Contrast Effects in Backward Evaluative Conditioning:

Exploring Effects of Affective Relief/Disappointment versus Instructional Information

Luke J. S. Green

Curtin University

Camilla C. Luck

Curtin University

Bertram Gawronski

University of Texas at Austin

Ottmar V. Lipp

Curtin University

Word count: 8353

Author Notes

This work was supported by an Australian Government Research Training Program Scholarship to Luke Green and grants DP180111869 and SR120300015 from the Australian Research Council to Ottmar Lipp.

Correspondence concerning this article should be sent to: Luke J S Green, School of Psychology, Curtin University, GPO Box U1987 Perth WA 6845, Australia. Email: luke.green2@postgrad.curtin.edu.au.

Abstract

Past studies of backward evaluative conditioning (EC) have found an assimilation effect, in that neutral conditional stimuli (CS) were found to acquire the valence of co-occurring unconditional stimuli (US). Recent studies employing a concurrent forward and backward conditioning paradigm with instructions suggesting a contrastive relation between the US and the backward CS have resulted in contrast effects, in that backward CSs acquired valence opposite to the US. The current research investigated whether these effects were in fact due to the instructions highlighting the contrastive relation between the US and CS, or whether affective relief/disappointment experienced at US offset could account for this result. Consistent with the hypothesized role of instructions, backward CS contrast effects occurred only when instructions highlighted the valence of the US and attributed control of that US to the CSs. In contrast to the affective relief/disappointment hypothesis, no backward CS contrast effects were found without such instructions.

Keywords: associative learning, backward conditioning, evaluative conditioning, attitudes, propositional learning

How we evaluate people, events, and stimuli has been shown to influence interpersonal relationships, voting behaviour, consumer behaviour, and career aspirations (e.g., Galdi, Arcuri, & Gawronski, 2008; Gibson, 2008; LeBel & Campbell, 2009). One method through which these evaluations are acquired and changed is known as *evaluative conditioning* (EC), which occurs when the evaluation of a neutral conditional stimulus (CS) is changed by its co-occurrence with a valenced unconditional stimulus (US; De Houwer, 2007). Prominent examples of EC include advertising campaigns that present a consumer product (CS) with a well-liked celebrity (US), leading to the product becoming positive. EC is of great interest to psychologists, because encountering co-occurring stimuli of differing valence is un-avoidable and ever-present in our daily lives.

Although past EC research has predominantly found assimilation effects (i.e., CS-US pairings produce CS evaluations that are in line with the valence of the co-occurring US), some studies have found contrast effects under certain conditions (i.e., CS-US pairings produce CS evaluations that are opposite to the valence of the co-occurring US). The main goal of the current research was to investigate the contribution of relief/disappointment learning and instructions about CS-US relations to the emergence of contrast effects in EC. We were particularly interested in whether relief/disappointment experienced at the offset of valenced USs could account for previously obtained contrast effects that have been interpreted to be the result of instructions about CS-US relations (Moran & Bar-Anan, 2013). If so, this would provide further support for the notion that EC and fear conditioning may share a common underlying mechanism.

Evaluative Conditioning

In a typical EC procedure, participants are presented with neutral stimuli (e.g., images of geometric shapes; CSs). Some of the neutral stimuli are presented together with pleasant stimuli (e.g., images of puppies; USpos), while others are presented together with unpleasant stimuli (e.g., images of snakes; USneg). After repeated pairings, evaluations of the neutral stimuli paired with pleasant stimuli tend to become more positive, whereas evaluations of the neutral stimuli paired with unpleasant stimuli tend to become more negative (De Houwer, Thomas, & Baeyens, 2001). This effect

is quite robust, in that it has been shown with stimuli from various modalities (i.e., visual, auditory, olfactory), when both the CSs and USs are from the same or different modalities, and when the CSs are presented with the same or different USs across trials (for a meta-analysis, see Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010).

EC-related changes in CS valence can be measured using either explicit or implicit measures. Explicit measures of CS valence involve asking participants to rate how much they like the CSs, or how pleasant they find the CSs (self-reported valence ratings). Implicit measures of CS valence, such as the implicit association test (IAT; Greenwald, McGhee, & Schwartz, 1998) or affective priming (Fazio, Jackson, Dunton, & Williams, 1995), are performance-based measures that infer CS evaluations from the speed of categorisation responses on different kinds of trials (De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Gawronski & De Houwer, 2014). Implicit measures exploit the fact that categorisation tends to be faster on valence-congruent trials than valence-incongruent trials. In affective priming, for example, responses to pleasant target words tend to be faster when they are preceded by a positive prime stimulus than when they are preceded by a negative prime stimulus (Fazio et al., 1995). In research on EC, implicit and explicit measures typically reveal similar patterns of results (e.g., Mallan, Lipp, & Libera, 2008; Olson & Fazio, 2001). However, as we will discuss below, dissociations between explicit and implicit measures have been found to emerge under certain circumstances (e.g., Hu, Gawronski, & Balas, 2017a, 2017b; Moran & Bar-Anan, 2013).

