the main study.

Statistical analyses. All reported means are accompanied by 95% confidence interval obtained with parametric bootstrapping (n = 5000). To test each of our hypotheses, we specified a statistical model and a null model that is reduced by excluding a test predictor from the full model. The models were then compared using a likelihood-ratio test.

To test whether the recognition memory differed across the six semantic categories, we calculated d-prime index per category that took into account unequal number of old vs new images (30:20), which differed from the main study (25:25). Hit rate was calculated as a total number of correct "definitely old" and "maybe old" responses divided by the the total

number of old items of the respective category. Likewise, false alarm rate was calculated as a total number of incorrect "definitely old" and "maybe old" responses divided by the the total number of new items of the respective category. Next, the inverse of the standard normal cumulative distribution function was evaluated at the probability values of P(hit) and P(false alarm) that were then subtracted from one another. If hit or false alarm rates were equal to one or zero, the scores were replaced with 1–1/(2 * number of responses) and 1/(2 * number of responses), respectively, where responses refer to the total number of old and new items from a respective category. We then fitted a LMM with an identity link function and a Gaussian error structure to test a fixed effect of a 6-level factor of category on the d-prime. The contrast matrix was composed using a standard dummy coding (i.e., treatment contrast). To account for the data non-independence, we included a random intercept nested within a subject (14 levels). To verify the results of the d-prime index, we additionally expressed the memory recognition on the original scale prior to the z-transformation, that is, as P(hit) - P(false alarm).

To additionally explore whether a position in which category-exemplars were presented in a triplet during the semantic judgment task affected reaction time (RT) and accuracy judgment in the absence of any underlying temporal structure among categories, we performed the same analysis procedure as in the main study. Specifically, we fitted a generalized linear mixed model (GLMM) with a *logit* link function and binomial error structure to investigate the fixed effect of a 3-level factor of position on the performance in the task. We used sum contrast to test successive difference between positions (i.e., 2-1, and 3-2). To account for the data non-independence, we included a random intercept nested within a subject (14 levels). The first model tested for the effect of position on the number of correct responses out of total number of valid trials, i.e., trials in which a response was given within the 1-s time limit. We additionally re-fitted the model to the number of correct responses out of total number of trials including omissions, i.e, trials in which a response was absent within the pre-specified time limit. In this model, omissions were treated as valid but incorrect responses. Responses that occurred before the image onset were treated as invalid, i.e., they were treated as NAs. To account for the varying number of valid trials

across participants, we included weights in both models that corresponded to the number of valid trials per condition. To verify the results, we fitted two other analogous GLMMs to binary responses in a single trial. For the RT, we fitted a linear mixed model (LMM) with an identity link function and a Gaussian error structure to investigate the effect of a fixed 3-level factor of position on the average RT to correct trials. To again account for the data non-independence, we included a random intercept nested within a subject in the model (14 levels).

Results

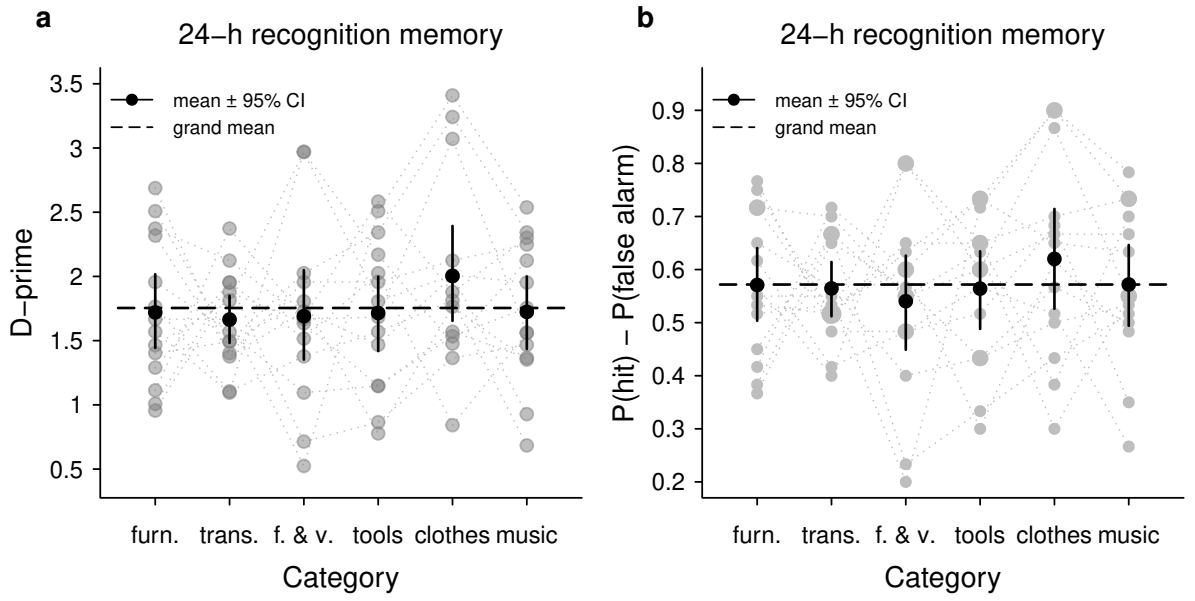

Figure S2. Baseline 24-hour recognition memory of six semantic categories (i.e., "furniture", "transportation", "fruits & vegetables", "tools", "clothes", and "music instruments") incidentally encoded during a semantic judgment task. **a)** Recognition memory expressed as d-prime was similar across all categories (likelihood ratio test: $\chi^2(5) = 3.93$, p =.557). **b)** Corrected recognition score calculated as P(hit) - P(false alarm) revealed a similar profile (likelihood ratio test: $\chi^2(5) = 2.68$, p = .749, marginal-$R^2 = 0.03$, conditional-$R^2 = 0.17$). Size of the circles is proportional to a number of participants with the same data point.

