

Memory specificity in evaluation conditioning: a ‘Who said what’ approach

Jéréline Benading¹ & Klaus Rothermund²

¹ Université Agnostique de Louvain-la-Jena

² Friedrich Schiller University Jena

Author Note

Correspondence concerning this article should be addressed to Jéréline Benading, .
E-mail: karoline.bading@uni-jena.de

Abstract

BLABLABLABLABLA

Keywords: keywords

Word count: X

Memory specificity in evaluation conditioning: a ‘Who said what’ approach

Please insert: introduction leading up to one the greatest studies of all time.

Amazing paraphraph

Please insert: a great paragraph

Another amazing paraphraph

Please insert: a great paragraph

Methods

We report how we determined our sample size, all data exclusions, all manipulations, and all measures. The preregistration, experiment program, data, and analyses are publicly available on the Open Science Framework at: <https://osf.io/rqkvy/>.

The study project was approved by the ethics committee of [*Institution Redacted*] (*ID Redacted*).

Participants and design

The experiment used a 2 (US valence: Positive vs. Negative vs. None) x 2 (Task order: Evaluation first vs. Memory first) mixed design, with US valence manipulated within participants and Task order manipulated between participants.

We recruited 172 participants (50% female; $M_{age} = 39.59$; $SD_{age} = 13.23$), our targeted sample size, online on Prolific. Participants were English speakers, had an approval rate of at least 90% and at least 100 previous submissions, and did not take part in previous evaluative conditioning studies we conducted on Prolific. Data were unavailable for one participant, and we excluded 5 participants declaring that they did not pay

attention or did not take their responses seriously. This resulted in a final sample size of 166 participants ($n = 70$ in the Evaluation first condition and $n = 96$ in the Memory first condition).

To determine sample size, we set α to .05, and we aimed for a statistical power of at least 80% to detect an Evaluative Conditioning effect as small as Cohen’s $d = 0.2$ in a one-sided paired sample t -test (IV: US valence; DV: evaluative ratings). An analysis with the R package *pwr* (version 1.3-0; Champely, 2020) showed that we needed 156 participants. A sample size of $N = 156$ also provided 80 power to detect correlations $rs \geq .22$ (e.g., between parameter estimates and evaluative conditioning scores). To avoid a final sample smaller than the targeted sample size after applying the exclusion criteria, we increased this estimate by 10%, resulting in 172 participants.

Materials and procedure

We programmed the experiment with lab.js (Henninger et al., 2022). We exported the study to an HTTPS-protected website with JATOS (Lange et al., 2015).

Stimuli. As neutral stimuli, we used 54 5- to 7-letter nonwords (all made of a combination of vowels and consonants; e.g., ‘*botsy*’; ‘*ampfong*’) used in a previous study (Stahl & Bading, 2020). On a participant basis, 24 nonwords were used as CSs (paired with USs in the learning phase) and 24 were new (i.e., presented only in the test phase).

The USs were 24 color images of animals (e.g., a cockroach), scenes (e.g., a rainbow), and objects (e.g., a knife) selected from the Open Affective Standardized Image Set (OASIS; Kurdi et al., 2017). Based on OASIS ratings (on a 7-point Likert scale), twelve images were positive ($M_{valence} = 5.88$; $SD_{valence} = 0.24$; $M_{arousal} = 4.10$; $SD_{arousal} = 0.50$) and 12 were negative ($M_{valence} = 2.05$; $SD_{valence} = 0.32$; $M_{arousal} = 4.27$; $SD_{arousal} = 0.52$). Positive USs were significantly more positive than negative USs, Welch’s $t(20.23) = 33.36, p < .001$. Positive and negative USs did not significantly differ in arousal,

Welch's $t(21.96) = 0.82, p = .419$.

Learning phase. After providing their informed consent, participants entered the learning phase. Participants were told that they would see pairs made of one nonword (the CSs; on the center-left of the screen) and one image (the USs on the center-right of the screen). Participants had to pay close attention to each pair.

Twenty-four CS-US pairs were presented three times each in a random order (block-wise) for 1000 milliseconds (separated by an inter-trial interval of 1000 ms), resulting in 72 trials. For each participant, the 24 CSs were randomly drawn from the pool of 54 nonwords. Each CS was paired with one specific US. The CSs were displayed in a sans-serif font (font-size: 40). The dimensions of the USs were 250 pixels (width) and 200 pixels (height).