Forward vs. Backward Conditioning

CS-US pairings can differ in terms of the sequence in which a CS co-occurs with a US. *Forward conditioning* involves cases in which a CS precedes a US (CS-US); *backward conditioning* involves cases in which a CS follows a US (US-CS); and *simultaneous conditioning* involves cases in which a US appears simultaneously with a CS (CS+US). Mallan et al. (2008) used a between-subjects design to compare EC effects in forward, backward, and simultaneous conditioning paradigms, using geometric shapes as CSs, and valenced pictures as USs. Participants were instructed to pay attention to the pictures, as they would be asked questions about them at the end of the experiment. After

conditioning, explicit valence ratings and affective priming revealed similar EC effects in all groups, such that CSs paired with positive USs became more positive than CSs paired with negative USs. These results suggest that both forward and backward conditioning lead to assimilation effects, such that the CS acquires the valence of the US it was paired with (see also Kim, Sweldens, & Hütter, 2016). This assimilation effect was evident on both explicit and implicit measures (see also Hofmann et al., 2010).

However, different from Mallan et al.'s findings (2008), Moran and Bar-Anan (2013) found a dissociation between explicit and implicit measures for backward conditioning in a study that compared forward and backward conditioning using a within-subjects design (CS-US-CS). On positive trials of the EC task, participants were presented with an image of one member of a family of alien creatures (forward CSpos), followed by a pleasant melody (USpos), and an image of a member of a second family of alien creatures (backward CSpos). On negative trials of the EC task, participants were presented with an image of a member of a third family of alien creatures (forward CSneg), followed by an aversive human scream (USneg), and an image of a member of a fourth family of alien creatures (backward CSneg). Before the EC task, participants were told that each alien family (CSs) had a different role to play: start the melody, stop the melody, start the human scream, or stop the human scream. Participants were instructed to learn the role of each family of aliens for a memory test at the end of the study. For forward CSs, Moran and Bar-Anan (2013) found assimilation effects on both explicit and implicit measures of CS valence. That is, CSs that preceded the pleasant melody were found to be more positive than CSs that preceded the unpleasant scream. For backward CSs, however, explicit and implicit measures revealed different results. Whereas implicit measures of backward CS valence showed assimilation effects (i.e., CSs presented after the pleasant melody elicited more favourable responses than CSs presented after the unpleasant scream), explicit measures of backward CS valence revealed contrast effects (i.e., CSs presented after the unpleasant scream were rated more positively than the CSs presented after the pleasant melody).

Moran and Bar-Anan (2013) interpreted the obtained dissociation in terms of two functionally distinct learning processes underlying evaluations on implicit and explicit measures. Drawing on the associative-propositional evaluation (APE) model (Gawronski & Bodenhausen, 2006, 2011, 2018), they suggested that implicit measures are more sensitive to effects of associative learning, reflecting the mere co-occurrence of a CS and a US regardless of their relation. In contrast, explicit measures were assumed to be more sensitive to effects of propositional learning, reflecting the particular relation between a CS and a co-occurring US. However, different from this interpretation in terms of two functionally distinct learning mechanisms, recent research suggests that the observed backward CS assimilation effect on implicit measures might be due to factors during the measurement of CS evaluations, in that backward CS contrast effects occur on both implicit and explicit measures when these factors are controlled (see Bading, Stahl, & Rothermund, 2019; Hu et al., 2017a; Moran & Bar-Anan, 2019; for a review, see Corneille & Stahl, 2019). These findings shift the explanatory burden from the reported dissociation between implicit and explicit measures to the question of what causes the backward CS contrast effects observed by Moran and Bar-Anan (2013).