24-hour recognition memory at baseline of each category is illustrated in the Figure S2a as the d-prime index and in the Figure S2b as corrected recognition (P[hit]-P[false alarm]). The LMM revealed no effect of category on recognition memory regardless whether it was indexed by dprime (likelihood ratio test: $\chi^2(5) = 3.94$, p = .559, marginal-$R^2 = 0.04$, conditional-$R^2 = 0.21$) or corrected recognition (likelihood ratio test: $\chi^2(5) = 2.68$, p = .749, marginal-$R^2 = 0.03$, conditional-$R^2 = 0.17$). Numerical comparison of the means revealed that the category "clothes" had the highest mean (2.00 [1.64, 2.41]) while the category "transport" had the lowest mean (1.67 [1.49, 1.86]). However, numerical

comparison on the original scale P(hit) - P(false alarm) revealed that the category "fruits & vegetables" had the lowest mean (0.56 [0.51, 0.62]). "Fruits & vegetables" also exhibited second largest variability (0.54 [0.45, 0.63]) after the category "clothes" (0.62 [0.52, 0.72]).

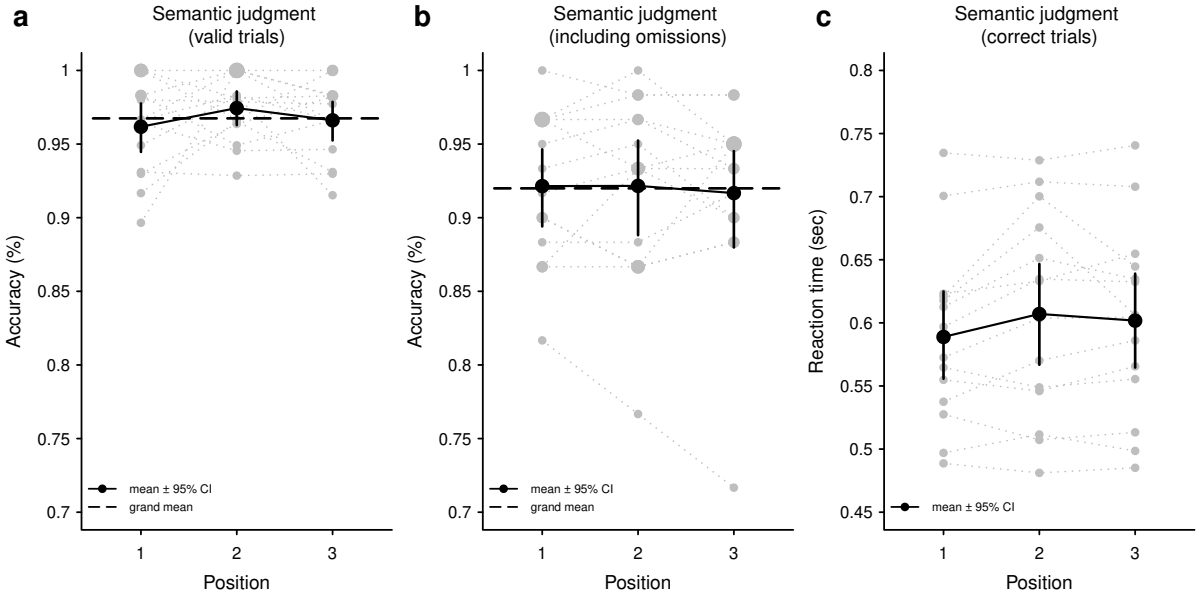

Figure S3. Response accuracy and reaction times during the semantic judgment task when a systematic order among semantic categories within a sequence is absent. **a)** Proportion of correct responses to the total number of valid trials and **b)** to the total number of trials including omissions is similar across their position in a sequence (likelihood ratio test: $\chi^2(2) = 2.34$, p = .310, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.09$ for the first model comparison and likelihood ratio test: $\chi^2(2) = 0.18$, p = .915, marginal-$R^2 = 0.00$, conditional-$R^2 = 0.13$ for the second model comparison). **c)** Average reaction time to correct trials is facilitated for the trials in the first position (likelihood ratio test: $\chi^2(2) = 7.09$, p = .029, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.94$). Size of the circles is proportional to a number of participants with the same data point.

Accuracy scores and RT of the the semantic judgment task are illustrated in the Figure S3. We observed no effect of position for the accuracy in neither of the two GLMMs (likelihood ratio test: $\chi^2(2) = 2.34$, p = .310, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.09$ for the first model comparison and likelihood ratio test: $\chi^2(2) = 0.18$, p = .915, marginal-$R^2 = 0.00$, conditional-$R^2 = 0.13$ for the second model comparison). Overall, participants were highly accurate in their responses for valid trials (log-odds estimate $\pm$ SE $= 3.54 \pm 0.20$) as well as when omissions treated as incorrect responses were included in the total number of trials (log-odds estimate $\pm$ SE $= 2.64 \pm 0.21$). Yet, both GLMMs exhibited tendency for underdispersion (0.69 and 0.94, respectively). Therefore, to verify their results, we ran two additional GLMMs fitted to individual binary responses. They revealed the same results (likelihood ratio test: $\chi^2(2) = 2.34$, p = .310, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.09$ for the first model comparison and likelihood ratio test: $\chi^2(2) = 0.18$, p = .915, marginal-$R^2 = 0.00$, conditional-$R^2 = 0.13$ for the second model comparison) but their

dispersion parameters were less underdispersed (0.91 and 0.94, respectively).

We observed an effect of position in the LMM fitted to the average RT (likelihood ratio test: $\chi^2(2) = 7.09$, p $= .029$, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.94$). Specifically, participants responded slower to the trials in the second compared to the first position in a triplet (estimate $\pm$ SE $= 0.018 \pm 0.007$), and with no difference in the second and third positions (estimate $\pm$ SE $=$ -0.005 $\pm$ 0.007).

Discussion

While we did not identify statistically significant differences in recognition memory across the six categories, we observed that the category "fruits & vegetables" and "clothes" exhibited high response variance and one of the biggest numerical difference. This left the "furniture", "tools", "vehicles", and "music instruments" as the four most similar categories to one another based on their 24-hour recognition memory.

Additionally, we identified that the semantic judgment task was very easy for participants as they exhibited ceiling performance and was independent of a trial position in a sequence. Surprisingly, participants were faster in their responses for the trials initiating the sequence of stimulus presentations, perhaps indicating an increased motivation to respond within a time limit at the onset of a sequence. These results are in contrast to the main study where participants exhibited facilitated performance for the second and third position in a sequence, indicating an online use of acquired relational knowledge, i.e., when a systematic order among semantic categories within a sequence was present.

