Of Two Minds: A registered replication

Tobias Heycke[†,1,2], Frederik Aust[†,1,9], Mahzarin R. Banaji[3], Jeremy Cone[8], Pieter Van Dessel[5], Melissa J. Ferguson[10], Xiaoqing Hu[6], Congjiao Jiang[4], Benedek Kurdi[3,10], 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] Williams College

[9] University of Amsterdam

[10] Yale University

This manuscript has been accepted in-principle as a registered report at *Psychological Science*. Data for Experiment 2 will be collected when the SARS-CoV-2 pandemic permits.

Author Note

All data, analysis scripts and materials are available at https://osf.io/8m3xb/; the supplementary online material (SOM) is available at https://osf.io/8w9bd/.

† Tobias Heycke and Frederik Aust contributed equally to this work.

Correspondence concerning this article should be addressed to Tobias Heycke, P.O. 122155, 68072 Mannheim, Germany. E-mail: tobias.heycke@gmail.com

Abstract

Several dual-process theories of evaluative learning posit two distinct implicit (or automatic) and explicit (or controlled) evaluative learning processes. As such, one may like a person explicitly but simultaneously dislike them implicitly. Dissociations between direct measures (e.g., Likert scales), reflecting explicit evaluations, and indirect measures (e.g., Implicit Association Test), reflecting implicit evaluations, support this claim. Rydell et al. (2006) found a striking dissociation when they brief flashed either positive or negative words prior to presenting a photograph of a person was with behavioral information of the opposite valence was presented: IAT scores reflected the valence of the flashed words whereas rating scores reflected the opposite valence of the behavioral information. A recent study, however, suggests that this finding may not be replicable. Given its theoretical importance, we report two new replication attempts ($n = 153$ recruited in Belgium, Germany and the USA; $n = TBD$ recruited in Hong Kong and the USA).

Keywords: evaluative learning, subliminal influence, implicit learning, replication

Of Two Minds: A registered replication

Are our explicit and implicit evaluations of an object or person always consistent with one another? Or is it possible that we like a person explicitly but simultaneously dislike them implicitly? One way to investigate this question is to compare two families of evaluative measures: direct measures (e.g., Likert scales) that assumedly elicit relatively more explicit, conscious, effortful, and controllable evaluations (hereafter explicit evaluations), on the one hand, and indirect measures (such as the Implicit Association Test [IAT]; Greenwald, McGhee, & Schwartz, 1998) that assumedly elicit relatively more implicit, unconscious, effortless, and uncontrollable evaluations (hereafter implicit evaluations), on the other hand. Indeed, several studies have shown dissociations between direct and indirect measures (see Gawronski & Brannon, 2019). Such evidence has been critical in supporting dual-process theories positing that explicit and implicit evaluations reflect different sets of attitudes that are acquired via two distinct processes.[1]

An influential dual-process theory is the Systems of Evaluation Model (SEM; McConnell & Rydell, 2014; McConnell, Rydell, Strain, & Mackie, 2008; Rydell & McConnell, 2006). This theory assumes that implicit evaluations emerge from mental associations that develop without conscious awareness or control, from the co-occurrence of stimuli with valenced events. For example, positive associations may develop simply because a person repeatedly wears a shirt in one's favorite color. In contrast, explicit evaluations are thought to reflect propositional representations that emerge from conscious, attention-demanding reasoning processes. For example, negative propositions may develop as a result of learning that the person holds political opinions that clash with one's own views. Hence, under this theory, a double dissociation between direct and indirect measures of evaluation is expected, with the former reflecting only consciously formed propositions and the latter reflecting only

[1] By *attitude* we mean latent knowledge representations that underlie the behavioral expression of *evaluations* on direct and indirect measures (Cunningham & Zelazo, 2007).

unconsciously formed associations.

As a test of this model, Rydell et al. (2006) contrasted two different learning pathways experimentally. In the experiment, participants learned about an unfamiliar person called Bob. Each trial started with a brief (25 ms) flash of a positive or negative word, not intended to be consciously registered by participants. Then a photograph of Bob was presented alone for 250 ms before a positive or negative behavioral statement was added to the display. The statement was clearly visible until participants made a guess as to whether the behavior was characteristic or uncharacteristic of Bob. Participants immediately received feedback, which implied that Bob was a good or bad person. Crucially, this behavioral information was always opposite in valence to the briefly flashed word. In line with the predictions of the SEM, explicit evaluations of Bob, measured via self-report, reflected predominantly the valence of the behavioral information. More intriguingly, implicit evaluations, measured via the IAT, reflected predominantly the valence of the words that had been briefly flashed prior to the photograph of Bob.

This finding has been influential in support of the SEM and other dual-process theories (e.g., Gawronski & Bodenhausen, 2011). However, beyond this prominent result, empirical evidence for dual evaluative learning processes remains weak overall (Corneille & Stahl, 2019). The absence of compelling evidence that implicit evaluations emerge from unconsciously formed associations has allowed for a different, more parsimonious, account to be popularized: that both implicit and explicit evaluations reflect propositional knowledge (e.g., De Houwer, 2018). Crucially, many prominent single-process propositional theories assume that propositional learning requires conscious awareness (Mitchell, De Houwer, & Lovibond, 2009). As such, the result reported by Rydell et al. (2006), where implicit evaluations reflected predominantly unconsciously formed associations, is particularly difficult to reconcile with these accounts. Under most propositional theories, both direct (self-report) measures and indirect measures (such as the IAT) should reflect propositional knowledge that emerges from conscious, attention-demanding reasoning processes.

Given the theoretical issues at stake, a replication of the double dissociation reported by Rydell et al. (2006) is critical. If the double dissociation is replicated, such a result would lend credence to strong forms of dual-process theories positing that implicit and explicit evaluations reflect different types of (associative and propositional) representations that are acquired via different learning pathways. Moreover, such a finding would provide evidence in favor of subliminal associative learning, a phenomenon for which current evidence is weak at best (Corneille & Stahl, 2019). On the other hand, if the finding by Rydell et al. (2006) does not replicate, and both direct and indirect measures are found to reflect the valence of the consciously processed behavioral information, such a result would strengthen confidence in single-process propositional theories of evaluation. After all, these theories argue that both implicit and explicit evaluations largely reflect the same consciously formed propositions.