Test phase. After the learning phase, participants entered the test phase. Participants performed two tasks: an evaluative rating task and a memory task. Participants were randomly assigned to one of the two Task order conditions. In the Evaluation first condition, participants performed the evaluative rating task and then the memory task. The order was reversed in the Memory first condition: participants performed the memory task and then the evaluative rating task. The exact wording of the tasks slightly differed as a function of Task order.

Evaluative rating task. In the evaluative classification task, the 24 CSs presented in the learning phase and 24 new nonwords (randomly drawn from the remaining pool of 30 nonwords) were displayed individually once in a random order without time limit. Participants rated how positive or negative they found the nonwords on a 8-point Likert scale ranging from 1 “very negative” to 8 “very positive.” The 48 trials were separated by 500-ms inter-trial intervals.

Memory task. In the memory task, the 24 CSs presented in the learning phase and 24 new nonwords (identical to the ones displayed in the evaluative rating task) were

displayed individually once in a random order without time limit.

Each of the trials began with a recognition memory task. Participants were asked whether the nonwords were part of the pairs presented in the learning phase (response options: “Yes (old)”; “No (new)”).

If participants responded “No (new),” the next recognition memory trial began after a 500-ms inter-trial interval. If participants responded “Yes (old),” a new screen including the same nonword appeared after a 500-ms blank screen, and participants had to select the specific image that the nonword was paired with. Participants were instructed to click on the specific image if they remembered it or to guess the correct image if they could not remember the exact image. Eight images (all from the learning phase) were displayed in two rows of four images. For nonwords that were correctly recognized (hits), the correct US was presented with 7 randomly selected distractors (3 images of the same valence as the correct US; 4 images of the opposite valence). All the USs were allocated to a random position. For nonwords that were incorrectly recognized (false alarms), 8 USs, half positive and half negative, were randomly selected and allocated to a random position.

Check measures. After the test phase, we administered an attention check and a seriousness check. In the attention check, we asked participants whether they paid attention to the nonwords and images presented throughout the study (Yes/No response). Participants were told that their response would not affect their payment.

In the seriousness check based on Aust et al. (2013), participants read:

“It would be very helpful if you could tell us at this point whether you have taken the requested responses seriously, so that we can use your answers for our scientific analysis, or whether you were just clicking through to take a look at the survey? (again, this will not affect your payment).”

The response options were *“I have taken the requested responses seriously”* and *“I have just clicked through, please discard my data”*. We used the answers to the attention

and seriousness checks as exclusion criteria (see the Participants and design section).

Participants then had the chance to comment on the study. Finally, participants were thanked and debriefed.

Data processing and analyses

BLABLABLA

Results¹

Preregistered analyses on evaluative ratings

We report the analyses we conducted on evaluative ratings. We averaged evaluative ratings as a function of US valence (Positive, Negative, None) and Task order (Evaluation first, Memory first). For each participant, we calculated an Evaluative Conditioning (EC) score, which is their mean evaluative rating on CSs paired with positive USs minus their mean evaluative rating on CSs paired with negative USs (negative scores indicate higher evaluations on negatively paired vs. positively paired CSs; positive scores indicate higher evaluations on positively vs. negatively paired CSs).

First, we conducted a between-participants ANOVA on EC scores with Task order as the only factor. We tested whether the grand mean was above 0 by calculating the F-test of the intercept. A grand mean above 0 would indicate that, overall, we replicated the EC effect. In line with the preregistration, we divided the p -value of this test by two to perform a one-tailed test, as the grand mean of EC scores was above 0 ($M = 0.72$; $SD = 0.97$). The F -test was significant, $F(1, 164) = 92.25$, $p < .001$, $\hat{\eta}_G^2 = .360$, 90% CI [.267, .444], showing that we replicated the EC effect. The effect of Task order was not

¹ All analyses were conducted with R (R Core Team, 2021). We used the packages XXX (), XXX (), XXX (), XXX (), XXX (), XXX (), XXX ()....

significant², $F(1, 164) = 0.94$, $p = .334$, $\hat{\eta}_G^2 = .006$, 90% CI [.000, .040], which means that performing the evaluative rating task before or after the memory task did not significantly change the EC effect.