Supplemental Experiment 3: Replicating memory prioritization by Pavlovian threat conditioning with non-repeating items

To replicate the sole effect of memory prioritization by Pavlovian threat conditioning when the overall demand on episodic memory is increased compared to a standard category-based Pavlovian conditioning task (de Voogd et al., 2016; Dunsmoor, Kragel, Martin, & LaBar, 2014; e.g., Dunsmoor et al., 2012, 2015; Kalbe & Schwabe, 2020), we ran a pilot study where an incidental encoding was extended by a pre-conditioning phase including four categories (i.e., "furniture", "tools", "transport", and "music instruments"). The conditioning phase followed the standard category-based Pavlovian threat conditioning with two categories (i.e., "fruits & vegetables" and "clothes") without repeating any of category exemplars.

Methods

Participants. We recruited healthy right- and left-handed volunteers between 18 and 35 years old with normal or corrected-to-normal vision from the general population in the city of Leipzig, Germany using MPI-CBS participant database. Eligibility criteria excluded pregnancy, current or a history of a neurological, psychiatric or drug-abuse disorder. Further eligibility criteria excluded use of medication (except for oral contraceptives and pain killers) and excessive alcohol consumption 24 hours before the experiment.

Twenty volunteers provided their written informed consent prior to the start of the study. One participant did not show up for the second day. One discontinued the experiment after the shock intensity level calibration on day 1. This led to the final sample of 18 participants (8 female) between 21 and 31 years old (mode = 23).

Behavioral tasks.

Semantic judgment task. The semantic judgment task served as a cover task for incidental encoding of 100 images of four categories (25 per each) selected for a pre-conditioning phase in the main study (i.e., "furniture", "tools", "vehicles", and "music instruments"). The task resembled the structure of the semantic judgment task used in the main study with the exception that the order of stimulus presentations was random, i.e., a

systematic order among semantic categories within a sequence of trials was absent, and a sequence consisted of four instead of three images due to the total number of stimuli. As in the main study, each stimulus presentation started with appearance of two category names at the bottom of the screen (2 s), followed by a fixation cross (0.5 s) and the presentation of the image (2.5 s). Participants had 1 s to respond during the image presentation, which terminated when the two options disappeared from the screen. Participants received visual feedback to their button press which vanished at the image offset, but no information about the accuracy of their response. Image presentations were spaced by 0.5 s interval within a sequence and sequences were spaced by a 6 $\pm$ 2 s variable interval.

Category-based threat conditioning task. Participants were exposed to basic-level exemplars (i.e., images) of two semantic categories (i.e., "fruits & vegetables" and "clothes") to learn through experience that one category (CS+) is predictive of unconditioned stimulus (US) – a mild electric shock to the fingers delivered via a constant current stimulator (Digitimer DS7A, Digitimer Ltd., the UK) – while another category (CS-) was never paired with shocks. The US consisted of a 200-ms train of 40 square pulses with individual pulse width of 0.2 ms at frequency of 200 Hz. The allocation of a category and condition (CS+/CS-) was counter-balanced across participants. There were 25 presentations of the CS+, 13 of which co-terminated with the US (52% reinforcement rate), and 25 presentations of the CS-. Pictures were presented for 4.0 s with a variable inter-trial interval of 7-11 s (the average ITI was set to 9 s). The order of the different trial types was pseudo-randomized so that no more than 3 trials of the same type could occur in a row. To facilitate learning, the first trial was always the CS+ that was reinforced with the US. Participants were told that when there was an image on the screen, they may receive a shock and that there was a relationship between the shocks and the categories. Participants were not explicitly instructed the conditioned category–unconditioned stimulus contingency but they were instructed to figure out the contingency through experience. They were also asked to provide a binary response whether they expected receiving a shock (yes or no) in every trial, but were instructed that their button presses did not affect the outcome on a trial to mitigate the potential to attribute the outcome to their choice or reaction time.

Surprised recognition memory test. The surprise recognition memory test consisted of 300 images in total – for each category there were 25 old images and 25 new category-matched foil images to control for false alarm rate (Dunsmoor et al., 2018, 2012, 2015). Images were randomly allocated as old or new across participants. The images were presented one at a time and the task was self-paced. On each trial, the image appeared on the screen for 1.5 s. Then, "definitely old", "maybe old", "maybe new" and "definitely new" response options appeared additionally under the picture to prompt the participant to indicate old-new judgment. The trial order was pseudo-randomized to ensure that participants did not encounter a long string of old or new images in a row. Additionally, the order was pseudo-randomized in such a way that none of the previously experienced sequences of images occurred during the memory test.

Procedure. Two experimental sessions took place on two consecutive days in the same test room and around the same time of day. Stimulus presentations were generated with the Psychophysics Toolbox (Brainard, 1997) for MATLAB 2016a (the MathWorks). Participants viewed the stimuli from a distance of approx. 60 cm on a 22" display (Ilyama Vision Master Pro 514, resolution 1280 x 1024, aspect ratio 5:4, refresh rate 80Hz) in a dimly lit room.

The first session (~60 min) consisted of a pre-conditioning phase (i.e., when the risk of receiving the unconditioned stimulus was absent) and a conditioning phase. In the pre-conditioning phase participants were exposed to a semantic judgment task. Afterwards, the session proceeded with the attachment of the electrodes to administer electric shocks (~10 min) in the same way as in the main study. Shock electrodes were attached to ring and pinky fingers of the right hand and the shock intensity level was calibrated using an ascending staircase procedure starting with a low voltage (near a perceptible threshold) to reach a level deemed "maximally uncomfortable without being painful" by the participant. Shock intensity was subsequently scored by the participant on a modified pain assessment scale from 0 (no sensation) to 10 (very high intensity), and ranged from 6 to 9 (mode = 8) in the current sample. To form Pavlovian threat memory, participants were subsequently exposed to the category-based differential delay threat conditioning task (Dunsmoor et al.,

2014, 2012, 2015) while concurrent measurements of their skin conductance were being collected. The procedure of the task was the same as in the main study with the exception that all images used in the conditioning were seen only once by participants. After the conditioning task, the shock electrodes were removed and the measurements of skin conductance were stopped. At the conclusion of the session, participants were asked to rate the intensity of the shock felt during the study on a scale from 0 (not at all unpleasant) to 10 (extremely unpleasant), and how much fear they felt during the experiment from 0 (not at all afraid) to 10 (extremely afraid). They were also asked to estimate how many shocks they received throughout the whole study (not counting shocks received during the calibration procedure) and report their understanding of what the contingency between the categories and shocks was that was rated on a 3-point scale by an experimenter (0-not aware; 1-sort of aware; 2-aware).