In two recent experiments, the double dissociation reported by Rydell et al. (2006) did not replicate (Heycke, Gehrmann, Haaf, & Stahl, 2018). Instead, both direct and indirect measures consistently reflected the valence of the behavioral information. At present, it is unclear whether these results point towards boundary conditions or call into question the replicability of the original study more generally. This ambiguity is due to the fact that materials were translated into German and stimuli were presented for a duration different from the original study. Here, we rigorously test the replicability of the double dissociation by closely adhering to the original procedure. To ensure its informativeness, the current replication attempt was conducted jointly by an international collective of experts on evaluative learning and implicit measures. Among the collaborators were the first author of the original study and authors of the previous replication attempts. To explore the generality of our results, we collected data in multiple countries and languages. A first, already concluded, experiment was conducted in Belgium, Germany and the USA. In a second experiment, for which the data is yet to be collected, we will use the insights from the first experiment to adjust the procedure to closely replicate the psychological conditions of the original study.

Experiment 1

Because the procedural modifications made by Heycke et al. (2018) may have caused the diverging results, we conducted a replication study using the unmodified experimental procedure of the original study.

Methods

The first author of the original study verified that our materials and procedure faithfully reproduced the original. The experiment was preregistered (https://osf.io/xe8au/) and data were collected at the University of Cologne (Germany), Ghent University (Belgium), and Harvard University (USA). All data files, materials, and analysis scripts are available at https://osf.io/8m3xb/. To give a vivid impression of the experimental procedure, an examplary video recording is available at https://osf.io/hmcfg/.

Material & Procedure. The experimental procedure consisted of three components: a learning task, evaluation task, and recognition task.

As in the original study, the learning task was a modified version of the evaluative learning paradigm by Kerpelman and Himmelfarb (1971). We briefly flashed a valent word followed by a longer presentation of a photograph of Bob together with a behavioral statement. Presentation durations differed across labs due to the availability of different refresh rates of the CRT monitors (85 Hz at Harvard and 75 Hz at Ghent and Cologne). In the following we will describe the setup of a trial with the presentation durations at a 75 Hz-refresh rate; deviating durations for a 85 Hz-refresh rate are given in brackets.

On each trial, a central fixation cross was displayed for 200 ms followed by a valent word flashed for 27 ms (24 ms; 2 frames). The screen background was black and text was white and set in Times New Roman font. The briefly flashed word was immediately replaced by the photograph of Bob, which served as a backward mask. Next, we provided behavioral information about Bob consisting of a behavioral statement and the additional information whether this behavior was characteristic or uncharacteristic of Bob. The photograph of Bob was presented in the center of the screen for 253 ms (247 ms) before a behavioral statement was added underneath. Participants' task was to press the "c" (= "characteristic") or "u" (= "uncharacteristic") key to guess whether the behavioral statement was characteristic or uncharacteristic of Bob. After every guess, the photograph of Bob, the behavioral statement, and the key labels were replaced with either the word "Correct" displayed in green letters or the word "False" in red letters, displayed for 5000 ms. Each trial ended with a blank screen presented for 1000 ms.

As the valence of briefly flashed words was manipulated within participants, they completed two 100-trial-blocks of the learning task. Each block consisted of trials with either only positive or negative words and the order of the blocks was randomized. The valence of the behavioral information was always opposite to the valence of the briefly flashed word. In blocks with positive words, positive behavioral statements were uncharacteristic of Bob and negative statements were characteristic. These contingencies were reversed in the blocks with negative words. We used 10 positive and 10 negative words; each of which was presented 10 times. For behavioral statements, we used 100 positive and 100 negative statements; 50 positive and 50 negative statements were randomly selected for the first block, the remaining statements were assigned to the second block. The order of briefly flashed words and behavioral information was randomized for each participant anew, whereas the order of blocks was counterbalanced across participants. A different photograph of Bob was randomly selected from six photographs of white males for each participant. The remaining five images were used in the implicit association test (see below). All materials were taken from the original study[2], with the sole exception that briefly flashed words, behavioral statements, and instructions were translated to German and Dutch for use in Germany and Belgium.

[2] The original manuscript lists the words "love", "party", "hate", and "death" as examples for briefly flashed words. The words "hate" and "love", however, were neither used as briefly flashed words in the original, nor our replication studies.

After each block, we measured evaluations of Bob directly and indirectly using Likert-scale ratings and the IAT, respectively. As in the original study, the order of the measures was the same for both blocks but counterbalanced across participants.

As direct measure of evaluation, we used three rating scales: First, participants rated Bob's likableness on a 9-point slider with the anchors labelled *Very Unlikable* and *Very Likable*. Next, again using 9-point sliders, they judged Bob on the dimensions *Bad–Good*, *Mean–Pleasant*, *Disagreeable–Agreeable*, *Uncaring–Caring*, and *Cruel–Kind*. Finally, they judged Bob on a "feeling thermometer" by entering a number between 0 (*Extremely unfavorable*) and 100 (*Extremely favorable*). Deviating from the original protocol, we collected rating scale responses as part of the computer task rather than using a paper-pencil questionnaire.

As indirect measure of evaluation, we used an IAT. Participants initially completed two types of training blocks with 20 trials each to familiarize themselves with the task. In one block, images of Bob and other white men had to be classified as Bob vs. not-Bob; in another block, positive and negative words had to be classified as positive vs. negative. In a subsequent critical block with 40 trials we intermixed the two classification tasks: Participants used one key to respond to both the images of Bob and negative words; they used another key to respond to images of other white men and positive words. After the first critical block, participants completed another training block with 20 trials of Bob vs. not-Bob with reversed key position and afterwards a second critical block with 40 trials with the reversed key mapping compared to the first critical block. It was counterbalanced whether participants completed the IAT as described above or with key mappings in reversed order (for a detailed description see Heycke et al., 2018, p. 1712). We instructed participants to respond quickly without making too many errors. In case of erroneous responses we displayed a red X as feedback and instructed participants to quickly correct their response to start the next trial.

