Of Two Minds: A registered replication

Tobias Heycke[1, 2], Frederik Aust[1], Mahzarin R. Banaji[3], Pieter Van Dessel[5], Xiaoqing Hu[6], Congjiao Jiang[4], Benedek Kurdi[3], 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

Manuscript under review (10/2019), Version 1.

Author Note

All data, analysis scripts and materials can be found at https://osf.io/8m3xb/

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

Abstract

Findings of dissociations between implicit (i.e., automatic) and explicit (i.e., non-automatic) evaluations that are based on distinct associative (i.e., co-occurrence based) and propositional (i.e., rule-based) learning procedures have fueled the dominance of dual-process theories of evaluative learning for decades. Arguably the most influential evidence has been found in a study by Rydell, McConnell, Mackie, and Strain (2006) in which participants learned about a person named Bob. It was observed that implicit evaluations reflected the valence of brief pairings of valenced words with the image of Bob whereas explicit evaluations reflected the (opposite) valence of the behavioral statements that were instructed to be characteristic of Bob. A recent study by Heycke and colleagues (2018) was unable to reproduce this data pattern independently. Given the theoretical importance of the findings by Rydell and colleagues, we present a series of additional replication attempts conducted by an international collective of researchers including the first author of the original finding.

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

Of Two Minds: A registered replication

Evaluative conditioning (EC) refers to the robust finding of a change in liking of a stimulus, referred to as conditioned stimulus (CS), that is due to its pairing with a stimulus of positive or negative valence, referred to as unconditioned stimulus (US) (De Houwer, 2007; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010). Dual-process theories postulate that EC can depend both on a mechanism via which co-occurrences between stimuli result in the automatic formation of mental associations as well as a second mechanism via which propositional beliefs are formed (e.g., Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006). Though dual-process theories have had a huge impact, it has been argued that there is actually little evidence for automatic association formation as a distinct mechanism for EC (see Corneille & Stahl, 2019). This conclusion is in line with propositional single-process views (e.g., Mitchell, De Houwer, & Lovibond, 2009) which postulate that all associative (evaluative) learning (including EC) is based on propositional processes.

One finding, however, is particularly difficult to reconcile with propositional single-process accounts: Rydell et al. (2006) reported that participants exhibited conflicting explicit and implicit evaluations (e.g., positive explicit and negative implicit evaluations) of a single target person after the presentation of rapid stimulus-stimulus pairings and behavioral statements that implied the opposite valence. In this study, participants engaged in a learning task about a person named Bob in which they determined whether Bob was a good or bad person by guessing whether positive and negative behavioral statements were characteristic of Bob. They received feedback about the accuracy of each of their guesses. Unbeknownst to the participants, on each trial, a valent word (US) was presented briefly (25 ms) prior to the presentation of a picture of Bob and the behavioral statement. Critically, the valence of the US words was always opposite to the behavioral information (e.g., USs were always positive when negative behavioral statements were characteristic of Bob). Following the learning task, explicit evaluations of Bob, assessed with a trait rating task,

reflected the valence of the learned behavioral information, while implicit evaluations, assessed with an Implicit Association Test (IAT: Greenwald, McGhee, & Schwartz, 1998) reflected the valence of the USs. These results were conceptually replicated by the original authors using similar methods (Rydell & McConnell, 2006; Rydell, McConnell, & Mackie, 2008). The Rydell et al. (2006) article has been referred to as providing strong evidence for numerous dual-process theories (Gawronski & Bodenhausen, 2011; Rydell & McConnell, 2006) and has been cited more than 320 times overall (Google Scholar, 10.15.2019).

Judging from its impact, the reported effect is generally assumed to be robust and replicable. A recent study, composed of two experiments, however, could not replicate the dissociative effect (Heycke, Gehrmann, Haaf, & Stahl, 2018). Instead, both implicit and explicit evaluations consistently reflected the valence of the learned behavioral information. At present, it remains unclear whether these unsuccessful attempts call into question the replicability of the original study or point towards boundary conditions. This ambiguity is due to the modifications made to the original materials and procedure in the replication studies: First, some words were replaced to ensure that no words presented in the learning phase served as distractors in the US memory assessment at the end of the study (see Heycke et al., 2018). Second, one experiment used German translations of the verbal material (rather than the original English material). Third, the other experiment, while using the original material, was conducted with a modified procedure: Because a subset of participants correctly identified USs after the learning task in the first experiment, the presentation duration of USs was reduced from 25 ms to 16 ms to ensure at-chance post-learning US recognition.

