Supplementary online material for *Of Two Minds: A registered replication*

Tobias Heycke[1,2], Frederik Aust[1], Mahzarin R. Banaji[3], Jeremy Cone[9], Pieter Van Dessel[5], Melissa J. Ferguson[8], Xiaoqing Hu[6], Congjiao Jiang[4], Benedek Kurdi[3,8], Robert Rydell[7], Lisa Spitzer[1], Christoph Stahl[1], Christine Vitiello[4], & Jan De Houwer[5]

[1] University of Cologne
[2] GESIS - Leibniz Institute for the Social Sciences
[3] Harvard University
[4] University of Florida
[5] Ghent University
[6] The University of Hong Kong
[7] Indiana University
[8] Cornell University
[9] Williams College

Experiment 1

In the following we report details on our power analysis, the model specification used for the Bayesian model comparisons, and additional secondary analyses. We report results from the linear mixed model analysis of the IAT response times, from prior sensitivity analyses for the Bayesian model comparisons, and from an exploratory analysis of the relationship between the recognition accuracy of briefly flashed words and associative learning. Table 1 summarizes the participants' demographics separately for each location of data collection.

Table 1
Participant demographics by location.

Location	Age	Female (%)	n
Cologne	24.61 [18, 64]	70.59	51
Ghent	21.92 [17, 50]	82.00	50
Harvard	19.58 [18, 22]	57.69	52

Note. Mean age is given with range in brackets.

Power analysis

The prediction, which is supported by all previous empirical reports, is a crossed disordinal interaction between the factor *learning block* and the control factor *valence order*. Therefore, our power analyses focuses on this interaction and the theoretically relevant simple effects of *learning block*. We estimate the sensitivity of our design using the R-package *Superpower* (Caldwell & Lakens, 2019) for the contrast analyes using the R-package *emmeans* (Lenth, 2018).

Rydell, McConnell, Mackie, and Strain (2006) found the smallest learning block difference for IAT scores, $\hat{\eta}_p^2 = .100$, $d_z \approx 0.47$. The learning block differences reported by Heycke, Gehrmann, Haaf, and Stahl (2018) were of similar magnitude but with an opposite sign. Across all labs ($N = 152$), our design has 95% power to detect effects as small as $\eta_p^2 = .081$ ($\delta_z = 0.42$) and 80% power to detect effects as small as $\eta_p^2 = .050$ ($\delta_z = 0.32$) in two-sided tests. Thus, our design is sufficiently sensitive to detect (or rule out) differences 11% smaller than the smallest learning block difference reported by Rydell et al. (2006). Figure 1 illustrates the implied sensitivity in units of Cohen's δ depending on the assumed repeated-measures correlation ρ.

Mixed model analysis

The ANOVA of IAT scores reported in the main text ignores potential systematic trial-to-trial variability in IAT response latencies due to stimuli. Any such systematic but unaccounted-for variance can inflate test statistics and yield underestimated p values as well as underestimated confidence intervals. We, therefore, also conducted a linear mixed model analysis of response times with crossed random effects for participants and items to ensure that our conclusion are not contingent on inadvertent stimulus effects (for details see