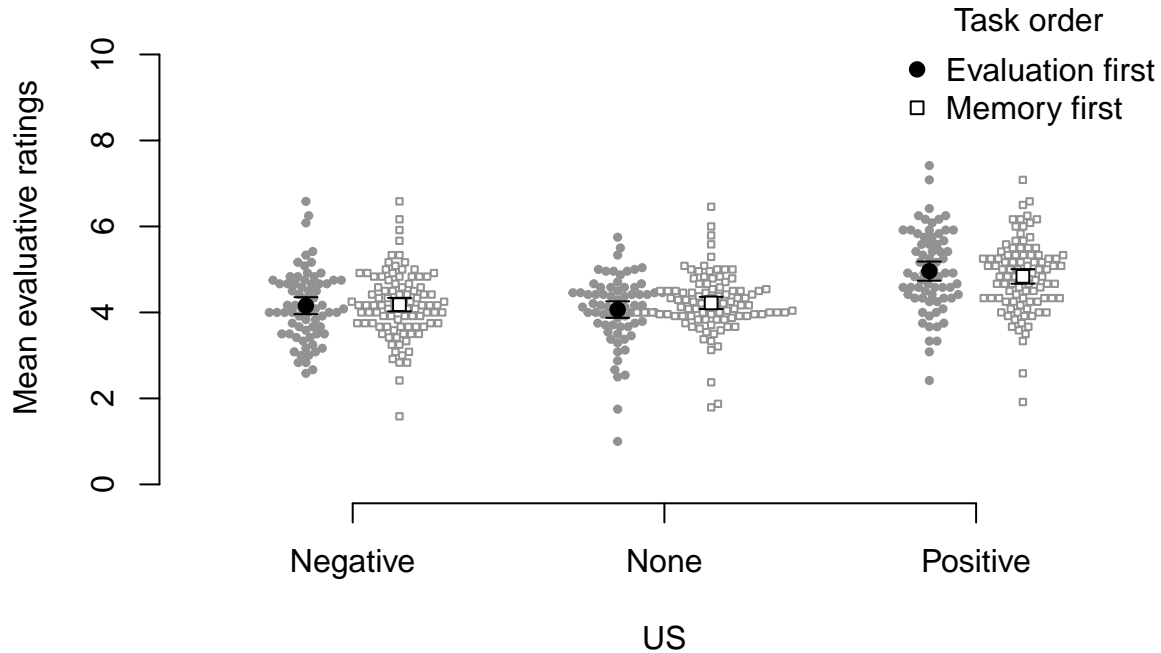


Figure 1. (#fig:plot_us_valence_order) Mean evaluative ratings (and 95% error bars) as a function of US valence and Task order.

Complementarily, we also conducted a 3 (US valence) x 2 (Task order) mixed ANOVA on evaluative ratings (see Figure @ref(fig:plot_us_valence_order)). Different from the ANOVA above, evaluative ratings on unpaired nonwords can be compared with

² As preregistered, we divided the p -value by two to perform a one-tailed version of the test (similar to t -tests) because EC scores are descriptively larger in the Evaluation first ($M = 0.80$; $SD = 0.87$) than Memory first ($M = 0.66$; $SD = 1.03$) condition. EC scores were above 0 both in the Evaluation first condition, $t(69) = 7.76$, $p < .001$, $d = 0.93$, 90% CI = [0.69, 1.16] and in the Memory first condition, $t(95) = 6.22$, $p < .001$, $d = 0.63$, 90% CI = [0.45, 0.82].

evaluations in other conditions. The main effect of US valence was significant, $F(2, 328) = 68.83, p < .001, \hat{\eta}_G^2 = .156, 90\% \text{ CI } [.099, .214]$. We followed-up on the ANOVA by conducting multiple comparisons (Bonferroni-corrected) based on the full model: evaluative ratings were higher for positively paired CSs compared with new non-words, $t(164) = -9.94, p < .001, d = 0.75, 95\% \text{ CI } = [0.58, 0.92]$, and compared with negatively-paired CSs, $t(164) = -9.61, p < .001, d = 0.74, 95\% \text{ CI } = [0.57, 0.91]$. Evaluative ratings were not significantly different for negatively-paired CSs and for nonwords, $t(164) = -0.39, p = .922, d = 0.02, 95\% \text{ CI } = [-0.13, 0.17]$. The main effect of Task order was not significant, $F(1, 164) = 0.04, p = .844, \hat{\eta}_G^2 = .000, 90\% \text{ CI } [.000, .010]$, nor was the interaction between US valence and Task order, $F(2, 328) = 1.81, p = .166, \hat{\eta}_G^2 = .005, 90\% \text{ CI } [.000, .021]$.

Preregistered analyses on memory performance

XXX

Preregistered analyses on evaluations as a function of MPT parameter estimates

XXX

MPT model

Amazing result 2

Relationships between ratings and parameters

Amazing result 3

General discussion

Amazing paragraph

smart conclusion 1

Another amazing paragraph

smart conclusion 2

Appendix

Amazing paragraph

additional information on something amazing

Another amazing paragraph

more information on another amazing thing