In the current study we will rigorously test the replicability of the original findings reported by Rydell et al. (2006) by closely adhering to the original procedure. To ensure its informativeness, the current replication attempt was a joint effort that involved researchers from different labs that all have expertise in studying evaluative conditioning and implicit

measures. Moreover, both the first author of the original study and authors of the unsuccessful replication attempts collaborated on the studies presented here. To explore the generality of our results, data were collected in multiple countries and languages. In a first, already concluded, experiment we have conducted a close replication in three different locations. In a second experiment, for which the data has not been collected yet, we will use the knowledge gained from the first experiment to adjust the procedure to closely replicate the psychological conditions of the original study.

Experiment 1

As the results reported by Heycke et al. (2018) could have been the result of the modifications to the experimental procedure, we conducted a replication study using the unmodified experimental procedure of the original study.

Methods

Materials and procedure were inspected by the first author of the original study to ensure that they closely corresponded to the original. The experiment was preregistered (https://osf.io/xe8au/) and data were collected in Germany, Belgium, and the United States. All data files, materials and analysis files are available at https://osf.io/8m3xb/.

Material & Procedure. The experimental procedure consisted of three components: learning task, evaluation task, and US 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 presented a US followed by a longer presentation of a picture of Bob together with a statement that described the behavior of Bob. Presentation durations differed across labs due to the availability of different refresh rates of the CRT monitors (85 Hz in the United States and 75 Hz in Belgium and Germany). 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 US displayed for 27 ms (24 ms; 2 frames). USs were immediately replaced by the picture of Bob, which served as a backward mask. The picture of Bob was presented in the center of the screen for 253 ms (247 ms) before the behavioral information was added underneath. The screen background was black and text was colored white and set in Times New Roman font. Participants' task was to press the "c" (= "characteristic") or "u" (= "uncharacteristic") key to judge whether the behavioral information was characteristic or uncharacteristic of Bob, which was displayed on the screen during the judgment phase. After every judgment, the image 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 US valence 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 USs and the order of USs was randomized. In blocks with positive USs, positive behavioral information was always uncharacteristic of Bob and negative information was characteristic. These contingencies were reversed in the blocks with negative USs. We used 10 positive and 10 negative words as USs each of which was presented 10 times. For behavioral information, 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 USs and behavioral information was randomized for each participant anew, whereas the order of blocks was counterbalanced across participants. A different picture of Bob was randomly selected from six pictures of white males for each participant. The remaining five images were used in the implicit evaluation measure (see below). All materials were taken from the original study, with the sole exception that US words, behavioral information, and instructions were translated to

German and Dutch for use in Germany and Belgium.

After each block, participants completed implicit and explicit evaluation measures. As in the original study, the order of the measures was the same for both blocks but counterbalanced across participants.

The explicit evaluation consisted of three parts, each presented on a single screen: 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 explicit evaluations as part of the computer task rather than using a paper-pencil questionnaire.

We assessed the implicit evaluations using the IAT described by Rydell et al. (2006). 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 with the key mappings described above or with key mappings in reversed order (for a detailed description see Heycke et al., 2018, p. 1712).

After completing the explicit and implicit evaluation measures, participants completed the second learning block and again completed implicit and explicit evaluations. Following the second set of evaluations, participants completed a surprise US recognition task. We presented 40 words in random order on a computer screen. Half of the words were the briefly presented USs from the learning task, the other half were new distractor words. We informed participants that 20 words were presented briefly during the learning task, asked them to select the briefly presented 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 explicit evaluation and US recognition tasks 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.

Participants. We set out to collect $N = 50$ participants at each location (Rydell et al., 2006). For each evaluation measure, the data from all three labs ($N = 150$) provides 95% power to observe two-way interaction effects as small as $f = 0.15$ ($d = 0.30$) and contrasts between blocks of the learning task as small as $d_z = 0.27$ ($\alpha = \beta = .05$, repeated-measures correlation $r = .5$, Nosek, Greenwald, & Banaji, 2007). Our design was, thus, adequately powered to detect effects half the size of those reported by Rydell et al. (2006). 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 failure. Hence, the following results are based on data from 153 participants. We compensated all participants with either € 8/10 (Cologne/Ghent), or partial course credit (Cologne/Harvard).