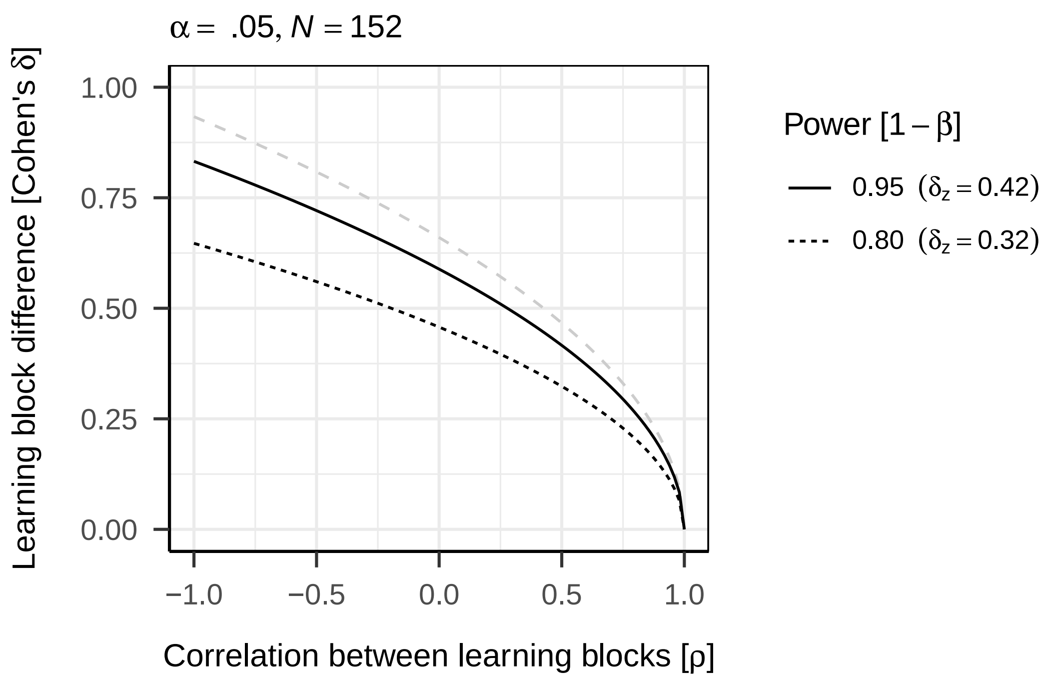

Figure 1. Sensitivity curves for learning block contrasts in Experiment 1 depending on the assumed repeated measures correlation. The grey dashed line at the top represents the smallest estimate of the learning block difference reported by Rydell et al. ($d_z = 0.47$; 2006).

Wolsiefer, Westfall, & Judd, 2017). For this analysis we excluded participants with error rates across all blocks larger than 50% or who responded faster than 300 ms on at least 10% of all trials. We additionally discarded trials in which responses were faster than 400 ms or slower than 10 s. These exclusion criteria are the same as those used by Wolsiefer et al. (2017).

We analyzed standardized response latencies, that is, the time that elapsed between stimulus presentation and *correct* response divided by the standard deviation of all response latencies in a given block, Figure 2. To assess the reversal of the response mapping effect, we contrasted the common response mapping of Bob and negative words with the common mapping of Bob and positive words. Hence, larger values represent more favorable indirect evaluations.