In the second session (~60 min), participants were first screened whether they consumed alcohol or used medication in the past 24 hours. Next, participants started with a surprise memory recognition test that followed the same procedure as in the main study. To assess whether participants expected the surprise memory test, they were asked whether they have had any expectations for the experiment just prior to the memory test ("Do you have any expectations of what the next task might be about: yes or no?"). Participants were then told that there would be a test of their memory for the pictures they saw earlier, and were asked to indicate on a 5-point scale how surprised they were by a memory test, from 1 ("I did not expect a memory test at all") to 5 ["Yes, I knew there would be a memory test"; Dunsmoor et al. (2015)]. In the current sample, the surprise was rated from 1 to 4 (mode = 1).

At the conclusion of the session, participants were asked to fill out four self-report questionnaires: (a) the trait inventory of the State-Trait Anxiety Inventory-STAI (Laux et al., 1981; Spielberger et al., 1983); (b) Childhood Trauma Questionnaire-CTQ (Bernstein et al., 2003; Klinitzke et al., 2012); (c) Intolerance of Uncertainty Scale-IUS (Buhr & Dugas, 2002; Gerlach et al., 2008); and (d) Berkman-Syme Social Network Index-SNI (Berkman &

Syme, 1979).

Skin conductance acquisition and response estimation. Skin conductance was collected at a sampling rate of 500 Hz using BIOPAC MP35 (Biopac Systems Inc., CA, USA) with disposable snap Ag/AgCl electrodes. Skin conductance pre-processing and response estimation was done in the same way as in the main study. Continuous raw skin conductance time series were analyzed in the window starting from 10 s prior to the onset of the first stimulus and terminating 16 s after the offset of the last stimulus. The raw time series was filtered with a 1st order butterworth high-pass filter at 0.05 Hz and smoothed with a low-pass moving average filter at 5 Hz. To avoid any remaining high-frequency fluctuations, the time series was down-sampled at 10 Hz and interpolated using linear interpolation.

SCR scoring was performed given the following criteria: the trough-to-peak deflection occurs within 0.5-4.0 s following CS onset and lasts between 0.5 and 5.0 s (or until the rise of the skin conductance due to the US onset for the CS+ trials co-terminating with the US). If an SCR did not meet these criteria, then the trial was scored as a zero. The scoring procedure was done semi-automatically by the Autonomate software (Green, Kragel, Fecteau, & LaBar, 2014). The quality of individual response identification was visually inspected and manually adjusted if required (e.g., when a response to the shock was misidentified as a conditioned response or the response did not clearly follow the shape of the canonical skin conductance response). Raw SCR values were scaled by the maximum SCR evoked by the shock, and the SCR magnitude was calculated as the mean value across condition-specific trials. The first trial was always a CS+, which co-terminated with the US, and therefore was excluded from analyses to account for orienting responses. SCR values were square-root transformed for group analysis.

Statistical analyses. To test each of our hypotheses, we specified a statistical model and a null model that was reduced by excluding a test predictor from the full model. The models were then compared using a likelihood-ratio test.

First, to test whether participants acquired Pavlovian conditioned threat responses, we specified statistical models that quantified the effect of condition (CS+ vs. CS-) on shock

expectancy ratings and square-root transformed SCR magnitude. To test the fixed effect of a two-level factor of condition on the number of responses indicating "yes" (i.e., 1) out of total number of trials per condition (i.e., 25), we fitted a GLMM with a *logit* link function and binomial error structure including an additional control predictor of category identity. Missing responses were treated as "no" (i.e., 0). To verify the results of the model which exhibited a large overdispersion parameter, we fitted an additional GLMM with a *logit* link function and binomial error structure to binary responses in a single trial. To test the fixed effect of a two-level factor of condition on square-root transformed SCR magnitude, we fitted a LMM with an identity link function and a Gaussian error structure including an additional control predictor of category identity. To account for data non-independence, we included a random intercept nested within a subject (18 levels). For the models fitted to single-trial data we additionally included a random effect of condition nested within a subject. For the GLMM model to converge, we excluded the random intercept and kept the random slope. All models used "sum contrast" (coded with 0.5).

Second, to analyze data of the recognition memory task, we aggregated the number of correct response to the total number of responses in the task per category, which is expressed as proportion of correct responses, i.e., when an old item is identified as "old" and a new item is identified as "new" for the respective semantic category. To keep the number of responses equal across participants, the "old"/"new" responses were collapsed across the two confidence levels. This way the proportion of correct responses corresponds to (P[hit] + (1-P[false alarm]))/ 2 in a signal detection theory framework. Given that the number of old items was equal to the number of new items, the proportion of correct responses was equivalent to the corrected recognition score, expressed as P(hit) - P(false alarm) – the correlation of the two indices is equal to 1 (Figure S4).