According to Moran and Bar-Anan (2013), the backward CS contrast effect observed in their study is the direct result of the instructional manipulation, which stated that the backward CSs would stop the preceding USs. An alternative mechanism that may account for the backward CS contrast effects reported by Moran and Bar-Anan (2013) without the need for an instructional manipulation is affective relief/disappointment. Research on relief learning has shown that presenting a CS at the offset of an aversive stimulus (US-CS) can result in this backward CS gaining positive valence (Andreatta, Mühlberger, Yarali, Gerber, & Pauli, 2010; Gerber et al., 2014; Luck & Lipp, 2017). This occurs because feelings of relief from the aversive stimulus ending become conditioned to the backward CS, which is presented simultaneously with feelings of relief. Preliminary research suggests that the same contrastive process also occurs at the offset of a positive stimulus, resulting in disappointment (Green, Luck, & Lipp, submitted). Although Andreatta et al. (2010) investigated contrast effects in fear conditioning rather than evaluative conditioning, these findings raise the

possibility that affective relief/disappointment may also affect the direction of backward EC effects. If the negative/positive US used by Moran and Bar-Anan (2013) was sufficiently aversive/pleasant to drive relief/disappointment learning, affective relief/disappointment may be sufficient to explain the backward CS contrast effect in their study. Although the boundary conditions of relief and disappointment learning are still not well understood, it is conceivable that affective relief/disappointment at the offset of the aversive/pleasant US in Moran and Bar-Anan (2013) may have contributed to the observed backward CS contrast effects.

A second alternative explanation is that Moran and Bar-Anan (2013) assessed forward and backward conditioning concurrently within subjects (CS-US-CS), whereas Mallan et al. (2008) assessed forward and backward conditioning between groups (CS-US vs. US-CS). Research by Andreatta, Mühlberger, Glotzbach-Schoon, and Pauli (2013) suggests that the presence of a forward CS may moderate backward conditioning effects. These researchers found that making an aversive electro-tactile US predictable by presenting a forward CS resulted in positive explicit valence ratings of a backward CS paired with this US, but a backward CS was rated negatively when no forward CS was presented (see also Andreatta & Pauli, 2017). This occurs because the forward CS becomes more aversive when the onset of the US is predictable. The result is a larger discrepancy between the conditioned valence of the forward and backward CSs, possibly making the backward CS appear to be the opposite valence of the forward CS. This finding was replicated by Green et al. (submitted). Using both positive and negative USs and the same stimuli and instructions as Moran and Bar-Anan (2013), Green et al. found assimilation effects for backward CSs when no forward CS was present. Thus, counter to Moran and Bar-Anan's (2013) argument that the backward CS contrast effect in their study is the result of instructions about contrastive CS-US relations, the concurrent forward and backward conditioning procedure may be conducive to backward CS contrast effects without any instructions about contrastive CS-US relations. In either of the aforementioned cases, contrast effects should be observed using a similar paradigm without the instructions employed by Moran and Bar-Anan (2013).

Based on these considerations, the main goal of the current research was to investigate whether presenting within-subjects forward and backward conditioning without instructions could result in backward CS contrast effects as a result of the combination between US predictability and affective relief/disappointment.[1]

Pilot Studies

In addition to the main experiment reported below, we conducted two pilot studies to determine the instructions and experimental parameters required to address whether within-subjects forward and backward conditioning and affective relief/disappointment alone would elicit backward CS contrast effects (for more details, see Supplementary Materials). Pilot Study 1 compared the instructions used by Moran and Bar-Anan (2013) and Mallan et al. (2008) in a within-subjects forward and backward conditioning picture-picture paradigm using stimulus presentation and timing parameters adapted from Mallan et al. (2008). No evidence of backward CS contrast effects on explicit measures were found in either group. Thus, we performed a direct replication of Moran, Bar-Anan, and Nosek (2016) to ensure that backward CS contrast effects could be obtained in a picture-picture paradigm using instructions. The instructions used in this second pilot study differed from the ones in the first pilot study, in that they highlighted the agency of the CSs in controlling the US (i.e., the CSs control which event happens to you, either gold bars or garbage) and the valence of the outcome (i.e., whether this event is happy [gold bars] or sad [garbage]). We also investigated whether the lack of backward CS contrast effects observed in Pilot Study 1 was the result of presenting explicit valence ratings and affective priming before the learning phase. It is possible that evaluating stimulus valence before the learning phase puts participants in an 'evaluative mindset', thus resulting in amplified or erroneous effects that could strengthen assimilative EC, and thereby conceal potential contrast effects (Gast & Rothermund, 2011). Pilot Study 2 revealed that backward CS contrast effects could in fact be obtained in a picture-picture paradigm using these instructions, regardless of whether explicit valence ratings and affective priming were presented before the learning phase. These results

[1] All materials, data, and analysis files are available at https://osf.io/ur5kd/.

suggest that an ‘evaluative mindset’ was not responsible for the lack of backward CS contrast effects observed in Pilot Study 1.