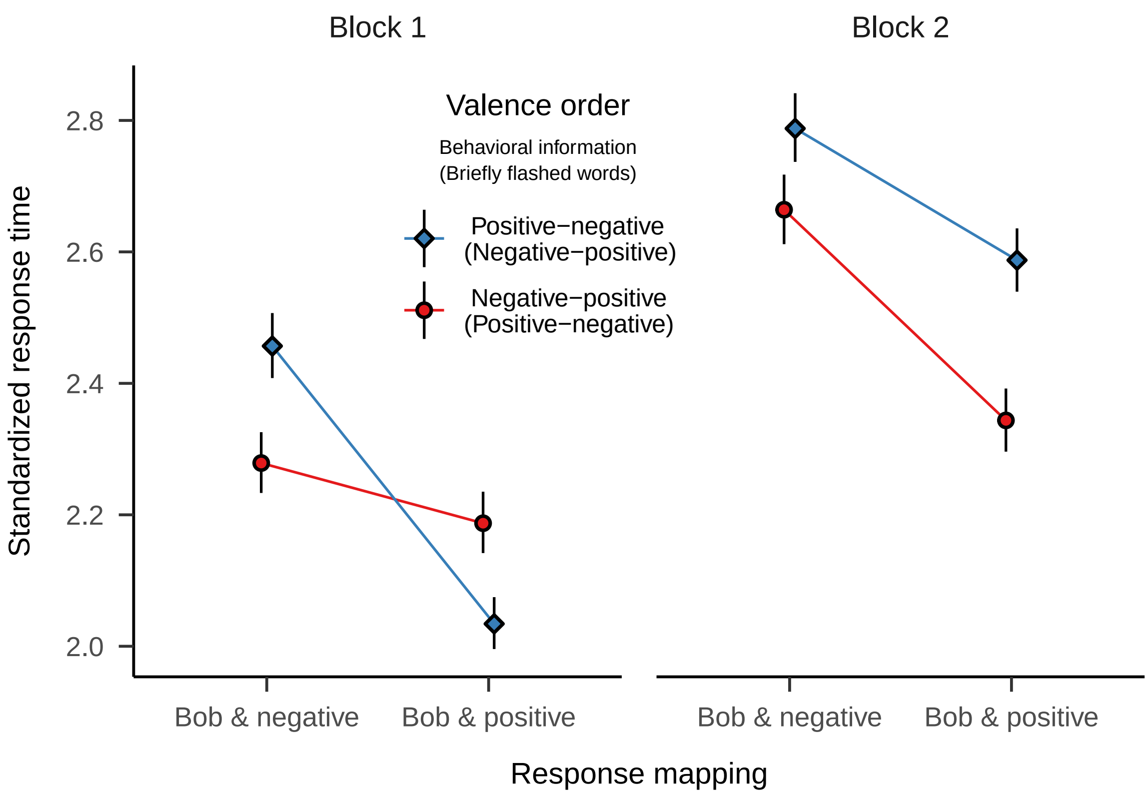

Figure 2. Standardized IAT response latencies across learning blocks. Black-rimmed points represent condition means, error bars represent 95% bootstrap confidence intervals based on 10,000 samples.

Table 2

Fixed effect estimates of the linear mixed model analysis of standardized IAT response times.

Effect	*b*	SE	*t*	*df*	*p*
Intercept	2.41	0.06	38.60	166.58	$<$ .001
Response mapping	0.13	0.01	9.52	53.48	$<$ .001
Learning block	-0.18	0.04	-4.95	151.65	$<$ .001
Valence order	-0.04	0.06	-0.66	154.12	.510
Category	-0.17	0.02	-7.47	19.58	$<$ .001
Word type	0.05	0.02	2.10	23.02	.047
Image type	-0.10	0.01	-11.02	19,358.08	$<$ .001
Response mapping × Learning block	0.00	0.01	-0.43	58.88	.667
Response mapping × Valence order	-0.02	0.01	-1.69	62.60	.096
Learning block × Valence order	0.04	0.04	1.07	151.95	.288
Response mapping × Category	-0.01	0.01	-1.62	11.28	.134
Response mapping × Word type	0.00	0.01	0.22	28.30	.826
Response mapping × Image type	-0.04	0.01	-3.79	14,941.38	$<$ .001
Learning block × Category	0.00	0.01	0.34	15.42	.741
Learning block × Word type	0.00	0.01	-0.51	47.78	.613
Learning block × Image type	0.01	0.01	1.12	12,072.69	.262
Valence order × Category	0.00	0.01	0.36	14.49	.725
Valence order × Word type	0.00	0.01	0.17	31.23	.866
Valence order × Image type	-0.01	0.01	-1.58	17,717.18	.115
Response mapping × Learning block × Valence order	-0.06	0.01	-7.02	81.48	$<$ .001
Response mapping × Learning block × Category	0.00	0.01	-0.16	42.09	.875
Response mapping × Learning block × Word type	0.01	0.01	1.26	150.14	.209
Response mapping × Learning block × Image type	0.00	0.01	0.53	19,080.54	.598
Response mapping × Valence order × Category	-0.01	0.01	-0.62	15.38	.542
Response mapping × Valence order × Word type	0.01	0.01	0.73	35.39	.471
Response mapping × Valence order × Image type	0.00	0.01	-0.15	15,412.01	.877
Learning block × Valence order × Category	0.01	0.01	0.73	34.05	.469

Table 2 continued

Effect	*b*	SE	*t*	*df*	*p*
Learning block × Valence order × Word type	0.00	0.01	-0.02	118.08	.986
Learning block × Valence order × Image type	-0.01	0.01	-1.02	12,875.69	.310
Response mapping × Learning block × Valence order × Category	0.02	0.01	2.37	75.77	.021
Response mapping × Learning block × Valence order × Word type	-0.01	0.01	-0.90	299.15	.371
Response mapping × Learning block × Valence order × Image type	0.00	0.01	0.13	18,370.00	.897

Note. The model additionally included random participant and item effects with random intercepts and random slopes for all manipulations during the learning procedure and their interactions.

Table 3

Random effect estimates and correlations of the linear mixed model analysis of standardized IAT response times.

	% of variance	1.	2.	3.	4.	5.	6.	7.	8.
Participant									
1. Intercept	.69	0.72	-0.04	-0.23	0.15				
2. Response mapping	.02		0.13	-0.09	0.21				
3. Learning block	.26			0.44	0.06				
4. Response mapping × Learning block	.00				0.06				
Stimulus									
1. Intercept	.01	0.09	-0.49	0.32	0.06	-0.78	0.66	0.60	0.12
2. Response mapping	.00		0.02	0.20	-0.39	0.79	-0.31	-0.89	-0.57
3. Learning block	.00			0.02	0.55	-0.35	0.69	-0.23	-0.76
4. Valence order	.00				0.03	-0.61	0.68	0.32	-0.12
5. Response mapping × Learning block	.00					0.01	-0.82	-0.82	-0.26
6. Response mapping × Valence order	.00						0.03	0.49	-0.06
7. Learning block × Valence order	.00							0.01	0.76
8. Response mapping × Learning block × Valence order	.00								0.01

Note. We report the estimated standard deviations in the main diagonals and the correlations in the off-diagnoals. The percentages of variance for the random effects were calculated by dividing each variance component by the total random variance, i.e., the sum of the random-effect variances.