Data analysis. How to evaluate the success of a replication attempt statistically 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 assess 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) found that across both measurement times explicit ratings were congruent with the learned behavioral information about Bob, whereas IAT scores were incongruent with learned behavioral information ($\mathcal{H}_{\text{Two minds}}$). (2) In contrast, Heycke et al. (2018) observed the same pattern for explicit ratings and IAT scores; across both measurement times both measures were congruent with the learned 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 did not succeed. The prediction of any effects serves as a control and 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 evaluate the relative predictive accuracy of these models we

performed Bayesian model comparison using Bayes factors. To test whether recognition memory performance was above chance we used a one-tailed Bayesian t test with default Cauchy prior ($r = \sqrt{2}/2$; Rouder, Speckman, Sun, Morey, & Iverson, 2009). We also report the results of the analyses described by Rydell et al. (2006) to facilitate comparisons with previously reported statistics.

For both analyses, we processed the IAT response times as described in the original study (for details see Rydell et al., 2006; Heycke et al., 2018). To ensure that our conclusions are robust to stimulus effects, we additionally supplemented the ANOVA analysis of IAT scores by a frequentist linear mixed model analysis, see SOM. To simplify the presentation of the Bayesian model comparison results, we collapsed data across valence orders by reversing the coding of the blocks when learned behaviors were first negative and subsequently positive. Thus, for both explicit and implicit evaluation measures we contrasted the first with the second block, such that positive values represent a more positive evaluation following the block in which learned behaviors were positive and USs negative. We used R (Version 3.6.1; R Core Team, 2018) and the R-packages *afex* (Version 0.25.1; Singmann, Bolker, Westfall, & Aust, 2018), *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *emmeans* (Version 1.4.1; Lenth, 2018), and *papaja* (Version 0.1.0.9842; Aust & Barth, 2018) for all our analyses.

Results

In the following, *valence order* refers to the joint order of learned behaviors and briefly presented USs. Any time we refer to one valence order (e.g., positive-negative) we specify the order of the learned behavior; the corresponding briefly presented USs were always of the opposite valence.

In the joint analysis of implicit and explicit evaluations, we found the three-way interaction between valence order, learning block, and evaluation measure, $F(1, 147) = 210.82$, $MSE = 0.31$, $p < .001$, $\hat{\eta}^2_G = .173$ previously reported by Rydell et al.

(2006) (Figure 1). Moreover, we found that the three-way interaction was significant in each lab (all $F(1, 147) > 46.62$, $p < .001$), but differed slightly in magnitude, $F(2, 147) = 3.48$, $MSE = 0.31$, $p = .033$, $\hat{\eta}^2_G = .007$. The direction of the effect was consistent across labs. As in the original analysis we examined the significant three-way interaction with separate analyses of each evaluation measure.

Explicit evaluation. As in the previous studies, we found a two-way interaction between valence order and learning block, $F(1, 147) = 1,556.14$, $MSE = 7.56$, $p < .001$, $\hat{\eta}^2_G = .868$. This interaction was significant in each lab (all $F(1, 147) > 450.58$, $p < .001$), but also differed between labs, $F(2, 147) = 4.05$, $MSE = 7.56$, $p = .019$, $\hat{\eta}^2_G = .033$. In all labs, ratings of Bob were more favorable after the first than after the second block when learned behaviors were first positive and later negative, Cologne: $\Delta M = -12.51$, 95% CI [−13.99, −11.03]; Ghent: $\Delta M = -11.26$, 95% CI [−12.83, −9.69]; Harvard: $\Delta M = -13.58$, 95% CI [−15.08, −12.07]; all $t(147) < -14.19$, $p < .001$. Vice versa, in all labs, ratings of Bob were more favorable after the second than after the first block when learned behaviors were first negative and later positive, Cologne: $\Delta M = 11.26$, 95% CI [9.69, 12.83]; Ghent: $\Delta M = 12.10$, 95% CI [10.59, 13.61]; Harvard: $\Delta M = 13.75$, 95% CI [12.25, 15.26]; all $t(147) < 18.04$, $p < .001$. Hence, the predicted differences were consistently detectable in all labs but differed in magnitude. No other effects were significant.