Main Experiment

The primary aim of the main experiment was to determine whether affective relief/disappointment and US predictability without instructions would be sufficient to elicit backward CS contrast effects. A secondary aim was to test whether the ‘valence-agency’ instructions adopted from Moran et al. (2016) in Pilot Study 2 drive backward CS contrast effects when compared with the ‘observe instructions’ used in Pilot Study 1. If this were the case, it would suggest that an emphasis on ‘valence’ and ‘agency’ is essential for backward CS contrast effects in the picture-picture paradigm of Pilot Study 2. Another secondary aim was to determine whether the lack of backward CS contrast effects in both groups in Pilot Study 1 was due to features of the ‘Mallan paradigm’ not supporting backward CS contrast learning, regardless of instructions. It is possible that overlap between the US and the backward CS may assist in highlighting the fact that the backward CS controls the offset of the US, as there is generally overlap between stimuli when one of them is responsible for stopping an event (i.e. a good Samaritan intervening to bring resolution to an altercation between two parties). Without this, backward CS contrast learning may not be possible. In addition to this, the variability of the US may increase the affective relief/disappointment experienced at the offset of the US, because varying the US duration makes it difficult to predict when the US is going to end. Thus, without US offset being unpredictable, backward CS contrast learning may be less likely.

Method

Participants and design. Participants were recruited through M-Turk using TurkPrime after the Curtin University Human Research Ethics Committee approved this research protocol (Litman, Robinson, & Abberbock, 2017). The sample comprised 194 participants after duplicates and those failing to complete the experiment were removed (n = 33). The sample size was based on previous research to detect the within-subjects interaction of interest for each group (Moran & Bar-Anan, 2013; Moran et al., 2016). Moran and Bar-Anan (2013) and Moran et al. (2016) had sample sizes ranging

from 32 to 68 participants. In these studies, the within-subjects interaction of interest yielded large effects sizes between $\eta_p^2 = .15$ and $\eta_p^2 = .60$. Based on these effect sizes, we anticipated that approximately 50 participants per group would provide sufficient power to detect the effects of interest. The 'observe instructions, Mallan paradigm' group consisted of 49 participants (27 female), the 'observe instructions, Moran paradigm' group comprised 48 (24 female), the 'valence-agency instructions, Mallan paradigm' group comprised 50 (28 female), and the 'valence-agency instructions, Moran paradigm' group comprised 47 participants (21 female). Five participants failed to provide demographic information. The mean age of the 189 participants who provided demographic information was 36.35, *SD* = 11.085. Groups did not differ on gender, χ^2 (3) = 2.220, $p = .528$, ethnicity, χ^2 (15) = 13.029, $p = .600$, or age, F (1, 185) = .012, $p = .913$, $\eta_p^2 = .000$, $BF_{incl} = 0.23$.

Explicit valence ratings. In the 'Mallan paradigm' groups, each CS was presented one-by-one and participants were asked to rate how pleasant they found the stimulus on a 9-point scale ranging from 1 (*unpleasant*) to 9 (*pleasant*). In the 'Moran paradigm' groups, each CS family was presented alone and participants were asked "Based on your very first emotional response, how much do you like the creatures in the picture? Click the appropriate answer below: dislike strongly, dislike moderately, dislike slightly, like slightly, like moderately, like strongly".

Affective priming task. In the 'Mallan paradigm' groups, each of the four CSs were presented once with 10 positive target words and 10 negative target words for a total of 80 trials. In the 'Moran paradigm' groups, two creatures from each family were presented with positive and negative words twice, and two creatures from each family were presented with positive and negative words three times, for a total of 10 positive and 10 negative word pairings per family. This resulted in 80 trials. For both groups, a fixation cross was presented for 500ms, followed by the CS prime for 200ms, and then the target word until the participant provided their response. Participants were instructed to press the *I* key if the target word was positive and the *E* key if the target word was negative. Target words were taken from Hu et al. (2017a, 2017b). The positive words were *pleasant, good, outstanding,*

beautiful, magnificent, marvellous, excellent, appealing, delightful, and *nice*. The negative words were *unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly.*