Table 4

Post-hoc tests of changes in response mapping effects across blocks separately for pictures and words for standardized IAT response times.

Valence order	ΔM	95% CI	t	df	p
Pictures					
Negative-positive	-0.18	$[-0.32, -0.04]$	-2.96	43.24	.010
Positive-negative	0.14	$[-0.01, 0.29]$	2.26	32.11	.060
Words					
Negative-positive	-0.30	$[-0.43, -0.17]$	-5.22	198.78	$< .001$
Positive-negative	0.28	$[0.15, 0.41]$	4.81	161.48	$< .001$

Note. p values were Tukey-corrected for two comparisons.

In line with the ANOVA results, we found the expected three-way interaction between *Response mapping*, *Valence order*, and *Learning block*; the interaction was moderated by the type of stimulus that participants responded to (pictures of Bob and non-Bobs vs. positive and negative words; *Category*), Table 2 and 3. The three-way interaction prompted us to test the differences between response mapping effects in the first and second learning block for each valence order.

In line with the conventional ANOVA analysis, we found that response time differences suggested more favorable evaluations of Bob after the first than after the second block when the behavioral information was first positive and later negative, $\Delta M = 0.21$, 95% CI $[0.12, 0.30]$, $t(61.46) = 4.52$, $p < .001$. Vice versa, response time differences suggested more favorable evaluations after the second than after the first block when descriptions of Bob were first negative and later positive, $\Delta M = -0.24$, 95% CI $[-0.33, -0.15]$, $t(74.80) = -5.25$, $p < .001$. Again, these results indicate that the self-reported evaluations and IAT scores were consistent.

Due to the significant four-way interaction, we additionally explored these contrasts separately for responses to pictures of Bob vs. non-Bobs and positive vs. negative words, Table 4. We found consistent changes in response mapping effects for both pictures and words, albeit the effects were larger for words.

Bayesian model comparison

We implemented the unconstrained model as a hierarchical linear model that encompasses each of the other models as special cases:

$$\begin{aligned}\hat{y}_{ijk} = &\mu + \nu_i + \eta_l x_{1il} + \\ &(\alpha + \tau_l x_{1il}) x_{2j} x_{3k} + \\ &(\beta + \upsilon_l x_{1il})(1 - x_{2j}) x_{3k}\end{aligned}$$

The model predicts the ith participant's response to evaluation measure j in the experimental block k. Responses are predicted as a combination of a grand mean μ, random participant intercepts ν_i (i.e., habitually higher or lower evaluations), a main effect of the labs η_l, and simple effects of learning block for rating scores (α) and IAT score (β). Additionally, we allowed the simple effects to be moderated by the labs (τ_l and υ_l represent the lab-specific deviations from the overall simple effects). The model does not include a main effect of evaluative measure because any mean differences between evaluative measures were leveled by the by-measure z standardization. x_{1il} represents l effect coded variables that indicate which lab participant i belongs to; x_{2j} indicates the evaluative measure (1 for rating score and 0 for IAT score), such that $\alpha + \tau_l$ is only relevant for rating scores and $\beta + \upsilon_l$ is only relevant for IAT scores; x_{3k} is an effect coded variable that is set to 0.5 for block 1 and -0.5 for block 2.

