Memory specificity in evaluation conditioning: a 'Who said what' approach

Jéréline Benading[1] & Klaus Rothermund[2]

[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

KB:

- introduce EC, explain importance of memory: recognizing CS, recollecting US valence and/or US identity
- introduce current methodological approaches and describe problems (measurement error, biases, . . .) –> need for formal model
- introduce who-said-what model as candidate for estimating recognition, US valence recollection, US identity recollection (without measurement error) and separating memory from response biases
- describe advantages for experimental and correlative research
- introduce present research: develop procedure, fit model, validate parameters via correlations

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

Results[1]

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

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

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, $F(1, 164) = 92.25$, $p < .001$, $\hat{\eta}^2_G = .360$, 90% CI [.267, .444], showing that we replicated the EC effect. The effect of Task order was not significant[2], $F(1, 164) = 0.94$, $p = .334$, $\hat{\eta}^2_G = .006$, 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.

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 evaluations in other conditions. The main effect of US valence was significant, $F(2, 328) = 68.83$, $p < .001$, $\hat{\eta}^2_G = .156$, 90% 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, $t(164) = -9.94, p < .001, d = 0.75, 95\%$ CI $= [0.58, 0.92]$, and compared with negatively-paired CSs, $t(164) = -9.61, p < .001, d = 0.74, 95\%$ CI $= [0.57, 0.91]$. Evaluative ratings were not significantly different for negatively-paired CSs and for nonwords, $t(164) = -0.39, p = .922, d = 0.02, 95\%$ CI $= [-0.13, 0.17]$. The main effect of Task order was not significant, $F(1, 164) = 0.04$, $p = .844$, $\hat{\eta}^2_G = .000$, 90% CI [.000, .010], nor was the interaction between US valence and Task order, $F(2, 328) = 1.81$, $p = .166$, $\hat{\eta}^2_G = .005$, 90% CI [.000, .021].

---

[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].

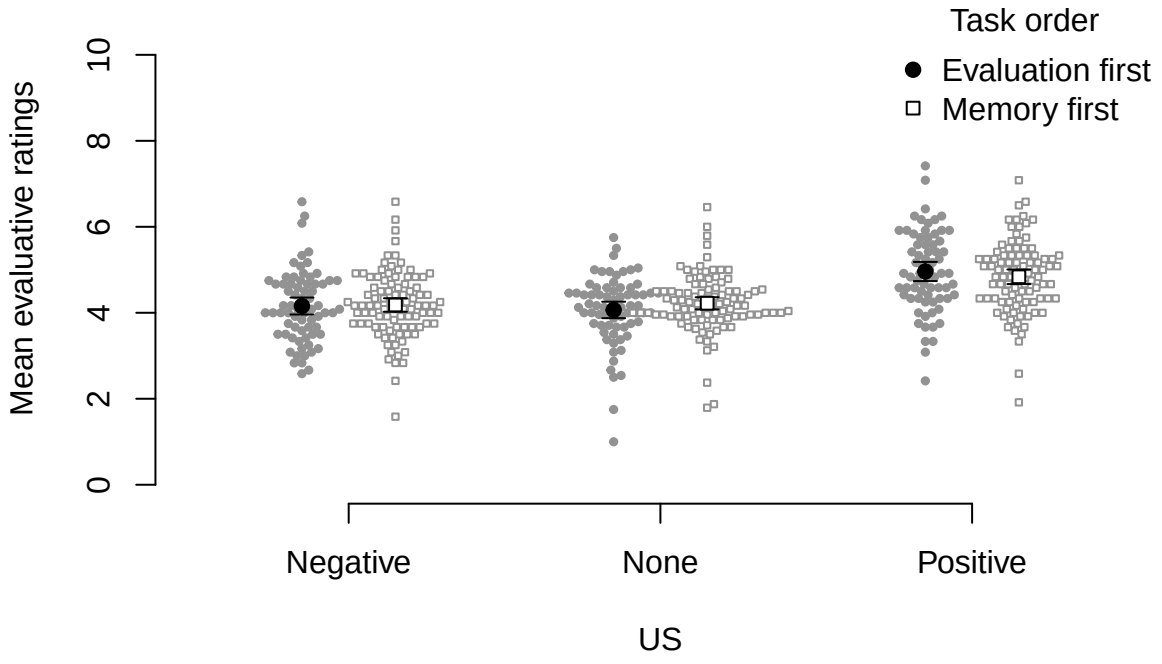

Figure 1. (#fig:plot_us_valence_order)Mean evaluative ratings (and 95% error bars) as a function of US valence and Task order.