To test whether Pavlovian conditioning prioritized memory for the CS+ category compared with the CS- category, we fitted a GLMM with a *logit* link function and binomial error structure to the data of two categories presented in the conditioning phase (i.e., "fruits & vegetables" and "clothes"). We included in the model a control predictor of category

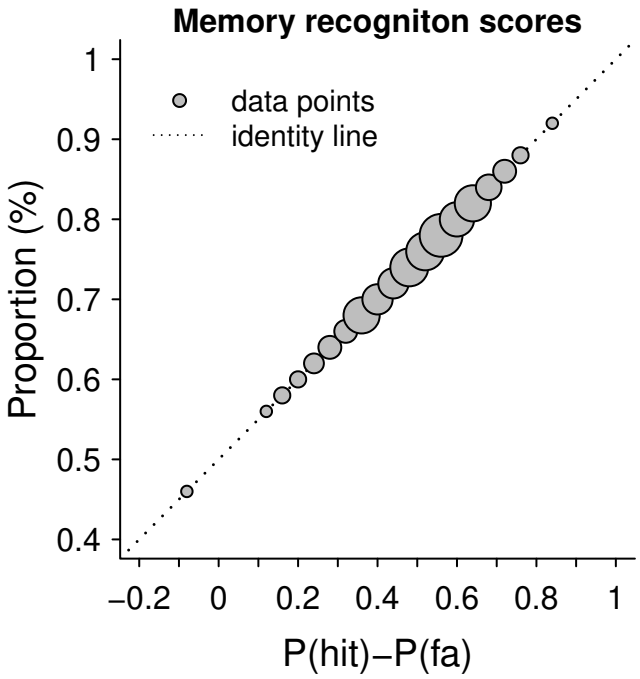

Figure S4. Linear relation between corrected recognition calculated as P(hit) - P(false alarm) and proportion of correct responses (percent correct) that can be expressed in terms of signal detection theory as (P[hit] + (1-P[false alarm]))/ 2. Data points come from the recognition memory test performed in the current sample. Size of the circles is proportional to a number of participants with the same data point.

identity and a random intercept nested within a subject (18 levels). We also fitted a GLMM with a *logit* link function and binomial error structure to single trial data where response was coded as either 1 or 0.

To verify the results of the binomial model, we expressed the recognition memory as d-prime index. Hit rate was calculated as a total number of correct "definitely old" and "maybe old" responses divided by the the total number of old items of the respective category. Likewise, false alarm rate was calculated as a total number of incorrect "definitely old" and "maybe old" responses divided by the the total number of new items of the respective category. Next, the inverse of the standard normal cumulative distribution function was evaluated at the probability values of P(hit) and P(false alarm) that were then subtracted from one another. If hit or false alarm rates were equal to one or zero, the scores were replaced with 1–1/(2 * number of responses) and 1/(2 * number of responses), respectively, where responses refer to the total number of old and new items from a respective category. We then fitted a LMM with an identity link function and a Gaussian error structure to test the fixed effect of the two-level factor of condition (CS+ vs. CS-) on

the d-prime. The model included a control predictor of category identity and a random slope nested within a subject (18 levels) since the model including a random intercept failed to converge.

To explore whether 24-hour recognition memory differed across the four non-conditioned categories, we expressed the recognition memory score as the proportion of correct responses and d-prime index, computed in the same way as for the previous analysis, and fitted analogous (G)LMMs with a corresponding link function (logit/identity) and error distribution (binomial/Gaussian) with a test predictor, i.e., a 4-level factor of category coded as a sum contrast using the first category ("furniture") as a reference. Random intercept was nested within a subject. The GLMM fitted to binary responses extended the random effects structure by including one of the slopes (without correlation with the intercept).

To additionally explore whether a position in which category exemplars were presented during the semantic judgment task affected reaction time (RT) and accuracy judgment in the absence of any underlying temporal structure among categories, we performed an analogous analysis procedure as in the previous pilot study and the main study. Specifically, we fitted a GLMM with a *logit* link function and binomial error structure to investigate the fixed effect of a 4-level factor of position on the number of correct responses out of total number of valid trials, i.e., trials in which a response was given within the 1-s time limit. We additionally re-fitted the model to the number of correct responses out of total number of trials including omissions, i.e, trials in which a response was absent within the pre-specified time limit. In this model, omissions were treated as valid but incorrect responses. Responses that occurred before the image onset were treated as invalid, i.e., they were treated as NAs. To account for the varying number of valid trials across participants, we included weights in both models that corresponded to the number of valid trials per condition. To verify the results, we fitted two other GLMMs to binary responses to a single trial. For the RT, we fitted a linear mixed model (LMM) with an identity link function and a Gaussian error structure to investigate the effect of a fixed 4-level factor of position on the average RT to correct trials. To again account for the data non-independence, we included a random

intercept nested within a subject in the model (18 levels). All these models coded 4-level factor of position using a sum contrast with the first position treated as a reference.

Results

Shock expectancy ratings and SCR are illustrated in the Figure S5. The GLMM revealed an expected effect of condition on a number of responses indicating shock expectancy (likelihood ratio test: $\chi^2(1) = 710.78$, $p < .001$, marginal-$R^2 = 0.91$, conditional-$R^2 = 0.99$). Shock expectancy was increased in the CS+ condition compared with the CS- condition (log-odds estimate $\pm$ SE $= 5.99 \pm 0.43$). While the model revealed a large overdispersion parameter (2.77), its results were corroborated by the additional GLMM fitted to individual binary responses (likelihood ratio test: $\chi^2(1) = 33.52$, $p < .001$, marginal-$R^2 = 0.68$, conditional-$R^2 = 0.78$), which dispersion parameter was within a reasonable range (0.92). Similarly, the LMM revealed an expected effect of condition on SCR (likelihood ratio test: $\chi^2(1) = 17.99$, $p = < .001$, marginal-$R^2 = 0.22$, conditional-$R^2 = 0.77$). SCR was increased in the CS+ condition compared with the CS- condition (estimate $\pm$ SE $= 0.10 \pm 0.02$). Taken together, participants exhibited acquisition of conditioned threat responses in both modalities, as expected.

Results of the 24-hour recognition memory test for non-conditioned and conditioned categories are illustrated in the Figure S6a as percent correct and in the Figure S6b as the d-prime index. Conditioned categories exhibited numerical differences in the proportion of correct responses (CS+ = 0.76 [0.73, 0.78]; CS- 0.69 [0.65, 0.73]) as well as d-prime (CS+ = 1.62 [1.50, 1.74]; CS- 1.23 [0.96, 1.48]). The GLMM revealed the expected effect of conditioning on recognition memory (likelihood ratio test: $\chi^2(1) = 8.99$, $p = .003$, marginal-$R^2 = 0.25$, conditional-$R^2 = 0.46$, $\phi = 0.99$). Probability of a correct response in the recognition memory test was increased for CS+ items compared with the CS- items (log-odds estimate $\pm$ SE $= 0.32 \pm 0.11$). The results were corroborated with the LMM fitted to d-prime (likelihood ratio test: $\chi^2(1) = 5.91$, $p = .015$, marginal-$R^2 = 0.17$, conditional-$R^2 = 0.20$, estimate $\pm$ SE $= 0.39 \pm 0.15$).