This model allowed us to place priors on the simple effects (in units of standardized mean differences d) for each evaluative measure and implement the theoretically motivated order constraints:

$$
\begin{aligned}
\mathcal{M}_{\text{No effect}}: &\ \delta_\alpha = 0 \\
&\ \delta_\beta = 0 \\
\mathcal{M}_{\text{One mind}}: &\ \delta_\alpha \sim \text{Positive-Half-Cauchy}(r = \sqrt{2}/2) \\
&\ \delta_\beta \sim \text{Positive-Half-Cauchy}(r = \sqrt{2}/2) \\
\mathcal{M}_{\text{Two minds}}: &\ \delta_\alpha \sim \text{Positive-Half-Cauchy}(r = \sqrt{2}/2) \\
&\ \delta_\beta \sim \text{Negative-Half-Cauchy}(r = \sqrt{2}/2) \\
\mathcal{M}_{\text{Any effect}}: &\ \delta_\alpha \sim \text{Cauchy}(r = \sqrt{2}/2) \\
&\ \delta_\beta \sim \text{Cauchy}(r = \sqrt{2}/2)
\end{aligned}
$$

Additionally, we placed default multivariate Cauchy priors ($r = \sqrt{2}/2$) on lab main effects η_l as well as on lab effects on evaluative differences between blocks for rating scores (τ_l) and IAT scores (υ_l).

To formally assess whether the data from all labs exhibited consistent effects we added another model that enforced the order constraint of $\mathcal{M}_{\text{One mind}}$ and $\mathcal{M}_{\text{Two minds}}$ not only for the average block effects (α and β) but for each lab individually (i.e., $\alpha_l = \alpha + \tau_l$ and $\beta_l = \beta + \upsilon_l$; $\mathcal{M}_{\text{One mind everywhere}}$ and $\mathcal{M}_{\text{Two minds everywhere}}$).

For the analyses we drew 1 million samples to estimate the postrior distribtution of model parameters. Because the draws from the posterior distribution are used to estimate the Bayes factors for model comparisons that involve order constraints (Klugkist et al., 2005b), the number of draws implies upper and lower bounds on some of the reported Bayes factors. Most notably, as a direct consequence of the number MCMC samples the $\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Two minds}}} \in [\frac{1}{1\times10^6}, 1\times10^6]$.

Prior sensitivity analysis. Bayesian model comparison by Bayes factors are by definition sensitive to the specified prior distributions. To ensure that our inference is not contigent on our choice of piors we conducted prior sensitivity analyses for our key results.

Table 5
Results of the prior sensitivity analysis for the Bayesian model comparisons of primary interest.

r_α	r_β	$\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Two minds}}}$	$\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Any effect}}}$
0.50	0.35	1.00×10^6	4.00
0.96	0.35	1.00×10^6	4.00
0.96	0.53	1.00×10^6	4.00
0.96	0.71	1.00×10^6	4.00
1.41	0.35	1.00×10^6	4.00
1.41	0.53	1.00×10^6	4.00
1.41	0.71	1.00×10^6	4.00

Note. The Bayes factor (BF) in favor of $\mathcal{M}_{\text{One mind}}$ relative to $\mathcal{M}_{\text{Any effect}}$ is bounded within the range of $[0, 4]$ (see footnote 1 in the main article). r_α and r_β denote the scale for the Cauchy prior on the simple effects of learning block for rating scores (α) and IAT scores (β), respectively (in units of standard deviations).

Direct and indirect evaluations. Our choice of piors for the simple effects of learning block for rating scores (α) and IAT score (β) could be viewed as either overly optimistic or pessimistic. The prior on simple rating score effects places considerable probability mass on effects $d < 0.707$ although the previously reported effects were very large. Similarly, placing the same prior on the simple effects for rating and IAT scores could be criticized because the previously reported IAT score effects were considerably smaller than those of rating scores.

We, therefore, varied the scale for the Cauchy priors on the simple effects in the ranges of $0.50 < r_\alpha < 1.41$ and $0.35 < r_\beta < 0.71$ for rating and IAT scores, respectively. Considering results previous studies, we limited our reanalysis to combinations where the prior scale was larger for rating than for IAT effects. The results of the prior sensitvity analysis reassure us that our inference is robust to a wide range and combination of scales of the default Cauchy priors, see Table 5. The Bayes factors were not affected by the scale of the priors to any meaningful degree. This is because our data are informative enough to overwhelm the priors and because these Bayes factors primarily depend on the shape and location of the posterior distribution, not the prior distributions (Klugkist et al., 2005a).

Recognition task. To test the robustness of our inference regarding participants recognition accuracy we varied the scale r of the Cauchy prior in a wide interval of $[0.50, 1]$. The resulting Bayes factors were $3.89 \times 10^6 < \text{BF}_{10} < 4.91 \times 10^6$ and thus varied by a factor of 1.26. These results again reassure that our inference is robust to a wide range of scales of the default Cauchy prior.