Following the first round of evaluations, participants completed the second learning block and again evaluated Bob directly and indirectly. After the second round of evaluations, participants completed a surprise recognition test for the briefly flashed words. We presented 40 words in random order on a computer screen. Half of the words were the briefly flashed words from the learning task, the other half were new distractor words. We informed participants that 20 words were flashed briefly during the learning task, asked them to select the briefly flashed words from the list, and encouraged them to guess if they did not know the correct answer. Participants could only proceed with the experiment once they had selected exactly 20 words.

The experiment ended with a demographic questionnaire (age, field of study/profession, gender, goal of the experiment, and comments). Our procedure was identical to the original procedure, with the exception that participants completed self-reported evaluations and the recognition task at the computer rather than using paper and pencil. In Belgium and Germany, we furthermore used Dutch and German translations of the original material. The procedure took approximately 50 minutes to complete.

Data analysis. In keeping with the original analysis strategy, we calculated composite rating scores and IAT scores as direct and indirect measures of evaluation. Rating scores were the average of the three z-standardized Likert-scale responses. To calculate IAT scores we logarithmized all response times after winsorizing responses faster than 300 ms or slower than 3,000 ms. IAT scores were the difference of mean transformed response times for blocks which combined Bob and negative words and blocks which combined Bob and positive words. Thus, for rating and IAT scores larger values indicate a more positive evaluation of Bob.

How to statistically assess the success of a replication attempt is subject of current debate (e.g., Fabrigar & Wegener, 2016; Simonsohn, 2013; Verhagen & Wagenmakers, 2014). Whether a pattern of results has been replicated is challenging to measure directly if the to-be-replicated pattern consists of more than two cells of a factorial design. One elegant approach is to instantiate a pattern of mean differences (i.e., the rank order of means), predicted by a theory or observed in a previous study, as order constraints in a statistical model (e.g., Hoijtink, 2012; Rouder, Haaf, & Aust, 2018). With the model in hand, replication success can be quantified as predictive accuracy of this model relative to a competing model, such as a null model or an encompassing unconstrained model (e.g., Rouder et al., 2018).

Based on previously reported results, there are two competing predictions for the current paradigm: (1) Rydell et al. (2006) reported that across both learning blocks ratings scores were congruent with the behavioral information about Bob, whereas IAT scores were incongruent with the behavioral information ($\mathcal{H}_{\text{Two minds}}$). (2) In contrast, Heycke et al. (2018) observed a consistent pattern for rating scores and IAT scores; both measures were congruent with the behavioral information ($\mathcal{H}_{\text{One mind}}$). We considered two additional predictions: no effect of the manipulation ($\mathcal{H}_{\text{No effect}}$) and the all-encompassing prediction of any outcome ($\mathcal{H}_{\text{Any effect}}$). If, of all predictions considered, our results are best described by the prediction of no effect, our experimental manipulations failed. The prediction of any effect reflects the possibility that we may observe an entirely unexpected outcome that is neither in line with the results reported by Rydell et al. (2006) or Heycke et al. (2018).

We implemented all predictions as order (or null) constraints in an ANOVA model with default (multivariate) Cauchy priors ($r = 0.5$ for fixed effects and $r = 1$ for random participant effects, see SOM for details; Rouder, Morey, Speckman, & Province, 2012; Rouder et al., 2018). To simplify the presentation of the Bayesian model comparison results, we collapsed data across valence orders such that we always contrasted blocks where the behavioral information was positive with those where it was negative. Thus, for both rating and IAT scores positive difference indicate that evaluations are congruent with the valence of the behavioral information, whereas negative values indicate that evaluations are congruent with the valence of the briefly flashed words. We assessed the relative predictive accuracy of these models by Bayesian model comparisons using Bayes factors. Note that comparisons of models where one model is a special order-constrained case of the other are asymmetric. Consider the example of $\mathcal{H}_{\text{One mind}}$, which is a special case of $\mathcal{H}_{\text{Any effect}}$. If the data are perfectly consistent with $\mathcal{M}_{\text{One mind}}$, they are inevitably also perfectly consistent with $\mathcal{M}_{\text{Any effect}}$. In this case $\mathcal{M}_{\text{One mind}}$ will be favored by the Bayes factor because $\mathcal{M}_{\text{One mind}}$ makes a more specific prediction—it predicts that 3/4 of the outcomes predicted by $\mathcal{M}_{\text{Any effect}}$ are impossible, Figure 2A. The degree to which the order-constrained model is more specific (more parsimonious) places an upper bound on the Bayes factor in its favor. On the other hand, there is no such bound on the Bayes factor in favor of the unconstrained model if the data are inconsistent with the order constraint—that is, the data fall outside of the predictive space deemd possible by the order-constrained model. It follows that $\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Any effect}}} \in [0, 4]$ because $\mathcal{M}_{\text{One mind}}$ limits its predictions to 1/4 of those of $\mathcal{M}_{\text{Any effect}}$. To guide their interpretation, we report the theoretical bounds on the reported Bayes factors alongside our results where applicable. Finally, we tested whether recognition memory accuracy using a one-tailed Bayesian t test with default Cauchy prior ($r = \sqrt{2}/2$; Rouder, Speckman, Sun, Morey, & Iverson, 2009).