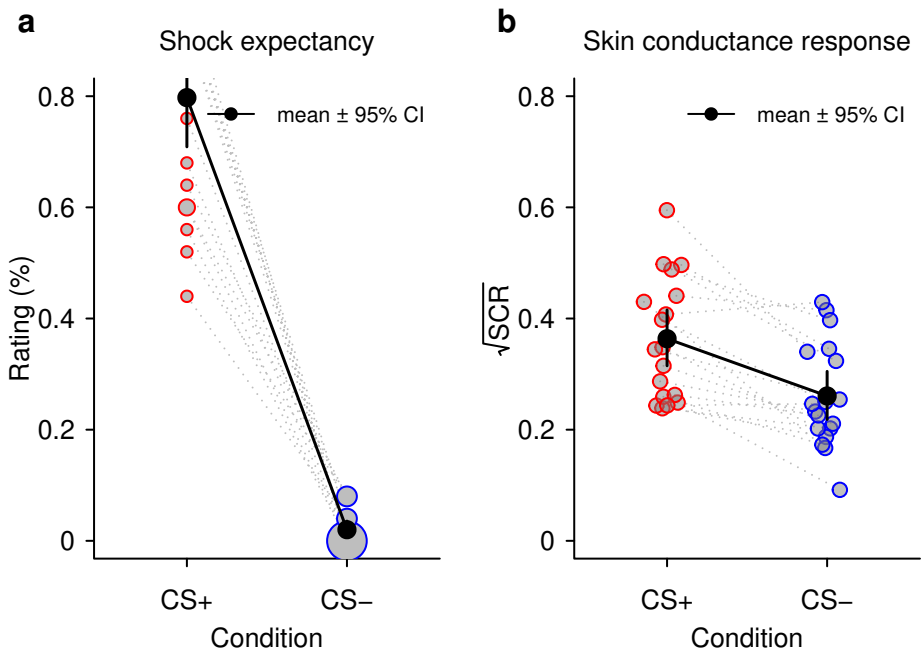

Figure S5. Shock expectancy and skin conductance response (SCR) during Pavlovian threat conditioning. **a)** Proportion of responses indicating shock expectancy revealed an expected effect of condition (likelihood ratio test: $\chi^2(1) = 710.78$, $p < .001$, marginal-$R^2 = 0.91$, conditional-$R^2 = 0.99$). Size of the circles is proportional to a number of participants with the same data point. **b)** Similarly, magnitude of SCR differed between the conditions (likelihood ratio test: $\chi^2(1) = 17.99$, $p < .001$, marginal-$R^2 = 0.22$, conditional-$R^2 = 0.77$).

We did not observe any significant difference in the recognition memory across non-conditioned categories when analysing proportion of correct responses (furniture = 0.76 [0.73, 0.79]; transportation = 0.76 [0.73, 0.79]; tools = 0.76 [0.73, 0.79]; music instruments = 0.73 [0.69, 0.77]) nor the d-prime index (furniture = 1.65 [1.46, 1.82]; transportation = 1.61 [1.38, 1.83]; tools = 1.54 [1.34, 1.75]; music instruments = 1.35 [1.07, 1.64]). Results of the GLMM (likelihood ratio test: $\chi^2(3) = 3.10$, $p = .377$, marginal-$R^2 = 0.04$, conditional-$R^2 = 0.43$, $\phi = 0.97$) were similar to the results of the LMM on d-prime (likelihood ratio test: $\chi^2(3) = 4.67$, $p = .197$, marginal-$R^2 = 0.05$, conditional-$R^2 = 0.25$). (G)LMMs fitted to the whole dataset of the memory recognition showed no differences between the two phases of the test, i.e., pre-conditioning vs conditioning in the proportion of correct responses (likelihood ratio test: $\chi^2(1) = 3.29$, $p = .070$, marginal-$R^2 = 0.14$, conditional-$R^2 = 0.47$, $\phi = 0.95$) nor the d-prime index (likelihood ratio test: $\chi^2(1) = 1.31$, $p = .253$, marginal-$R^2 = 0.10$, conditional-$R^2 = 0.24$).

Accuracy scores and RT of the the semantic judgment task are illustrated in the Figure S7. We observed no effect of position on the accuracy for valid trials (likelihood ratio