Recollective Memory Test. For exploratory purposes, the current study also included measures of recollective memory. In the 'Mallan paradigm' groups, participants were shown each CS and asked: "Circle the appropriate answer below. Was this picture presented: Together with pleasant pictures, together with unpleasant pictures, together with pleasant and unpleasant pictures, I did not see this picture, I could not tell?" In the 'Moran paradigm' group, participants were shown each CS and asked: "Circle the appropriate answer below. What is the role of this creature: To start pleasant pictures, to stop pleasant pictures, to start unpleasant pictures, to stop unpleasant pictures?" Using the sum of correct responses on the memory test, accuracy scores on the test could range from zero to four. Both groups were also presented with each US and each CS, and asked to indicate which CS came before or after each US. This procedure resulted in an accuracy score ranging from 0 to 16. Participants were classified as remembering the CS-US contingencies only if they scored 100% on both memory tests. In the 'Moran paradigm' groups, each CS family was presented alone and participants were asked "In the game, what was the role of the creatures in the picture? Click the appropriate answer below: Starting gold, starting garbage, stopping gold, stopping garbage?" The analysis of the recollective memory data did not add substantially to the current report, and is available in the Supplementary Materials.

Demographics questionnaire. Participants were asked to report their age, gender, and ethnicity, and to provide information about the environment in which they completed the task, and if they had any comments.

Apparatus/stimuli. In the 'Mallan paradigm' groups, four images of aliens, one from each of the four families of alien creatures created by Moran and Bar-Anan (2013), were used as CSs (see below; materials from Moran and Bar-Anan, 2013, available at https://osf.io/cqsnj/). Each alien differed in colour and head shape. Four positive and four negative pictures from the International Affective Picture System (IAPS; CSEA, 1999) were used as USs (1050, 1300, 1440, 1710, 5833,

6313, 6560, and 8190). In the 'Moran paradigm' groups, CSs and USs were those used by Moran et al. (2016; available at https://osf.io/v2trw/). CSs were four families of alien creatures, with each family comprising four creatures for a total of 16 CSs. The positive US was a picture of puppies, gold bars, and a baby, presented next to each other as a single image, and the negative US was a picture of an aggressive dog, garbage, and a crying child presented next to each other as a single image. Inquisit 4 Web by Millisecond Software ™ (2016) was used to run the experiment and to record responses in all tasks.

Procedure. Participants selected the HIT (human intelligence task) on M-Turk and read the description of the study. When participants began the study, they were presented with an information sheet outlining the tasks, informed that they could withdraw at any time by pressing 'ctrl + q', and then prompted to press 'continue' if they consented to participate. Informed consent was implied if participants pressed 'continue'. In all groups, the first explicit valence ratings and affective priming task was presented followed by the training phase. In the 'Mallan paradigm' groups, the training phase comprised 12 positive and 12 negative trials presented pseudo-randomly, with inter-trial intervals of 4, 6, and 8 seconds. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was adapted from Moran and Bar-Anan (2013), with some modifications based on Mallan et al. (2008). We used one CS from each of the four alien families, four positive and four negative pictures as USs, and each stimulus was presented for 4 seconds with onset and offsets coinciding (i.e., no overlap between CSs and USs). CSs were counter-balanced using a Latin square resulting in four CS orders, with each CS occurring in each role equally. In the 'Moran paradigm' groups, the training phase comprised 12 positive and 12 negative trials randomly presented with inter-trial intervals of 2 seconds. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was an exact replication of Moran et al. (2016). CSs were presented for 1.5 seconds and USs were presented in blocks of 1s flashes with a 200ms break between each flash for a total of 3 or 5 seconds of total US presentation time. Onset of the US coincided with offset of the forward CS, and onset of the backward

CS occurred 200ms after the last US appearance. One group in each of the ‘Mallan paradigm’ and ‘Moran paradigm’ groups received the *valence-agency* instructions and one group received the *observe* instructions.

In the *valence-agency* instructions groups, participants received the following instructions from Moran et al. (2016) before the training phase:

> *In the next game, you will get piles of shiny gold bars, but also some stinky garbage piles. Getting gold bars is a happy event, whereas getting garbage piles is a sad event. In the game, four families of creatures control whether happy or sad events happen to you. These are the four families. One family of creatures will always start the gold bars coming your way. A second family of creatures will always stop the gold bars. A third family of creatures will always start garbage piles coming your way. A fourth family of creatures will always stop the garbage piles. Your goal in this game is to learn which family of creatures starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage. We will test your learning later in the game, so please pay close attention. If you read and understood the instructions, hit the spare bar to continue. Please pay close attention to the images on the screen. Make sure you learn and remember which family does each of the four actions (start gold, stop gold, start garbage, stop garbage). Press space to start the game.*

After 12 trials, the following instructions were presented:

> *Do you know by now which family starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage? Try to memorize what each family does for a later test. Press space for a few more rounds to help you remember the roles of the families better.*

In the *observe* instructions group, participants received the following instructions adapted from Mallan et al. (2008):

> *In this task you will be presented with a series of pictures. Please pay attention to which pictures follow each other as you will be tested on this at the end of the experiment.*

After the training phase, the second explicit valence ratings and affective priming task was presented, followed by the memory test and demographics questionnaire. Participants then received a completion code to receive their compensation, and were thanked for their participation. The experiment took approximately 20 minutes on average to complete, and participants were compensated US-$5.70.