To facilitate comparisons with previously reported statistics, we also conducted the frequentist analyses described by Rydell et al. (2006). To ensure that our conclusions about indirectly measured evaluations are robust to stimulus effects, we supplemented the ANOVA analysis of IAT scores by a frequentist linear mixed model analysis, see SOM. We used R (Version 3.6.3; R Core Team, 2018) and the R-packages *afex* (Version 0.23.0; Singmann, Bolker, Westfall, & Aust, 2018), *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *emmeans* (Version 1.5.1; Lenth, 2018), and *papaja* (Version 0.1.0.9997; Aust & Barth, 2018) for all our analyses.

Participants. We set out to collect $n = 50$ participants at each location (N = 150). We recruited 155 participants (aged 17-64 years, $M = 22.02$; 69.93% female, 0.65% nonbinary; see supplementary online material [SOM] for details); two participants were excluded due to technical failures. Hence, the reported results are based on data from 153 participants. We compensated all participants with either € 8/10 (Cologne/Ghent), or partial course credit (Cologne/Harvard).

Statistical power. 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*. Our assessment of the statistical senstivitiy of our design focused on the tests of simple *learning block* effects, because they are of primary theoretical interest and less sensitive than the test of the interaction. We estimate the sensitivity for the frequentist analyses described by Rydell et al. (2006) using the R-package *Superpower* (Caldwell & Lakens, 2019). The smallest simple effect of learning block reported by Rydell et al. (2006) was $d_z \approx 0.47$ ($\hat{\eta}_p^2 = .100$) for IAT scores.[3] Across all locations, our planned contrasts had 95% power to detect learning block effects as small as $\delta_z = 0.42$[4] ($\eta_p^2 = .081$; $N = 152$, $\alpha = .05$, two-sided tests). Thus, our design is sufficiently sensitive to detect (or rule out) differences 11% smaller than the smallest learning block difference reported in the original study.

Results

In the following, *valence order* refers to the joint order of briefly flashed words and behavioral information. Any time we refer to one valence order (e.g., positive-negative) we specify the order of the behavioral information; briefly flashed words were always of the opposite valence.

To reiterate, Rydell et al. (2006) reported that across learning blocks ratings scores were congruent with the behavioral information about Bob, whereas IAT scores were incongruent with the behavioral information. This pattern of results implies (1) a three-way

[3] The learning block differences reported by Heycke et al. (2018) were of similar magnitude but with an opposite sign.

[4] We report the implied sensitivity in units of Cohen's δ depending on the assumed repeated-measures correlation ρ in the supplementary material.

Table 1

Means and 95% confidence intervals of rating and IAT scores in Experiment 1 broken down by valence order, learning block, and lab location.

	Rating score		IAT score	
ValenceBlock	Learning block 1	Learning block 2	Learning block 1	Learning block 2
Cologne				
Negative-positive	-0.89 [-1.02, -0.76]	0.72 [0.56, 0.87]	0.02 [-0.05, 0.10]	0.15 [0.09, 0.21]
Positive-negative	0.97 [0.85, 1.09]	-0.82 [-0.97, -0.67]	0.18 [0.11, 0.25]	0.06 [0.00, 0.11]
Ghent				
Negative-positive	-0.81 [-0.94, -0.69]	0.91 [0.77, 1.06]	0.06 [-0.01, 0.13]	0.15 [0.09, 0.20]
Positive-negative	0.81 [0.68, 0.93]	-0.80 [-0.96, -0.65]	0.20 [0.12, 0.27]	0.11 [0.05, 0.17]
Harvard				
Negative-positive	-1.03 [-1.16, -0.91]	0.93 [0.78, 1.08]	0.03 [-0.04, 0.10]	0.10 [0.05, 0.16]
Positive-negative	0.99 [0.86, 1.11]	-0.95 [-1.10, -0.80]	0.12 [0.05, 0.19]	0.05 [0.00, 0.11]

interaction of *measure of evaluation*, *valence order*, and *learning block* in a joint analysis of all evaluations, (2) two opposite crossed disordinal interactions of *valence order* and *learning block* for separate analyses of rating and IAT scores, (3) larger rating scores following learning blocks in which the behavioral information was positive compared to when it was negative, and, finally, (4) smaller IAT scores following learning blocks in which the behavioral information was positive compared to when it was negative. We first report the results of the frequentist analyses described by Rydell et al. (2006). Busy readers interested in an integrative replicability assessment may wish to skip ahead to the Bayesian model comparisons.