Preregistered analyses on memory performance

Model fit and parameter estimates. The who-said-what MPT model (with $D_{\text{positive}} = D_{\text{negative}} = D_{\text{new}}$ and $1/n = .25$) was fit to the whole sample. Task order (memory first vs. evaluation first) was included as a categorical predictor of individual parameter estimates. The model fit the data well, $T_1^{\text{observed}} = 0.05$, $T_1^{\text{expected}} = 0.04$, $p = .403$, $T_2^{\text{observed}} = 2.53$, $T_2^{\text{expected}} = 2.45$, $p = .461$, and was therefore used as the baseline model to assess loss of model fit due to parameter restrictions. Measures of (absolute and relative) model fit for the baseline model (and restricted model variants) are reported in Table 1. Parameter estimates based on the baseline model can be found in Table 2.

The discrimination parameter D was .454 on average. A restricted who-said-what

Table 1

Experiment 1: Absolute fit and WAIC for the hierarchical extensions of unrestricted and restricted variants of the who-said-what model.

	Unrestricted	$D = 0$	$C_{\text{pos}} = 0$	$C_{\text{neg}} = 0$	$d_{\text{pos}} = 0$	$d_{\text{neg}} = 0$	$b = .5$	$a = .5$
Goodness of fit: Means								
T_1^{observed}	0.05	74.02	9.09	10.07	0.04	0.05	6.92	0.05
T_1^{expected}	0.04	0.05	0.04	0.04	0.04	0.04	0.05	0.04
p	.403	$< .001$	$< .001$	$< .001$	.592	.373	$< .001$	.389
Goodness of fit: Covariances								
T_2^{observed}	2.53	724.63	121.70	110.23	2.50	2.66	39.18	2.66
T_2^{expected}	2.45	2.19	2.45	2.47	2.50	2.48	2.68	2.49
p	.461	$< .001$	$< .001$	$< .001$	.484	.417	$< .001$	.420
Relative predictive accuracy								
WAIC	3,695.99	11,200.27	5,094.10	5,288.79	3,695.01	3,695.48	5,632.61	3,696.15
SE	71.33	304.31	163.84	164.93	71.43	71.54	94.70	71.59

model setting the D parameter to zero produced inadequate fit, $T_1^{\text{observed}} = 74.02$, $T_1^{\text{expected}} = 0.05$, $p < .001$, $T_2^{\text{observed}} = 724.63$, $T_2^{\text{expected}} = 2.19$, $p < .001$, and a WAIC of $11,200.27$. The WAIC of the restricted model was more than 10 points higher than the WAIC of the baseline model ($\text{WAIC}_{baseline} = 3,695.99$), indicating that the D parameter cannot be set to zero without loss of model adequacy.

The C parameter for positively (negatively) paired CSs was .666 (.864) on average. A restricted who-said-what model setting C_{positive} (C_{negative}) to zero produced inadequate fit and a WAIC of $5,094.10$ ($5,288.79$). The WAIC values were again more than 10 points higher than the WAIC of the baseline model, indicating that neither C_{positive} nor C_{negative} parameter can be set to zero without loss of model adequacy.

The d parameter for positively (negatively) paired CSs was .049 (.118) on average. A restricted who-said-what model setting d_{positive} (d_{negative}) to zero produced adequate fit and a WAIC of $3,695.01$ ($3,695.48$). The WAIC values were almost identical to the WAIC of

Table 2

Parameter estimates (with 95% credible intervals) based on a hierarchical extension of the unrestricted who-said-what model with task order as categorical predictor of individual parameter estimates.

	overall		evaluation first		memory first		
Parameter	M	95% CI	M	95% CI	M	95% CI	p
D	.454	[.413, .495]	.396	[.341, .452]	.513	[.454, .572]	.002
C_{pos}	.667	[.574, .757]	.641	[.514, .760]	.690	[.565, .802]	.272
C_{neg}	.808	[.713, .899]	.793	[.661, .906]	.818	[.690, .920]	.368
d_{pos}	.049	[.001, .176]	.055	[.000, .290]	.080	[.000, .365]	.382
d_{neg}	.118	[.003, .382]	.167	[.001, .692]	.121	[.000, .516]	.423
b	.135	[.107, .165]	.212	[.162, .266]	.081	[.056, .110]	$< .001$
a	.482	[.439, .524]	.470	[.414, .527]	.493	[.437, .549]	.269

Note. One-sided p values

the baseline model, indicating that both d_{positive} and d_{negative} can be set to zero without loss of model adequacy.

The b parameter (indicating a bias for responding "old" in the recognition task) was .135 on average. A restricted who-said-what model setting the b parameter to .5 produced inadequate fit and a $WAIC$ of $5,632.61$. The WAIC difference between the models (restricted vs. baseline) was again larger than 10, indicating that the b parameter cannot be set to .5 without loss of model adequacy.

Finally, the a parameter (indicating a bias for selecting a positive US in the recollection task) was .482 on average. A restricted who-said-what model setting the a parameter to .5 produced adequate fit and a WAIC of $3,696.15$. The WAIC was almost identical to the WAIC of the baseline model, indicating that the a parameter can be set to

.5 without loss of model adequacy.