Implicit evaluation. For IAT scores, we found a two-way interaction between valence order and learning block, $F(1, 147) = 44.68$, $MSE = 0.01$, $p < .001$, $\hat{\eta}^2_G = .075$; in this case we detected no differences between labs, $F(2, 147) = 1.19$, $MSE = 0.01$, $p = .308$, $\hat{\eta}^2_G = .004$. IAT scores for Bob were larger, indicating a more favorable attitude, after the first than after the second block when learned behaviors were first positive and later negative, $\Delta M = -0.09$, 95% CI [−0.13, −0.05], $t(147) = -4.64$, $p < .001$. Vice versa, IAT scores for Bob were larger after the second than after the first block when learned behaviors were first negative and later positive, $\Delta M = 0.09$, 95% CI [0.06, 0.13], $t(147) = 4.81$, $p < .001$. The results of the mixed model analysis were in line with the conclusions from the

ANOVA analysis, see SOM. Hence, evaluations of Bob assessed with the IAT corresponded to the valence of the learned behaviors and contradicted the valence of the briefly presented USs. Explicit ratings and IAT scores did not dissociate.

Differences between implicit and explicit evaluations. As in the original study, we additionally compared participants' z standardized evaluations toward Bob assessed by ratings and IAT scores.

Evaluations assessed by rating and IAT scores differed in all conditions. Ratings of Bob were more favorable than IAT scores in the first block, $\Delta M = 0.59$, 95% CI [0.34, 0.84], $t(147) = 4.64$, $p < .001$, but less favorable in the second block when learned behaviors were first positive and later negative, $\Delta M = -0.74$, 95% CI [−0.95, −0.52], $t(147) = -6.79$, $p < .001$. Conversely, ratings of Bob were less favorable than IAT scores in the first block, $\Delta M = -0.58$, 95% CI [−0.84, −0.33], $t(147) = -4.54$, $p < .001$, but more favorable in the second block when learned behaviors were first positive and later negative, $\Delta M = 0.71$, 95% CI [0.50, 0.93], $t(147) = 6.52$, $p < .001$. Given that evaluations were consistent in valence across measures, these differences indicate that evaluations assessed by ratings were more extreme than those assessed by IAT scores.

Bayesian model comparisons. The direct comparison of predictive accuracy indicated that our data overwhelmingly favored the result patterns reported by Heycke et al. (2018) over those reported by Rydell et al. (2006), $\mathrm{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Two minds}}} = 1.00 \times 10^{6}$, Table 1. Additional comparisons with the control models confirmed that the experimental manipulations were effective ($\mathrm{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{No effect}}} = 3.06 \times 10^{86}$) and did not produce an

unexpected result pattern ($\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Any effect}}} = 4.00, \in [0, 4]$[1]).

We additionally assessed whether all labs 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 block effects but on each lab's 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 $\text{BF}_{\mathcal{M}_{\text{One mind everywhere}}/\mathcal{M}_{\text{Any effect}}} = 11.05$ ($\in [0, 12]$[2]). Prior sensitivity analyses confirmed that the above results are robust to a wide range of priors, see SOM.

Recognition task. Finally, we examined participants' recognition memory for the words that we presented briefly as USs during the learning procedure. Recognition accuracy was better than chance, $M = .56$, 95% CI [.55, ∞], $t(152) = 6.24$, $p < .001$, $M = 0.56$ 95% HDI [0.54, 0.58], $\text{BF}_{10} = 4.59 \times 10^6$. Hence, we cannot assume that the stimulus presentation was subliminal. It remained unclear, however, whether recognition accuracy differed between labs, $F(2, 150) = 2.94$, $MSE = 0.01$, $p = .056$, $\hat{\eta}^2_G = .038$, $\text{BF}_{01} = 1.27$.

---

[1] The comparison of these models is asymmetric. If, as in this case, the data are perfectly consistent with $\mathcal{M}_{\text{One mind}}$, they are also consistent with $\mathcal{M}_{\text{Any effect}}$. The order restriction enforced by $\mathcal{M}_{\text{One mind}}$ limits the prediction of the model to 1/4 of the outcome space predicted by $\mathcal{M}_{\text{Any effect}}$ (i.e., the upper right quadrant of Figure 2). This four-fold greater parsimony of $\mathcal{M}_{\text{One mind}}$ constitutes the upper bound for $\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Any effect}}}$ (assuming all regions of the outcome space are equally likely a priori). Hence, for this model comparison we could not have obtained stronger evidence (given numerical imprecision of the MCMC sampling approach). Conversely, if the data had fallen outside the predicted outcome space of $\mathcal{M}_{\text{One mind}}$ there is no upper bound to the evidence in favor of $\mathcal{M}_{\text{Any effect}}$.