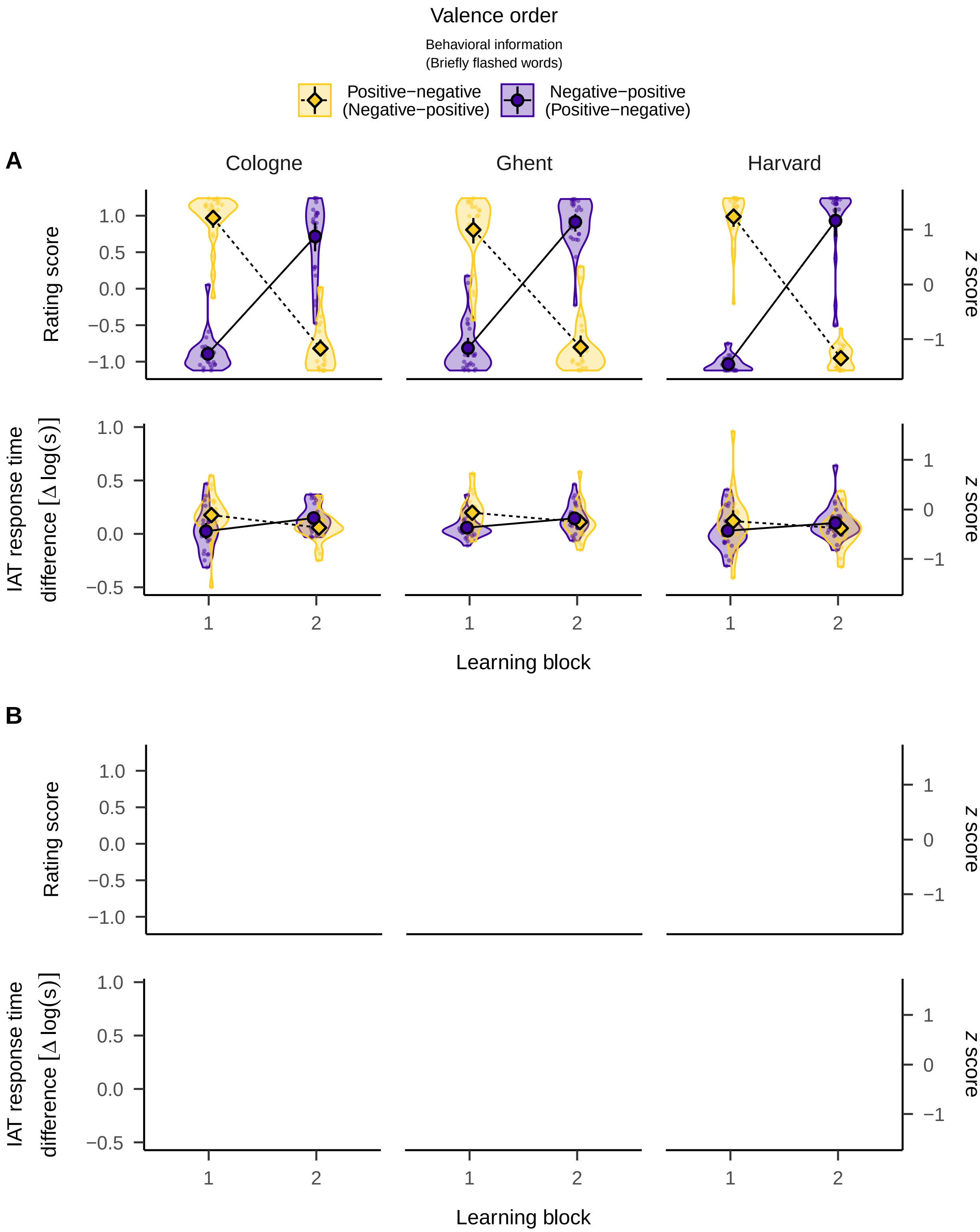

Figure 1. Mean evaluative rating and IAT scores for Experiments 1 (**A**) and Experiment 2 (**B**) broken down by valence order, learning block, and lab location. Black-rimmed points represent condition means, error bars represent 95% bootstrap confidence intervals based on 10,000 samples, small points represent individual participant scores, and violins represent kernel density estimates of sample distributions.

Joint analysis of rating and IAT scores. For a joint analysis, we separately z-standardized directly and indirectly measured evaluations and submitted them to a four-way ANOVA with the factors *measure of evaluation* (direct vs. indirect), *valence order* (positive or negative behavioral information first), *learning block* (first or second learning block), and *lab location* (Cologne, Ghent, Harvard). Table 1 summarizes the condition means. We found a significant three-way interaction between valence order, learning block, and measure of evaluation, $d = 2.40$, 90% $[1.97, 0.65]$, $F(1, 147) = 210.82$, $MSE = 0.31$, $p < .001$, Figure 1[5]. Moreover, we observed a significant four-way interaction indicating that the three-way interaction differed between lab locations, $\hat{\eta}^2_p = 0.05$, 90% $[0.00, 0.10]$, $F(2, 147) = 3.48$, $MSE = 0.31$, $p = .033$. Follow-up tests indicated that the three-way interaction was significant in each lab (all $F(1, 147) > 46.62$, $p < .001$) and the direction of the effect was consistent across labs. In line with the original analysis, we next examined the interaction between valence order, learning block, and lab location in separate analyses of rating and IAT scores.

Direct measure: Evaluative rating scores. As in the previous studies, for rating scores we found a two-way interaction between valence order and learning block, $d = 6.51$, 90% $[5.69, 0.93]$, $F(1, 147) = 1,556.14$, $MSE = 0.15$, $p < .001$. This interaction was significant in each lab (all $F(1, 147) > 450.58$, $p < .001$), but also differed in magnitude, $\hat{\eta}^2_p = 0.05$, 90% $[0.00, 0.11]$, $F(2, 147) = 4.05$, $MSE = 0.15$, $p = .019$. In all labs, rating scores corresponded to the valence of the behavioral information. Rating scores indicated *more* favorable evaluations after the first than after the second block when behavioral information was first positive and later negative, Cologne: $d_z = -1.34$, 95% CI $[-1.56, -1.12]$; Ghent:

[5] Figure 1 may give the impression that the difference between valence orders was of similar magnitude at learning block 1 and 2 in rating scores but differed in IAT scores. However, we found differences between valence orders at learning blocks 1 and 2 in both measures of evaluation (all $t(147) > 2.51$, $p < .013$) and we did not find these differences between valence orders to vary between evaluative measures, $d = 0.16$, 90% $[-0.16, 0.04]$, $F(1, 147) = 0.94$, $MSE = 0.76$, $p = .334$.

$d_z = -1.21$, 95% CI $[-1.42, -0.99]$; Harvard: $d_z = -1.45$, 95% CI $[-1.69, -1.22]$; all $t(147) < -14.19$, $p < .001$. Conversely, rating scores indicated *less* favorable evaluations after the first than after the second block when behavioral information was first negative and later positive, Cologne: $d_z = 1.21$, 95% CI $[0.99, 1.42]$; Ghent: $d_z = 1.29$, 95% CI $[1.08, 1.51]$; Harvard: $d_z = 1.47$, 95% CI $[1.24, 1.71]$; all $t(147) > 14.19$, $p < .001$. Hence, in all labs directly measured evaluations corresponded to the valence of the behavioral information and were opposite to the valence of the briefly flashed words.