Effects of task order on parameter estimates. MPT parameters (from the baseline model) as a function of task order are reported in Table 2. For each MPT parameter, we calculated the posterior difference between group means and its associated Bayesian p values (see Table 2). For the D parameter, the group mean in the "memory first" condition was substantially higher than the group mean in the "memory first" condition. For the b parameter, the effect of task order was reversed: here, the group mean was substantially higher in the "evaluation first" condition than in the "memory first" condition. For the remaining parameters, the posterior difference between group mean was insubstantial.

Preregistered analyses on evaluations as a function of MPT parameter estimates

EC effects. As pre-registered, we calculated several linear models including different sets of MPT parameters as predictors. None of the reported models included C_{pos} and C_{neg} as separate predictors; instead, we used the individual mean of the two parameter estimates (denoted as C hereafter). This departure from the pre-registration was made to reduce multicollinearity between predictors (since the bivariate correlation between C_{pos} and C_{neg} turned out to be extremely high, $r = .96$, 95% CI [.95, .97], $t(164) = 45.21$, $p < .001$). As pre-registered, we also calculated variance inflations factors (VIFs) for all predictors included in a given model. If the VIF of D, b or a exceeded 5, we calculated a follow-up model without the respective MPT parameter(s) as predictor(s) of individual EC effects. This analytical strategy was adopted to reduce multicollinearity with the MPT parameters that should be most directly related to EC (C, d_{pos} and d_{neg}).

Baseline-corrected CS evaluations.

Table 3

Model 1: Linear model predicting individual EC effects from D, C, d_{pos}, d_{neg}, b and a.

Predictor	*b*	95% CI	*t*	*df*	*p*	VIF
Intercept	0.72	[0.60, 0.84]	11.95	159	< .001	—
D	2.91	[1.60, 4.22]	4.38	159	< .001	4.14
C	0.88	[-0.16, 1.93]	1.67	159	.097	4.26
d_{pos}	-4.46	[-16.21, 7.28]	-0.75	159	.454	5.16
d_{neg}	10.70	[4.69, 16.72]	3.51	159	< .001	3.41
b	-1.00	[-2.23, 0.23]	-1.60	159	.112	1.97
a	0.40	[-16.37, 17.16]	0.05	159	.963	7.34

Table 4

Model 1b (follow-up): Linear model predicting individual EC effects from D, C, d_{pos}, d_{neg} and b.

Predictor	*b*	95% CI	*t*	*df*	*p*	VIF
Intercept	0.72	[0.60, 0.84]	11.98	160	< .001	—
D	2.89	[1.72, 4.07]	4.86	160	< .001	3.35
C	0.90	[0.08, 1.72]	2.17	160	.032	2.64
d_{pos}	-4.24	[-10.99, 2.52]	-1.24	160	.217	1.72
d_{neg}	10.61	[6.11, 15.11]	4.66	160	< .001	1.92
b	-0.99	[-2.18, 0.20]	-1.65	160	.102	1.85

Table 5

Model 1c (exploratory): Linear model predicting individual EC effects from D, C, d_{pos} and d_{neg}.

Predictor	*b*	95% CI	*t*	*df*	*p*	VIF
Intercept	0.72	[0.60, 0.84]	11.92	161	$< .001$	—
D	2.70	[1.54, 3.86]	4.61	161	$< .001$	3.22
C	1.06	[0.26, 1.86]	2.60	161	.010	2.5
d_{pos}	-2.87	[-9.45, 3.72]	-0.86	161	.392	1.62
d_{neg}	8.49	[4.76, 12.23]	4.49	161	$< .001$	1.31

Table 6

Model 1d (exploratory): Linear model predicting individual EC effects from C, d_{pos} and d_{neg}.

Predictor	*b*	95% CI	*t*	*df*	*p*	VIF
Intercept	0.72	[0.59, 0.84]	11.24	162	$< .001$	—
C	2.23	[1.57, 2.89]	6.65	162	$< .001$	1.51
d_{pos}	3.29	[-3.11, 9.69]	1.02	162	.311	1.36
d_{neg}	5.97	[2.18, 9.76]	3.11	162	.002	1.2

Table 7

Model 2: Linear model predicting individual EC effects from D, C, d_{pos} and d_{neg} (including two-way interactions between D and C, d_{pos} or d_{neg})

Predictor	*b*	95% CI	*t*	*df*	*p*	VIF
Intercept	0.61	[0.44, 0.77]	7.23	158	$< .001$	—
D	2.72	[1.53, 3.90]	4.53	158	$< .001$	3.38
C	1.34	[0.48, 2.20]	3.07	158	.003	2.89
d_{pos}	-5.69	[-13.25, 1.87]	-1.49	158	.139	2.14
d_{neg}	7.13	[3.08, 11.19]	3.47	158	$< .001$	1.55
$D \times C$	3.48	[-1.36, 8.32]	1.42	158	.158	2.85
$D \times d_{\text{pos}}$	-9.73	[-43.38, 23.91]	-0.57	158	.569	1.84
$D \times d_{\text{neg}}$	-5.98	[-29.17, 17.22]	-0.51	158	.612	1.59

Additional analyses

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