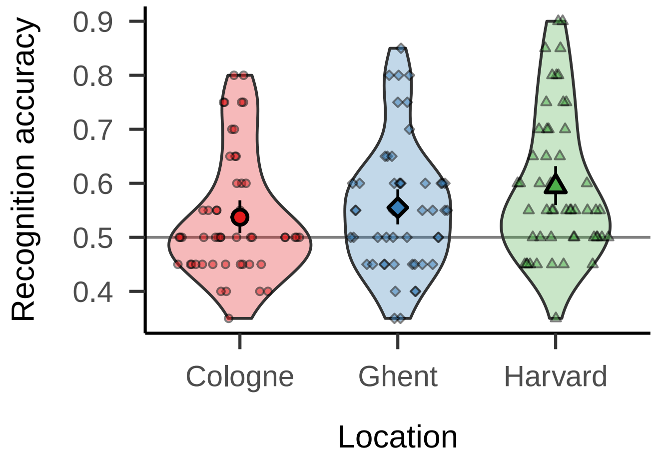

Figure 3. Black-rimmed points represent condition means, error bars represent 95% bootstrap confidence intervals based on 10,000 samples. Small points represent individual participants' accuracy. Violins represent kernel density estimates of sample distributions.

Word recogniton and indirect evaluations

In contrast to the original results reported by Rydell et al. (2006), the recognition accuracy of briefly flashed words was above chance in this study, Figure 3. Memory for these words may, thus, have interfered with the associative learning process and prevented the predicted reversal of the IAT score differences. We, therefore, performed an exploratory regression analysis of the recognition accuracy of briefly flashed words and the IAT score difference between blocks used in the Bayesian analysis above. Positive IAT score differences represent a more favorable evaluation after the block in which Bob was paired with positive behavioral information and briefly flashed negative words. Conversely, negative IAT score differences between blocks indicate that the IAT effects reflect the valence of the briefly flashed words. If word recognition indeed obstructed the associative learning process, we would expect to observe a positive relationship between the recognition accuracy and IAT score differences between blocks: When the recognition for briefly flashed words is high, IAT score differences should reflect the valence of the behavioral information but not with the word valence. We would expect to observe smaller and eventually negative IAT score differences as the recognition accuracy declines and associative learning takes over. To account for measurement error in the recognition accuracy of briefy flashed words we fit an errors-in-variable regression model (Klauer, Draine, & Greenwald, 1998). Because the model assumes that predictor values are sampled from a Gaussian distribution truncated at 0, we probit-transformed the recognition accuracy.[1]

We did not detect a relationship between the recognition accuracy of briefly preseted words and IAT score differences between learning blocks, $b = -0.15$, 95% CI $[-0.52, 0.60]$, $t(151) = -0.52$, $p = .605$. Moreover, the positive intercept of the regression line indicates a positive IAT score difference despite at-chance word recognition accuracy, $b = 0.56$, 95% CI $[5.91, 0.00]$, $t(151) = 5.91$, $p < .001$. Hence, even for participants who exhibited no

[1]A standard linear regression analysis yielded the same results.