[2] The additional order constraints enforced by $\mathcal{M}_{\text{One mind everywhere}}$ limits the prediction of the model to 1/12 of the outcome space predicted by the unconstrained model. Hence, for this model comparison the upper bound for $\text{BF}_{\mathcal{M}_{\text{One mind everywhere}}/\mathcal{M}_{\text{Any effect}}}$ is 12 (assuming all regions of the outcome space are equally likely a priori). Baring the transitivity of Bayes factors in mind this implies that the upper bound for $\text{BF}_{\mathcal{M}_{\text{One mind everywhere}}/\mathcal{M}_{\text{One mind}}}$ is 3. Hence, in both model comparisons we could not have obtained much stronger evidence in favor of $\mathcal{M}_{\text{One mind everywhere}}$.

Discussion

We reproduced the procedure of Rydell et al. (2006), using the original material and the first author of the original study approved the procedure. Despite all efforts to keep the procedure as close as possible to the original procedure, we did not replicate the original finding. In contrast to Rydell et al. (2006), we observed that both implicit and explicit evaluations reflected the valence of the learned behavioral information. The briefly presented USs did not lead to contradicting implicit evaluations compared to explicit evaluations. In short, we found no evidence for an evaluative dissociation. Our findings mirror the results of the previous replication attempt by Heycke et al. (2018). Moreover, the observed results were consistent across three languages and countries indicating that neither inaccurate translations nor differences in sampled populations are likely to have caused the deviation of the results from the original findings. Thus, our findings raise more doubts about the general replicability of the dissociative evaluative learning effect that was observed by Rydell et al. (2006).

There is, however, one objection these findings cannot entirely dispel: The close physical recreation of the original procedure may not have faithfully reproduced the psychological conditions of the original learning task. In the original study, US recognition accuracy was not significantly different from chance (Rydell et al., 2006). However, like Heycke et al. (2018) we did observe better-than-chance US recognition accuracy. We, therefore, have to assume that participants consciously perceived the briefly presented USs, which might have influenced our results. Hence, it is possible that the conscious perception of briefly presented USs constitutes a critical departure from the to-be-reproduced learning conditions. Although an exploratory analysis suggested that there was no relationship between US recognition accuracy and implicit evaluations (see SOM), we decided to repeat the experiment and reduced US visibility experimentally to more closely mimic the psychological conditions of the original learning task.

Experiment 2

To address the concern that our previous replication may have been unsuccessful because USs were consciously perceived (as indicated by above-chance US recognition accuracy), we will conduct a second study and reduce the presentation duration of USs during the learning task.

Visibility pilot study

To identify a presentation duration that replicates the psychological conditions of the original learning task (i.e., at-chance recognition accuracy for briefly presented USs), we ran a pilot study with a reduced US presentation duration of 13 ms (one frame on a 75 Hz CRT monitor).[3] Because all subsequent studies will be conducted in English, the pilot study used the English material and instructions and was conducted at the University of Florida. Except for the shorter US presentation 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).

The US identification performance was not significantly better than chance, $M = 0.51$, 95% CI [0.50, ∞], $t(59) = 1.31$, $p = .098$, however the data provide no evidence for at-chance accuracy $M = 0.51$, 95% HDI [0.50, 0.53], $\text{BF}_{01} = 1.76$. Although these results do not formally confirm that the shortened presentation duration yielded at-chance US visibility, the estimates of US memory accuracy are very similar to those from Rydell et al. (2006) ($M = .48$, 95% CI [.45, .51]). Further reduction of the presentation duration may eventually yield conclusive evidence for at-chance visibility, but it could inadvertently cause stimuli to become practically invisible. We will, therefore, conduct the final studies using a US

[3] We ran a series of pilot studies in Dutch, which also yielded above-chance US memory. These pilot studies employed a shortened procedure, used Dutch material, or were conducted immediately after an unrelated priming study, which also used briefly presented words. We, therefore, decided a posteriori, that above-chance performance 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 of 13 ms.

Method

Participants. We will run the experiment in Florida, Indiana, and Hong Kong. As in Experiment 1, 50 students will participate at each of three locations, yielding a total of $N = 150$ participants. All participants who sign up for the study before the intended $n = 50$ has been reached will be allowed to participate; hence, the final sample size could be slightly larger. We will exclude any participants from the analysis who abort the experiment or who ask us to remove their data (e.g., because they did not pay attention or did not follow the instructions). We will recruit additional participants to replace those excluded.