Indirect measure: IAT scores. For IAT scores, we found a two-way interaction between valence order and learning block, $d = 1.10$, 90% $[0.75, 0.33]$, $F(1, 147) = 44.68$, $MSE = 0.01$, $p < .001$; in this case we detected no differences across labs, $\hat{\eta}_p^2 = 0.02$, 90% $[0.00, 0.04]$, $F(2, 147) = 1.19$, $MSE = 0.01$, $p = .308$. In all labs, IAT scores corresponded to the valence of the behavioral information. IAT scores indicated *more* favorable evaluations after the first than after the second block when behavioral information was first positive and later negative, $d_z = -0.38$, 95% CI $[-0.55, -0.21]$, $t(147) = -4.64$, $p < .001$. Conversely, IAT scores indicated *less* favorable evaluations after the first than after the second block when behavioral information was first negative and later positive, $d_z = 0.40$, 95% CI $[0.23, 0.57]$, $t(147) = 4.81$, $p < .001$. The results of the mixed model analysis corroborated the conclusions from the ANOVA analysis, see SOM. Hence, in all labs indirectly measured evaluations corresponded to the valence of the behavioral information and were opposite to the valence of the briefly flashed words. Directly and indirectly measured evaluations did not dissociate.

Differences between rating and IAT scores. In keeping with our preregisted analysis plan, we also compared z-standardized directly and indirectly measured evaluations—despite the consistent pattern of results—and found that they differed across measures in every condition. When behavioral information was first positive and later negative, rating scores indicated a more favorable evaluation than IAT scores in the first block, $d_z = 0.41$, 95% CI $[0.23, 0.59]$, $t(147) = 4.64$, $p < .001$, but a less favorable evaluation in the second block, $d_z = -0.51$, 95% CI $[-0.67, -0.35]$, $t(147) = -6.79$, $p < .001$. Conversely, when behavioral information was first negative and later positive rating scores indicated a less evaluation than IAT scores in the first block, $d_z = -0.40$, 95% CI $[-0.59, -0.22]$, $t(147) = -4.54$, $p < .001$, but a more favorable evaluation in the second block, $d_z = 0.49$, 95% CI $[0.33, 0.65]$, $t(147) = 6.52$, $p < .001$. These results, corroborate that directly and indirectly measured evaluations were consistent, but indicate that directly measured evaluations were more extreme than indirect measured evaluations.

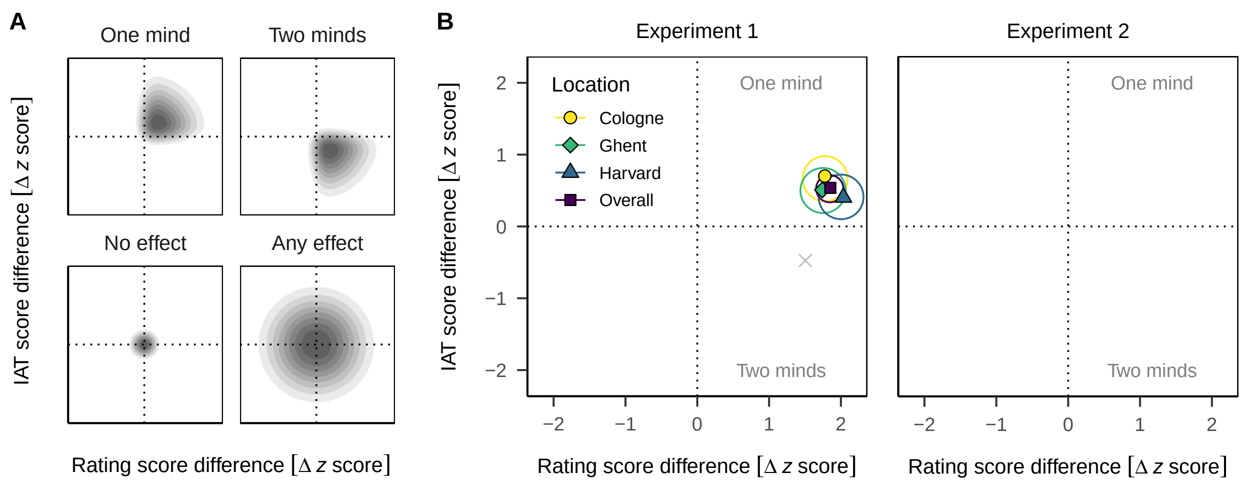

Figure 2. Predictions of the four models of primary interest (**A**) and results of Experiment 1 and Experiment 2 (**B**). Black-rimmed points represent mean differences in evaluations between the two learning blocks. To simplify the presentation of the results, we collapsed data across valence orders such that we always contrasted blocks where the behavioral information was positive with those where it was negative. Thus, for both rating and IAT scores positive difference indicate that evaluations correspond to the valence of the behavioral information, whereas negative values indicate that evaluations correspond to the valence of the briefly flashed words. Ellipses represent 95% Bayesian credible intervals based on the unconstrained model $\mathcal{M}_{\text{Any effect}}$. For comparison, the grey × represents the learning block differences reported in the original study.

Bayesian model comparisons. The direct comparison of predictive accuracy indicated that our data overwhelmingly favored the qualitative pattern reported by Heycke et al. (2018) over that reported by Rydell et al. (2006), $\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Two minds}}} = 1.00 \times 10^6$, Table 2. Additional comparisons with the control models confirmed that the experimental manipulations were effective ($\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{No effect}}} = 3.06 \times 10^{86}$) and did not produce an unexpected result, $\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Any effect}}} = 4.00 \in [0, 4]$.

We additionally assessed whether all labs consistently produced the same result pattern. We implemented a model that enforced the order-constraint of $\mathcal{M}_{\text{One mind}}$ not only on the average learning block effects but on each lab's learning block effect. Our data provide strong evidence for consistent result patterns across labs relative to the less-constrained models, $\text{BF}_{\mathcal{M}_{\text{One mind everywhere}}/\mathcal{M}_{\text{One mind}}} = 2.76 \in [0, 3]$ and $\text{BF}_{\mathcal{M}_{\text{One mind everywhere}}/\mathcal{M}_{\text{Any effect}}} = 11.05 \in [0, 12]$. As noted in the Data analysis section, due to the upper bounds on the Bayes factors, we could not have obtained much stronger evidence in favor of $\mathcal{M}_{\text{One mind everywhere}}$. Prior sensitivity analyses confirmed that our results are robust to a wide range of priors, see SOM.