Statistical analyses. Frequentist analyses were performed using IBM SPSS Statistics 25. We also report the results Bayesian analyses conducted in JASP 0.10.0.0 to supplement the frequentist analyses. BF_{10} values from the model comparison are reported for main effects, and $BF_{inclusion}$ (BF_{incl}) values from the effects analysis (across matched models) are reported for interactions. The $BF_{inclusion}$ (across matched models) compares models that contain the effect of interest with equivalent models that have had the effect of interest removed. The result is a model that provides only the effect of the interaction of interest without contributions from lower order effects (known as the Baws approach; see Mathôt, 2017, for a discussion).

The explicit valence ratings in the 'Moran paradigm' groups were transformed from a 6-point scale to a 9-point scale ([X – 1] * 1.6 + 1), so that ratings could be compared with the 'Mallan paradigm' groups. EC scores were calculated as the difference between ratings of CSs paired with positive USs and ratings of CSs paired with negative USs. EC scores were calculated separately for forward vs. backward conditioning and for pre-training vs. post-training. Positive EC scores represent an assimilation effect and negative EC scores represent a contrast effect. In the affective priming task, trials on which target words were categorised incorrectly were scored as error trials. Trials on which reaction times were shorter than 300ms and longer than 1000ms were categorised as outliers, as they were deemed to be outside the window of a valid response (see Koppehele-Gossel, Hoffmann, Banse, & Gawronski, in press). Participants with a percentage of invalid trials greater than 25% on the affective priming task were excluded from the priming analyses ('observe instructions, Mallan paradigm', n = 12, 'observe instructions, Moran paradigm', n = 11, 'valence-agency instructions, Mallan paradigm', n = 5, 'valence-agency instructions, Moran paradigm', n = 7). In the final sample at

pre-test, 7.47% of trials were incorrect categorisations of target words and 6.85% of trials were outliers. At post-test, 7.82% of trials were incorrect categorisations of target words and 8.60% of trials were outliers. For the 'Moran paradigm' groups, responses following CSs within the same family in the affective priming task were averaged to provide overall means for each family. Priming scores of EC effects were calculated as the difference in response times between incongruent and congruent trials: (CSs paired with positive USs/negative target words + CSs paired with negative USs/positive target words) – (CSs paired with positive USs/positive target words + CSs paired with negative USs/negative target words). Priming scores were calculated separately for forward vs. backward conditioning and for pre-training and post-training. Positive priming scores suggest an assimilation effect, while negative scores suggest a contrast effect. EC scores from explicit valence ratings and affective priming scores were subjected to separate Frequentist and Bayesian 2 (Instructions: observe vs. valence-agency instructions; between-participants) × 2 (Paradigm: Mallan paradigm vs. Moran paradigm; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participants) × 2 (Time: pre-test vs. post-test; within-participants) mixed ANOVAs. Significant interactions from the Frequentist analyses were followed-up with pairwise comparisons and one sample *t*-tests where appropriate. Pillai's trace values of the multivariate solution are reported for main effects and interactions ($\alpha = .05$). These analyses were also performed as two-sided paired and independent samples Bayesian *t*-tests using the default settings in the BayesFactor package in R. BF_{10} values are reported for all Bayesian follow-up analyses. The reliability of the priming task was $\alpha = .11$ at pre-test and $\alpha = .21$ at post-test. The analyses of the error data from the affective priming task did not add substantially to the current report and are available in the Supplementary Materials.