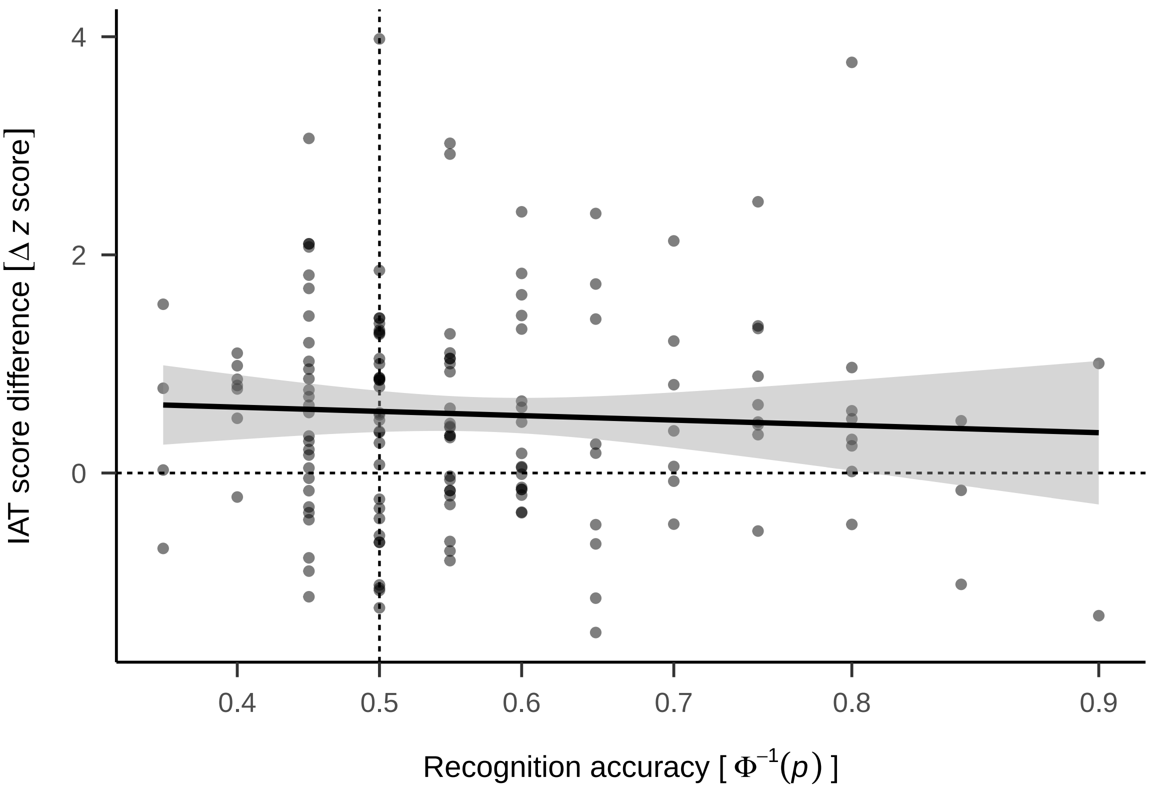

Figure 4. Scatterplot of the recognition accuracy of briefly flashed words (on probit scale) and evaluative differences in IAT scores between learning blocks in which Bob was presented with positive descriptions and those in which he was paired with negative descriptions. The regression line and confidence band represents predictions of the errors-in-variables model (Klauer et al., 1998).

memory for briefly flashed words, IAT score differences reflected the valence of the behavioral information, see Figure 4. These results provide no indication that the deviation of our findings from those reported by Rydell et al. (2006) are attributable to the above-chance recognition of briefly flashed words in this study.

Experiment 2

In the following we report the details of our a priori power analysis, provide details for the model specification used for the Bayesian model comparisons, and additional secondary analyses.

Power analysis

As for Experiment 1, our power analyses focuses on this interaction and the theoretically relevant simple effects of *learning block*. We estimate the sensitivity of our design using the R-package *Superpower* (Caldwell & Lakens, 2019) for the contrast analyes using the

R-package *emmeans* (Lenth, 2018). Rydell et al. (2006) found the smallest learning block difference for IAT scores, $\hat{\eta}_p^2 = .100$, $d_z \approx 0.47$. The learning block differences reported by Heycke et al. (2018) were of similar magnitude but with an opposite sign.

Across all locations we will recruit N = 320 participants (not including participants from the pilot study). To maximize the power of the planned contrasts, we will test whether valence order moderates the learning block contrasts by testing the main effect of learning block ($\alpha = \beta = .05$ for $\delta_z = 0.20$ and $N = 320$). If we detect no main effect of learning block, we will pool participants across valence orders by reversing the learning block coding in one group (as in the Bayesian model comparison of Experiment 1). Similarly, if the different prime presentation durations do not moderate the learning block contrasts, we will pool participants across presentation durations ($\alpha = \beta = .05$ for $\delta_z = 0.40$ and $N = 320$). This analysis plan results in four possible contrast analyses of the learning blocks: (1) 4 learning block contrasts by valence order and presentation duration, (2) 2 learning block contrasts collapsed across valence orders but separate for presentation durations, (3) 2 learning block contrasts collapsed across presentation durations but separate for valence orders, and (4) 1 learning block contrasts collapsed across valence orders and presentation durations.

The data from all locations (N = 320, not including participants from the pilot study) provides 95% power to detect learning block differences as small as $\eta_p^2 = .040$ ($\delta_z = 0.40$) or as small as $\eta_p^2 = .020$ ($\delta_z = 0.29$) and $\eta_p^2 = .010$ ($\delta_z = 0.20$) when pooling participants across one or both between-participant factors ($\alpha = .05$). Thus, our design is sufficiently sensitive to detect (or rule out) differences 13% smaller (39% or 57% when pooling participants across one or both between-participant factors, respectively) than the smallest learning block difference reported by Rydell et al. (2006). Figure 5 illustrates the implied sensitivity in units of Cohen's δ depending on the assumed repeated-measures correlation ρ.