Recognition of briefly presented words. Finally, we examined participants' recognition memory for the briefly flashed words at the end of the study. Recognition accuracy was better than chance, $M = .56$, 95% CI $[.55, \infty]$, $t(152) = 6.24$, $p < .001$, $\text{BF}_{10} = 4.59 \times 10^6$. Hence, we cannot assume that the stimulus presentation was outside of participants' conscious awareness. It remains unclear whether recognition accuracy differed between labs, $\hat{\eta}^2_p = 0.04$, 90% $[0.00, 0.09]$, $F(2, 150) = 2.94$, $MSE = 0.01$, $p = .056$, $\text{BF}_{01} = 1.27$ (see SOM for details).

Discussion

As confirmed by the first author of the original study, we faithfully reproduced the procedure of Rydell et al. (2006), but the original results did not replicate. We observed that both directly and indirectly measured evaluations reflected the valence of the behavioral

Table 2

Summary of Bayesian model comparisons.

	Experiment 1		Experiment 2	
Model ($\mathcal{M}_i$)	$\mathrm{BF}_{\mathcal{M}_i/\mathcal{M}_{\text{Any effect}}}$	NPP	$\mathrm{BF}_{\mathcal{M}_i/\mathcal{M}_{\text{Any effect}}}$	NPP
No effect	0.00	.00		
One mind	4.00	.25		
... everywhere	11.05	.69		
Two minds	0.00	.00		
... everywhere	0.00	.00		
Any effect		.06		

Note. As noted in the Data analysis section, the Bayes factors (BF) in favor of $\mathcal{M}_{\text{One mind}}$ and $\mathcal{M}_{\text{One mind everywhere}}$ relative to $\mathcal{M}_{\text{Any effect}}$ are bounded within the range of $[0, 4]$ and $[0, 12]$, respectively. Hence, in both model comparisons we could not have obtained much stronger evidence against $\mathcal{M}_{\text{Any effect}}$. The direct comparison of the models of primary interest overwhelmingly favored $\mathcal{M}_{\text{One mind}}$ over $\mathcal{M}_{\text{Two minds}}$, $\mathrm{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Two minds}}} = 1.00 \times 10^6$. The naive posterior probability (NPP) quantifies the probability of each model given the data assuming that all models are equally likely a priori.

information; the briefly flashed words did not produce a reversal of the indirectly measured evaluations. In short, we found no dissociation between directly and indirectly measured evaluations. Our findings mirror the results of the previous replication attempt by Heycke et al. (2018). Moreover, our results were consistent across three languages and countries indicating that neither inaccurate translations nor differences in sampled populations are likely to have caused the divergence from the original finding. Thus, our results raise more doubts about the replicability of the dissociative evaluative learning effect that was reported by Rydell et al. (2006).

There is, however, one objection our data cannot dispel: The close physical recreation of the original procedure does not guarantee a faithful reproduction of the psychological conditions of the original learning task. In the original study, recognition accuracy of the briefly flashed words was not significantly different from chance (Rydell et al., 2006). Like Heycke et al. (2018), however, we observed better-than-chance recognition accuracy. We have to assume that participants consciously perceived at least some of the briefly flashed words, which may have affected our results. Hence, it is possible that the conscious perception of briefly flashed words constitutes a critical departure from the to-be-reproduced learning conditions. Although an exploratory analysis suggested that there was no relationship between recognition accuracy and indirectly measured evaluations (see SOM), we decided to repeat the experiment and reduce the visibility of briefly flashed words to more closely mimic the psychological conditions of the original study.

Experiment 2

To address the concern that our previous replication may have been unsuccessful because briefly flashed words were consciously perceived, we will conduct a second study and reduce the presentation duration of the briefly flashed words during the learning task.

Pilot study

To identify a presentation duration that reproduces the psychological conditions of the original study (i.e., at-chance recognition accuracy for briefly flashed words), we ran a pilot study with a presentation duration reduced to 13 ms (one frame on a 75 Hz CRT monitor).[6] Because all subsequent studies will be conducted in English, the pilot study used the English material and was conducted at the University of Florida. Except for the shorter presentation duration the methods were the same as in Experiment 1. For the pilot study, we recruited 60 participants (aged 18-21 years, $M = 18.38$; 56.67% female).

Recognition accuracy for the briefly flashed words was not significantly better than chance, $M = 0.51$, 95% CI $[0.50, \infty]$, $t(59) = 1.31$, $p = .098$, but the Bayesian evidence for at-chance accuracy was inconclusive, $\mathrm{BF}_{01} = 1.76$. Based on these results we cannot rule out that, even with the shortened presentation duration, briefly flashed words were recognized above chance. To confirm that the recognition accuracy was comparable to the original study, we performed a nonsuperiority test. We compared the observed accuracy to the smallest deviation from at-chance accuracy that could have been detected in the original study, i.e., $M = 0.53$. The test confirmed that the recognition accuracy was comparable to that observed by Rydell et al. (2006), , $M = .48$, 95% CI [.45, .51], $t(59) = -2.05$, $p = .022$. Thus, we conclude that the visibility of words flashed for 13 ms is likely to be functionally comparable to that of the original study. Of course the presentation duration could be reduced further to obtain conclusive evidence for at-chance visibility, but this runs the risk of inadvertently causing stimuli to become practically invisible. To safeguard against the possibility that the 13 ms presentation duration is already too brief, we will add a second

[6] We ran a series of pilot studies in Dutch, which also yielded above-chance recognition of briefly flashed words. These pilot studies employed a shortened procedure, used Dutch material, or were conducted immediately after an unrelated priming study, which also used briefly flashed words. We, therefore, decided a posteriori, that above-chance accuracy in these studies may not be informative for our subsequent replication attempt, as we will use only English materials in the next studies.

presentation duration and flash words for for 20 ms in some locations[7]. This means that across both studies, briefly flashed words will have been presented for 13 ms, 20 ms 24 ms, and 27 ms.