Results

Explicit valence ratings. Figure 1 shows mean EC scores based on explicit valence ratings for forward and backward conditioning measured pre-training and post-training as a function of Instructions and Paradigm. The figure suggests assimilation effects for forward conditioning at post-training for all groups, and contrast effects for backward conditioning at post-training for the valence-

agency instructions group only, regardless of paradigm. The ANOVA revealed significant main effects of Conditioning Type, F (1, 190) = 119.82, p < .001, η_p^2 = .387, BF_{10} = 3.30 × 10^{17}, and Time, F (1, 190) = 93.24, p < .001, η_p^2 = .329, BF_{incl} = 1.29 × 10^{10}, a significant two-way interaction between Instructions and Conditioning Type, F (1, 190) = 60.53, p < .001, η_p^2 = .242, BF_{incl} = 3.28 × 10^{12}, a significant two-way interaction between Conditioning Type and Time, F (1, 190) = 186.36, p < .001, η_p^2 = .495, BF_{incl} = 2.67 × 10^{33}, and a significant two-way interaction between Paradigm and Time, F (1, 190) = 6.41, p = .012, η_p^2 = .033, BF_{incl} = 2.15, which were qualified by a significant three-way interaction between Instructions, Conditioning Type, and Time, F (1, 190) = 65.84, p < .001, η_p^2 = .257, BF_{incl} = 1.41 × 10^{13}. The four-way interaction between Instructions, Paradigm, Conditioning Type, and Time was not significant, F (1, 190) = 0.10, p = .749, η_p^2 = .001, BF_{incl} = 0.12. Decomposing the three-way interaction, follow-up analyses revealed that, for forward conditioning, EC scores for valence-agency instructions were significantly larger than EC scores for observe instructions at post-training, F (1, 190) = 53.94, p < .001, η_p^2 = .221, BF_{10} = 9.43 × 10^8, but not pre-training, F (1, 190) = 0.01, p = .924, η_p^2 < .001, BF_{10} = 0.16. In contrast, for backward conditioning, EC scores for valence-agency instructions were significantly smaller than EC scores for observe instructions at post-training, F (1, 190) = 41.51, p < .001, η_p^2 = .179, BF_{10} = 11.71 × 10^6, but not pre-training, F (1, 190) = 0.02, p = .879, η_p^2 < .001, BF_{10} = 0.16. One-sample t-tests further indicated that post-training EC scores in the observe instruction groups were significantly larger than zero for forward conditioning, t (96) = 6.96, p < .001, d = 0.71, BF_{10} = 22.96 × 10^6, and backward conditioning, t (96) = 2.66, p = .009, d = 0.27, BF_{10} = 3.13. In contrast, post-training EC scores in the valence-agency instruction groups were larger than zero for forward conditioning, t (96) = 19.23, p < .001, d = 1.95, BF_{10} = 2.47 × 10^{31}, and significantly smaller than zero for backward conditioning, t (96) = 6.11, p < .001, d = 0.62, BF_{10} = 53.77 × 10^4. EC scores for forward and backward conditioning did not significantly differ from zero at pre-training for any of the four groups, all ts < 1.30, all ps > .196, all ds < 0.13, BF_{10}s < 0.25. Decomposing the significant two-way interaction between paradigm and time, follow-up analyses showed that EC scores for the Mallan paradigm tended to be larger than

EC scores for the Moran paradigm at post-training, F (1, 190) = 3.32, p = .070, η_p^2 = .017, BF_{10} = 0.23, but not pre-training, F (1, 190) = 2.69, p = .103, η_p^2 = .014, BF_{10} = 0.57.

Affective priming. Figure 2 shows mean EC scores based on affective priming for forward and backward conditioning measured pre-training and post-training as a function of instructions and paradigm. The figure suggests an assimilation effect at post-training in the valence-agency instructions/Mallan paradigm group and a baseline priming score larger than zero in the observe instructions/Moran paradigm group. A marginal main effect of Time, F (1, 153) = 3.80, p = .053, η_p^2 = .024, BF_{10} = 1.11, was qualified by a marginal interaction between Instructions and Time, F (1, 153) = 3.82, p = .053, η_p^2 = .024, BF_{incl} = 0.92. The four-way interaction between Instructions, Paradigm, Conditioning Type, and Time was not significant, F (1, 153) = 0.20, p = .888, η_p^2 < .001, BF_{incl} = 0.22. Follow-up analyses revealed that priming scores in the valence-agency instructions group were significantly larger at post-training compared to pre-training, F (1, 153) = 8.41, p = .004, η_p^2 = .052, BF_{10} = 5.68. There was no significant difference between pre-training and post-training in the observe instructions group, F (1, 153) < 0.01, p = .998, η_p^2 < .001, BF_{10} = 0.09. One sample t-tests revealed that priming scores in the valence-agency instructions group were significantly larger than zero at post-training, t (85) = 3.15, p = .002, d = 0.34, BF_{10} = 13.95, but not pre-training, t (85) = 0.48, p = .633, d = 0.05, BF_{10} = 0.10. Moreover, priming scores in the observe instructions group were significantly larger than zero at pre-training, t (70) = 2.06, p = .044, d = 0.24, BF_{10} = 0.94, and marginally larger at post-training, t (70) = 1.73, p = .088, d = 0.21, BF_{10} = 0.49.