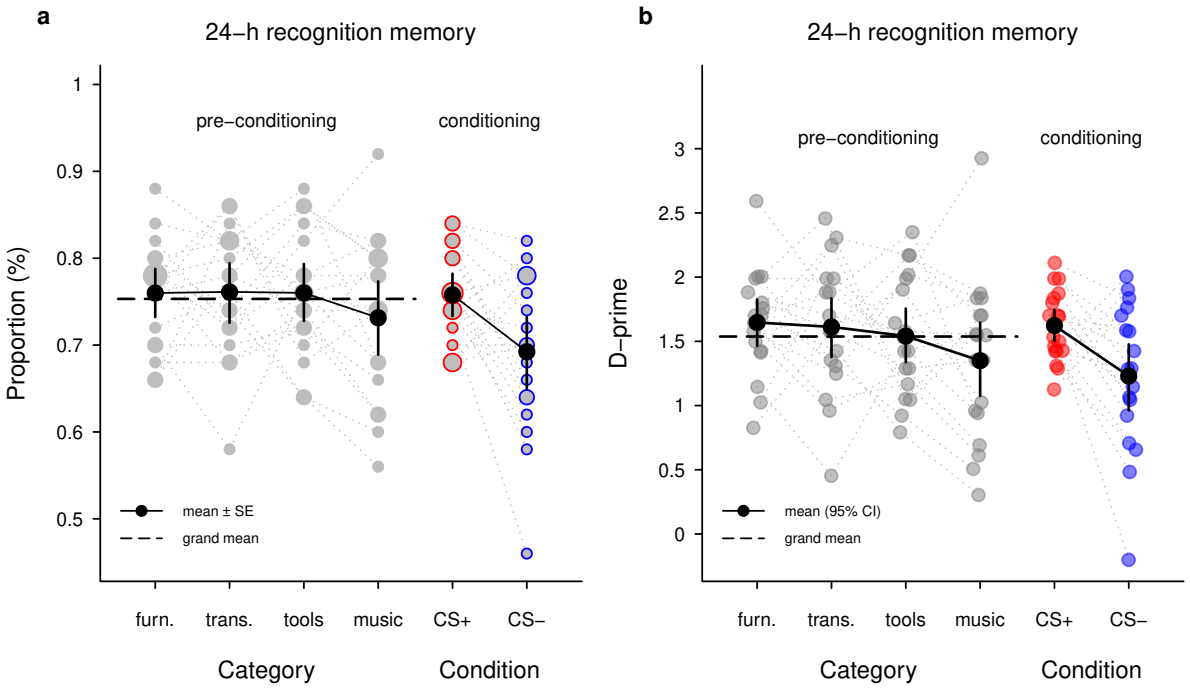

Figure S6. 24-hour recognition memory for non-conditioned and conditioned categories incidentally encoded during a semantic judgment task (pre-conditioning phase) and Pavlovian threat conditioning (conditioning phase). **a)** Probability of correctly recognizing items of conditioned categories differed between conditions (likelihood ratio test: $\chi^2(1) = 8.99$, p = .003, marginal-$R^2 = 0.25$, conditional-$R^2 = 0.46$) while it was similar across all four non-conditioned categories (likelihood ratio test: $\chi^2(3) = 3.10$, p = .377, marginal-$R^2 = 0.04$, conditional-$R^2 = 0.43$). Size of the circles is proportional to a number of participants with the same data point. **b)** Recognition memory expressed as d-prime revealed the same pattern in the conditioned (likelihood ratio test: $\chi^2(1) = 5.91$, p = .015, marginal-$R^2 = 0.17$, conditional-$R^2 = 0.20$) and non-conditioned categories (likelihood ratio test: $\chi^2(3) = 4.67$, p = .197, marginal-$R^2 = 0.05$, conditional-$R^2 = 0.25$).

test: $\chi^2(3) = 3.30$, p = .347, marginal-$R^2 = 0.02$, conditional-$R^2 = 0.11$, $\phi = 0.92$) exhibiting ceiling performance (log-odds estimate ± SE = 3.40 ± 0.21), but on the accuracy when omissions were included in the total number of trials (likelihood ratio test: $\chi^2(3) = 8.42$, p = .038, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.11$, $\phi = 0.86$). For these trials, accuracy was highest in the first position compared with the other positions (P2-P1: log-odds estimate ± SE = -0.47 ± 0.21; P3-P1: log-odds estimate ± SE = -0.45 ± 0.21; P4-P1: log-odds estimate ± SE = -0.56 ± 0.21). The additional two GLMMs fitted to individual binary responses revealed the same pattern of results (likelihood ratio test: $\chi^2(3) = 5.24$, p = .155, marginal-$R^2 = 0.02$, conditional-$R^2 = 0.11$, $\phi = 0.89$ for the first model and likelihood ratio test: $\chi^2(3) = 8.89$, p = .031, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.11$, $\phi = 0.95$ for the second model). We observed an effect of position in the LMM fitted to the average RT (likelihood ratio test: $\chi^2(3) = 11.71$, p = .008, marginal-$R^2 = 0.04$, conditional-$R^2 = 0.80$). LMM revealed that participants responded slower to the trials in the second compared to the first position (log-odds estimate ± SE = 0.03 ± 0.01), while

response to the third and fourth positions were rather similar (P3-P1: log-odds estimate ± SE = 0.02 ± 0.01; P4-P1: log-odds estimate ± SE = 0.00 ± 0.01).

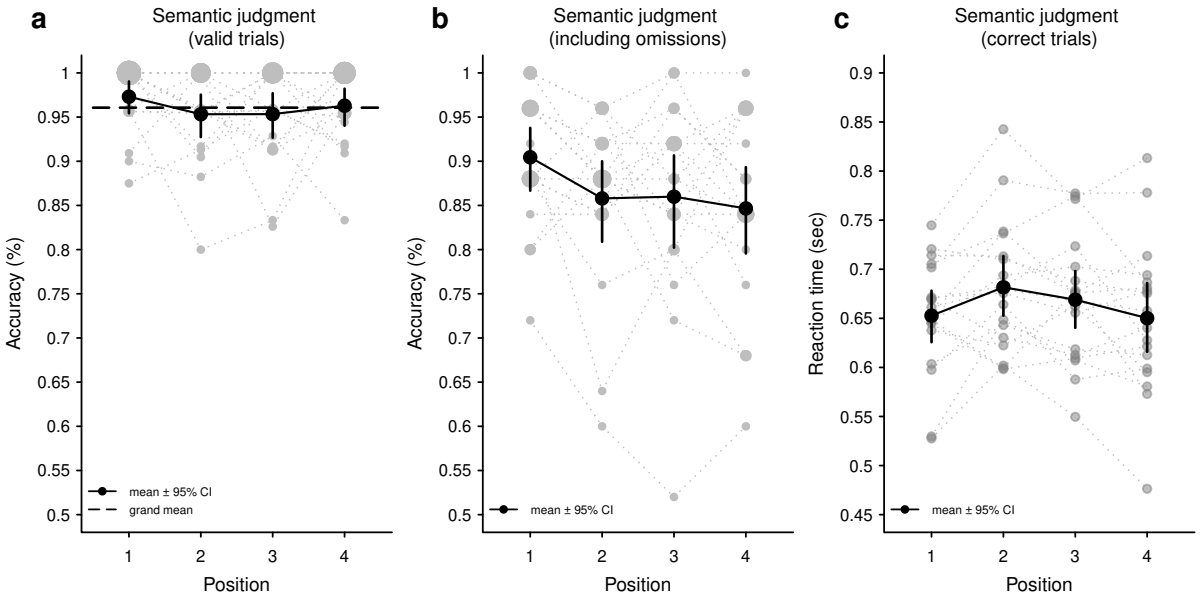

Figure S7. Response accuracy and reaction times during the semantic judgment task when a systematic order among semantic categories within a sequence is absent. **a)** Proportion of correct responses to the total number of valid trials is similar across their position in a sequence (likelihood ratio test: $\chi^2(3) = 3.30$, p = .347, marginal-$R^2 = 0.02$, conditional-$R^2 = 0.11$). **b)** However, it is better for the trials in the first position when their proportion is calculated to the total number of trials including omissions (likelihood ratio test: $\chi^2(3) = 8.42$, p = .038, marginal-$R^2 = 0.01$, conditional-$R^2 = 0.11$). **c)** Average reaction time to correct trials is facilitated for the trials in the first position (likelihood ratio test: $\chi^2(3) = 11.71$, p = .008, marginal-$R^2 = 0.04$, conditional-$R^2 = 0.80$). Size of the circles is proportional to a number of participants with the same data point.