Method

Material & Procedure. We will use the same materials and procedure as in Experiment 1 but flash words for 13 ms or 20 ms. Furthermore, all labs will use the same Python script to collect the data and only the English material will be used to match the official language at all locations.

Data analysis. The new data[8] from all locations will be submitted to analyses analogous to those of Experiment 1. We will, again, perform the analyses reported in the original study and assess replication success by performing Bayesian model comparisons. In contrast to Experiment 1, all labs will use the same stimulus material and lab location will be partially confounded with the presentation duration of the briefly flashed words. Thus, we will replace the lab location factor by presentation duration of the briefly flashed words in both analyses. Additionally, we will compare the data from Hong Kong to those from the American labs to explore whether our results are consistent across ethnicities and cultures. Given the consistent results in Experiment 1, we will omit the linear mixed model analysis of IAT response times.

To maximize the power of the planned contrasts in the frequentist ANOVA analyses,

[7] In case we can collect data in all five locations, the following sentence will be added to the manuscript: Three locations flashe words for 20 ms; only two locations flashed words for 13 ms because we also included the data of pilot study (N = 60) in the overall analysis, which also used a 13 ms presentation duration.

[8] To ensure valid results, the pilot study for Experiment 2 employed the complete experimental procedure, that is, we also collected evaluative ratings and IAT responses. As of now, only the word recognition accuracy was analyzed; we have not looked at evaluative ratings and IAT responses. Once the data of the second, preregistered experiment are in, we will add the data from the pilot study to our final analyses.

we will test whether valence order moderates the learning block contrasts by testing the main effect of learning block. 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 presentation durations of flashed words do not moderate the learning block contrasts, we will pool participants across presentation durations. All data and analysis code will be made available in the OSF repository and linked to in the manuscript.

Participants. If the current SARS-CoV-2 pandemic premits, we will recruit 80 participants at Yale University, the University of Florida, the University of Hong Kong, Indiana University Bloomington, and Williams College, but in no less than four of these locations. As in Experiment 1, all participants who sign up, before the planned sample size has been reached will be allowed to participate. We will, again, recruit additional participants to replace those excluded, unless data removal is requested after completion of the data collection.

Statistical power. As for Experiment 1, our assessment of the statistical senstivitiy of our design focused on the tests of simple *learning block* effects. Across the minimum of four locations, our planned contrasts will have 95% power to detect learning block effects as small as $\delta_z = 0.40$ ($\eta_p^2 = .040$) or as small as $\delta_z = 0.29$ ($\eta_p^2 = .020$) and $\delta_z = 0.20$ ($\eta_p^2 = .010$) when pooling participants across one or both between-participant factors ($N = 320$, $\alpha = .05$, two-sided tests).[9] The tests of the main effect of learning block and the three-way interaction, on which we will base our decision to pool participants across the between-subject conditions, will have 95% power to detect effecs as small as $\delta_z = 0.20$ ($\eta_p^2 = .010$) and $\delta_z = 0.40$ ($\eta_p^2 = .040$), respectively ($N = 320$, $\alpha = .05$, two-sided tests). 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,

[9] We report the implied sensitivity in units of Cohen's δ depending on the assumed repeated-measures correlation ρ in the supplementary material.

respectively) than the smallest learning block difference reported by Rydell et al. (2006). Note that these are conservative estimates as they do not take into account the additional 60 participants from our pilot study that we will include in the analysis and because we may collect data in five rather than four locations.

References

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

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

Corneille, O., & Stahl, C. (2019). Associative Attitude Learning: A Closer Look at Evidence and How It Relates to Attitude Models. *Personality and Social Psychology Review*, *23*(2), 161–198. https://doi.org/10.1177/1088868318763261

Cunningham, W. A., & Zelazo, P. D. (2007). Attitudes and evaluations: A social cognitive neuroscience perspective. *Trends in Cognitive Sciences*, *11*(3), 97–104. https://doi.org/10.1016/j.tics.2006.12.005

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin*, *13*(3). https://doi.org/10.5964/spb.v13i3.28046

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80. https://doi.org/10.1016/j.jesp.2015.07.009

Gawronski, B., & Bodenhausen, G. V. (2011). The AssociativePropositional Evaluation Model: Theory, Evidence, and Open Questions. *Advances in Experimental Social Psychology*, *44*, 59–128.

Gawronski, B., & Brannon, S. M. (2019). What is cognitive consistency, and why does it matter? In E. Harmon-Jones (Ed.), *Cognitive dissonance: Reexamining a pivotal theory in psychology (2nd ed.).* (pp. 91–116). Washington: American Psychological Association. https://doi.org/10.1037/0000135-005

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton: CRC.

Kerpelman, J. P., & Himmelfarb, S. (1971). Partial reinforcement effects in attitude acquisition and counterconditioning. *Journal of Personality and Social Psychology*, *19*(3), 301–305. https://doi.org/10.1037/h0031447

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

McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 204–218). The Guilford Press.

McConnell, A. R., Rydell, R. J., Strain, L. M., & Mackie, D. M. (2008). Forming implicit and explicit attitudes toward individuals: Social group association cues. *Journal of Personality and Social Psychology*, *94*(5), 792–807. https://doi.org/10.1037/0022-3514.94.5.792

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(02), 183. https://doi.org/10.1017/S0140525X09000855

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*(1), 41–56.

https://doi.org/10.1080/03637751.2017.1394581

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, *91*(6), 995–1008. https://doi.org/10.1037/0022-3514.91.6.995

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

Simonsohn, U. (2013). *Small Telescopes: Detectability and the Evaluation of Replication Results* (SSRN Scholarly Paper No. ID 2259879). Rochester, NY: Social Science Research Network.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *Afex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731