Material & Procedure. We will employ the same material and procedure as in Experiment 1 but use the shortened US presentation duration of 13 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 three locations. All materials are available at https://osf.io/8m3xb/.

Data analysis. The new data[4] from all locations will be jointly submitted to the same analyses as in Experiment 1. We will, again, perform the analyses reported in the original study, supplement the ANOVA analysis of IAT scores by a linear mixed model analysis, and assess the replication success by performing Bayesian model comparisons. All data and analysis code will be made available in the OSF repository and linked to in the manuscript.

---

[4] To ensure representative 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 US identification performance was analyzed; we have not looked at the explicit and implicit evaluation data. Once the data of the second, preregistered experiment are in, we will add the data from the pilot study to our final analyses.

References

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

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

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology*, *10*(02), 230–241.

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. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

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

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

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916

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

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

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In *Social psychology and the unconscious: The automaticity of higher mental processes* (pp. 265–292). New York, NY, US: Psychology Press.

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. (2008). Consequences of discrepant explicit and implicit attitudes: Cognitive dissonance and increased information processing. *Journal of Experimental Social Psychology*, *44*(6), 1526–1532. https://doi.org/10.1016/j.jesp.2008.07.006

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

Table 1

Summary of Bayesian model comparisons.

	Experiment 1		Experiment 2	
Model ($\mathcal{M}_i$)	$\text{BF}_{\mathcal{M}_i/\mathcal{M}_{\text{Any effect}}}$	NPP	$\text{BF}_{\mathcal{M}_i/\mathcal{M}_{\text{Any effect}}}$	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. 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 $[0, 4]$[1] and $[0, 12]$[2], 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 $\mathcal{M}_{\text{Two minds}}$, $\text{BF}_{\mathcal{M}_{\text{One mind}}/\mathcal{M}_{\text{Two minds}}} = 1.00 \times 10^6$. The naive posterior probability (NPP) quantifies the probability of each model given the data assuming that all models are equally likely a priori.