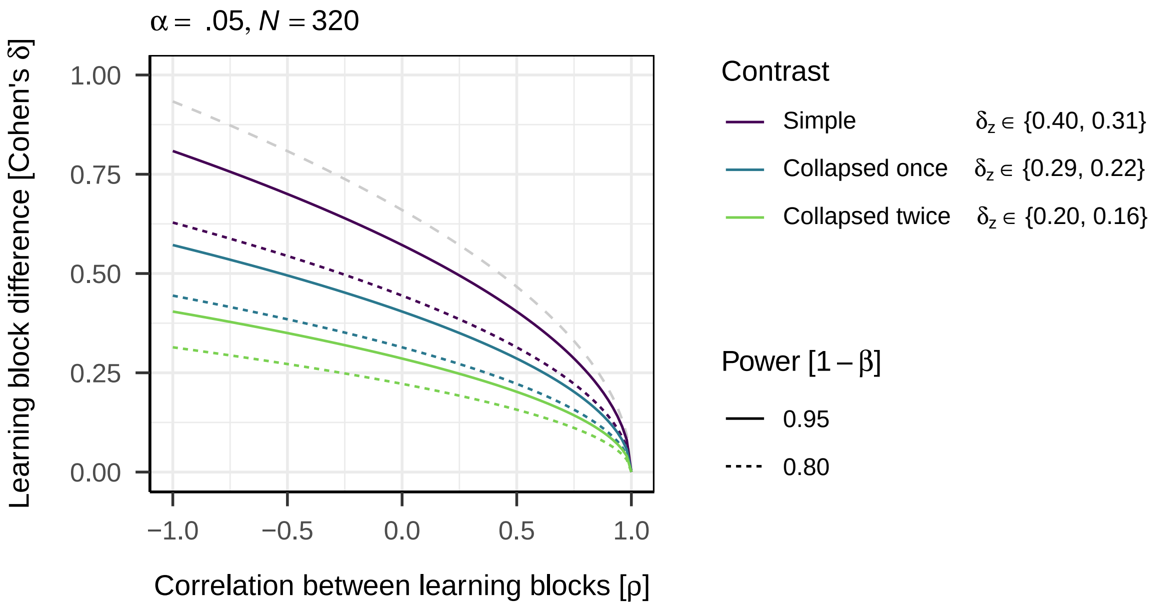

Figure 5. Sensitivity curves for learning block contrasts in Experiment 2 depending on the assumed repeated measures correlation. Simple contrast are estimated within each *flashed word presentation duration-valence order* combination (not adjusted for multiple comparisons). Collapsed contrasts pool participants across one or both of the between subject factors (i.e., *flashed word presentation duration* and *valence order*). δ_z-values in the legend correspond to 95% and 80% power. The grey dashed line at the top represents the smallest estimate of the learning block difference reported by Rydell et al. ($d_z = 0.47$; 2006).

References

Caldwell, A., & Lakens, D. (2019). *Power analysis with superpower.* Retrieved from https://arcaldwell49.github.io/SuperpowerBook

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? A registered replication of Rydell et al. (2006). *Cognition and Emotion*, *32*(8), 1708–1727. https://doi.org/10.1080/02699931.2018.1429389

Klauer, K. C., Draine, S. C., & Greenwald, A. G. (1998). An unbiased errors-in-variables approach to detecting unconscious cognition. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 253–267. https://doi.org/10.1111/j.2044-8317.1998.tb00680.x

Klugkist, I., Kato, B., & Hoijtink, H. (2005a). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69. https://doi.org/10.1111/j.1467-9574.2005.00279.x

Klugkist, I., Laudy, O., & Hoijtink, H. (2005b). Inequality Constrained Analysis of Variance: A Bayesian Approach. *Psychological Methods*, *10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Lenth, R. (2018). *Emmeans: Estimated marginal means, aka least-squares means.* Retrieved from https://CRAN.R-project.org/package=emmeans

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of Two Minds: Forming and Changing Valence-Inconsistent Implicit and Explicit Attitudes. *Psychological Science*, *17*(11), 954–958. https://doi.org/10.1111/j.1467-9280.2006.01811.x

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017). Modeling stimulus variation in three common implicit attitude tasks. *Behavior Research Methods*, *49*(4), 1193–1209. https://doi.org/10.3758/s13428-016-0779-0