Discussion

First, we replicated the effect of Pavlovian threat conditioning on episodic memory when the CS+ and CS- items were presented only once, which lead to increased recognition of the CS+ items compared with the CS- items (e.g., de Voogd et al., 2016; Dunsmoor et al., 2014, 2018, 2012, 2015; Dunsmoor & Kroes, 2019; Kroes, Dunsmoor, Lin, Evans, & Phelps, 2017). This effect was replicated when extra material was incidentally encoded before conditioning increasing the overall demand on episodic memory. Interestingly, the overall recognition memory for the neutral information in the pre-conditioning was as good as the memory for CS+ items, suggesting that CS+ items are prioritized over the CS- items but their memory is not enhanced as such.

Second, we replicated our previous finding from the supplemental experiment 2 showing that the four categories presented in the pre-conditioning phase exhibited similar 24-hour recognition memory to one another. This indicates that threat conditioning has no

effect on episodic memory of pre-conditioned categories if they are not embedded in a relational structure including conditioned categories.

Lastly, we observed a similar pattern of behavior during the semantic judgment task as previously found in the supplementary experiment 2. When a systematic order among semantic categories was absent, participants tended to be faster and slightly more accurate in their responses for the trials initiating the sequence of stimulus presentations. This contrasts with the main study, in which participants showed facilitated performance at the second and third positions in a sequence, suggesting that the predictable order of categories facilitated performance in the semantic judgment task.

References

Berkman, L. F., & Syme, S. L. (1979). Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents. *American Journal of Epidemiology*, *109*(2), 186–204. https://doi.org/10.1093/oxfordjournals.aje.a112674

Bernstein, D. P., Fink, L., Handelsman, L., Foote, J., Lovejoy, M., Wenzel, K., … Ruggiero, J. (1994). Initial reliability and validity of a new retrospective measure of child abuse and neglect. *The American Journal of Psychiatry*, *151*(8), 1132–1136. https://doi.org/10.1176/ajp.151.8.1132

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., … Zule, W. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse & Neglect*, *27*(2), 169–190. https://doi.org/10.1016/s0145-2134(02)00541-0

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Buhr, K., & Dugas, M. J. (2002). The Intolerance of Uncertainty Scale: Psychometric properties of the English version. *Behaviour Research and Therapy*, *40*(8), 931–945. https://doi.org/10.1016/s0005-7967(01)00092-4

de Voogd, L. D., Fernández, G., & Hermans, E. J. (2016). Awake reactivation of emotional memory traces through hippocampal-neocortical interactions. *NeuroImage*, *134*, 563–572. https://doi.org/10.1016/j.neuroimage.2016.04.026

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*(4), 447–456. https://doi.org/10.1016/j.jml.2007.11.004

Dunsmoor, J. E., Kragel, P. A., Martin, A., & LaBar, K. S. (2014). Aversive learning modulates cortical representations of object categories. *Cerebral Cortex (New York, N.Y.: 1991)*, *24*(11), 2859–2872. https://doi.org/10.1093/cercor/bht138

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., Kroes, M. C. W., Moscatelli, C. M., Evans, M. D., Davachi, L., & Phelps, E. A. (2018). Event segmentation protects emotional memories from competing experiences encoded close in time. *Nature Human Behaviour*, *2*(4), 291–299. https://doi.org/10.1038/s41562-018-0317-4

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

Dunsmoor, J. E., & Paz, R. (2015). Fear Generalization and Anxiety: Behavioral and Neural Mechanisms. *Biological Psychiatry*, *78*(5), 336–343. https://doi.org/10.1016/j.biopsych.2015.04.010

Gerlach, A. L., Andor, T., & Patzelt, J. (2008). Die Bedeutung von Unsicherheitsintoleranz für die Generalisierte Angststörung Modellüberlegungen und Entwicklung einer deutschen Version der Unsicherheitsintoleranz-Skala. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *37*(3), 190–199. https://doi.org/10.1026/1616-3443.37.3.190

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology*, *91*(3), 186–193. https://doi.org/10.1016/j.ijpsycho.2013.10.015

Hsieh, L.-T., Gruber, M. J., Jenkins, L. J., & Ranganath, C. (2014). Hippocampal activity patterns carry information about objects in temporal context. *Neuron*, *81*(5), 1165–1178. https://doi.org/10.1016/j.neuron.2014.01.015

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

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. https://doi.org/10.1037/xlm0000728

Klinitzke, G., Romppel, M., Häuser, W., Brähler, E., & Glaesmer, H. (2012). [The German Version of the Childhood Trauma Questionnaire (CTQ): psychometric characteristics in a representative sample of the general population]. *Psychotherapie, Psychosomatik, Medizinische Psychologie*, *62*(2), 47–51. https://doi.org/10.1055/s-0031-1295495

Kroes, M. C. W., Dunsmoor, J. E., Lin, Q., Evans, M., & Phelps, E. A. (2017). A reminder before extinction strengthens episodic memory via reconsolidation but fails to disrupt generalized threat responses. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/s41598-017-10682-7

Laux, L., Glanzmann, P., Schaffner, P., & Spielberger, C. D. (1981). *Das State-Trait-Angstinventar*.

Neath, I., & Surprenant, A. M. (2003). *Human memory: An introduction to research, data, and theory*. Belmont, CA: Wadsworth.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800