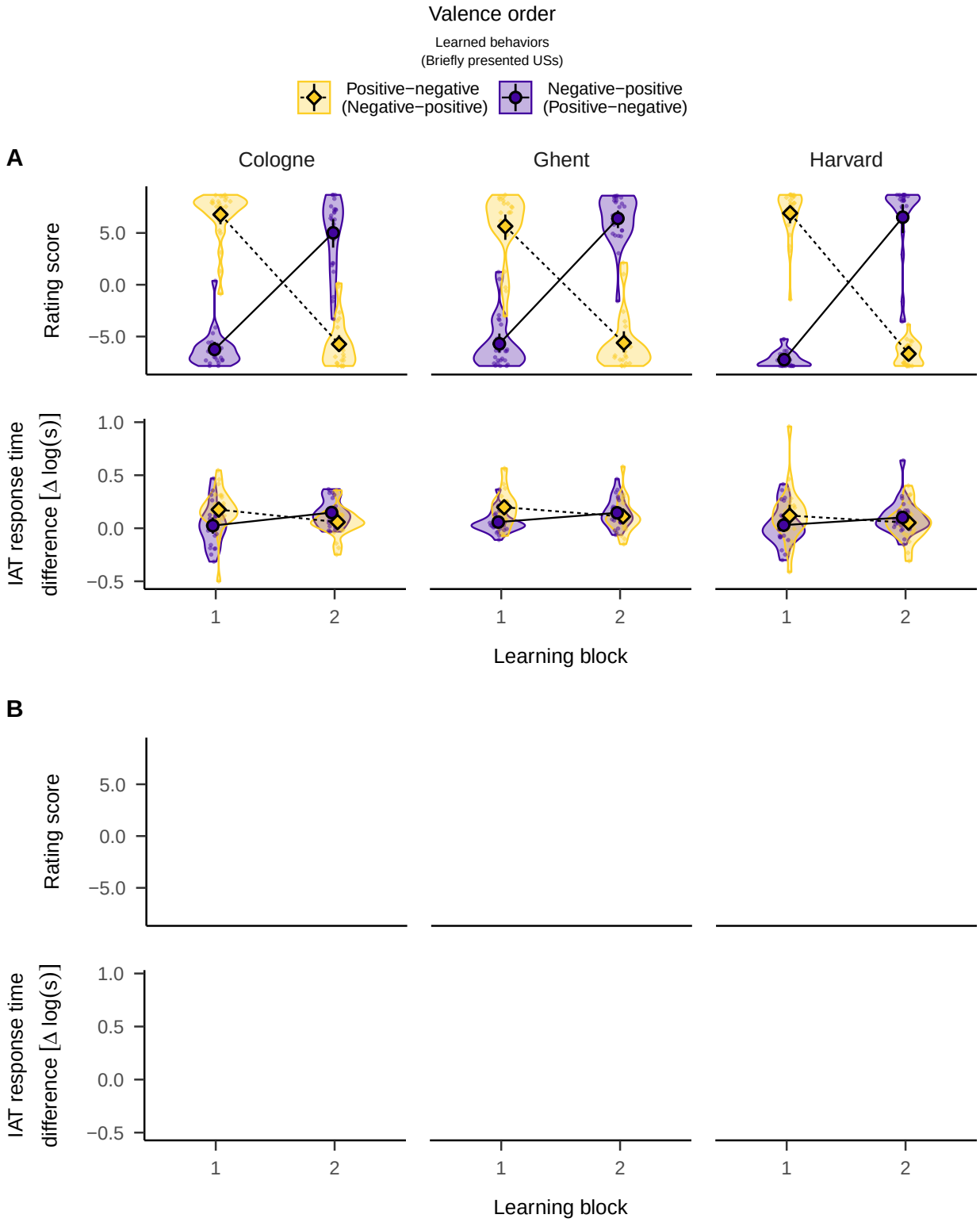

Figure 1. Evaluative rating and IAT scores for Experiments 1 (**A**) and Experiment 2 (**B**). Black-rimmed points represent condition means, error bars represent 95% bootstrap confidence intervals based on 10,000 samples, small points represent individual participants' responses, and violins represent kernel density estimates of sample distributions.

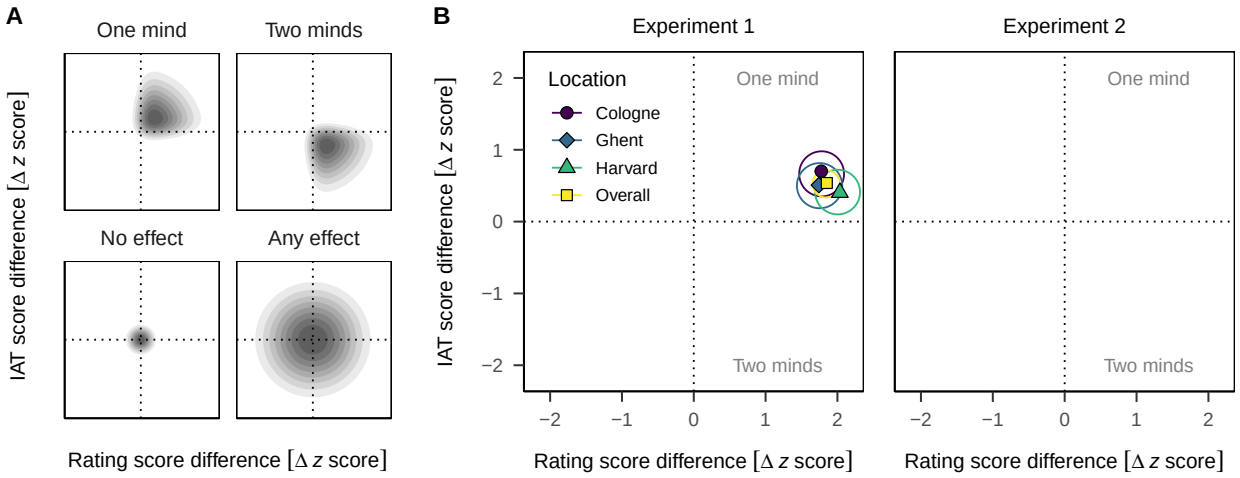

Figure 2. Predictions of the four models of primary interest (**A**) and results of Experiment 1 and Experiment 2 (**B**). Black-rimmed points represent means of observed attitude differences between learning blocks with positive and negative learned behaviors. Positive values indicate that attitudes correspond to the valence of learned behaviors and contradict the valence of the briefly presented USs. Ellipses represent 95% Bayesian credible intervals based on the unconstrained model $\mathcal{M}_{\text{Any effect}}$.