Discussion

The primary aim of this experiment was to determine whether affective relief/disappointment at the offset of an aversive/pleasant US when that US was predictable (CS-US-CS) could elicit backward CS contrast effects without the need for an instructional manipulation. Our secondary aim was to assess whether the absence of ‘valence’ and ‘agency’ components of the instructional manipulation could be responsible for the lack of backward CS contrast effects in our Pilot Study 1. Finally, we investigated whether the offset of the US was required to be unpredictable in order for

backward CS contrast effects to emerge (assessed by comparing paradigms with and without US offset predictability).

For forward conditioning, explicit valence ratings revealed assimilation effects regardless of instructions. In contrast, for backward conditioning, explicit valence ratings showed assimilation effects for 'observe instructions' and contrast effects for 'valence-agency instructions.' Moderate evidence in support of the null hypothesis for the four-way interaction including 'paradigm' supports the conclusion that this pattern emerged in both the 'Moran paradigm' and 'Mallan paradigm' groups. Unexpected baseline differences in measures from the affective priming task make their interpretation difficult, although it seems that assimilation effects occurred regardless of conditioning type and more strongly in the 'valence-agency instructions' group. However, this conclusion should be regarded with caution, especially considering the Bayes factor for the interaction between instructions and time was inconclusive.

Overall, these findings clearly demonstrate that affective relief/disappointment at the offset of an aversive/pleasant stimulus when the US is predictable is not sufficient to elicit backward CS contrast effects without an instructional manipulation. Moreover, these findings show that backward CS contrast effects using picture USs occur when the instructions emphasize 'valence' and 'agency' (Moran et al., 2016), suggesting that the different results in Pilot Studies 1 and 2 are driven by differences in instructions. Finally, these findings demonstrate that backward CS contrast effects can occur regardless of whether there is overlap between CSs and USs, or variability in US duration. This suggests that unpredictable US offset and presenting stimuli in a manner that appears as if the CSs are controlling US onset and offset is not necessary to observe backward CS contrast effects in presence of contrastive instructive instructions that emphasize valence and agency.

As a caveat to these conclusions, we would like to note that post-hoc power analyses revealed that our main experiment was underpowered to detect potential higher-order interactions. Explicit valence ratings showed backward CS contrast effects for the valence-agency instruction groups and assimilation effects for the observe instruction groups. Greater statistical power may have permitted

the detection of a significant four-way interaction involving the factor ‘paradigm’, given that the difference between forward and backward conditioning under observe instructions at post-test was somewhat smaller in the Moran paradigm group compared with the Mallan paradigm group. However, this difference would have been driven by smaller forward conditioning rather than a difference in backward conditioning in the observe instructions groups. Because one of our aims was to determine whether US offset needed to be predictable in the valence-agency instructions groups for backward CS contrast effects to emerge and additional power seems unlikely to change the pattern of backward conditioning effects observed here, our conclusions appear valid despite the fact that the experiment was underpowered for the detection of a significant four-way interaction. Moreover, Bayesian analysis of the four-way interaction provided moderate support for the null hypothesis, suggesting that even with higher power this interaction would not be meaningful.

Low statistical power may have also contributed to the affective priming task yielding less reliable results than expected. Cronbach’s α ranged between .11 and .21, which is lower than expected for an affective priming task using this outlier treatment (Koppehele-Gossel et al., in press). The large confidence intervals of the mean suggest a lack of sensitivity, consistent with concerns that affective priming is more susceptible to measurement error than other implicit measures (Gawronski & De Houwer, 2014). In retrospect, the task may have benefited from more trials, especially in the ‘Moran paradigm’ conditions that involved the presentation of multiple exemplars of each CS family during conditioning. However, because affective priming scores largely followed the results of Moran and Bar-Anan (2013) and Hu et al. (2017a), it seems unlikely that our conclusions would have been different if stronger priming results had been observed. Nevertheless, future research with larger samples and greater trial numbers in the affective priming task may help to corroborate our conclusions.

The hypothesis that affective relief/disappointment may result in backward CS contrast effects was not supported. Moreover, the presence of the forward CS making the US predictable did not lead to a greater contrast between the valence of the forward and backward CSs, which may have resulted

in a backward CS contrast effect. To explain why the predicted backward CS contrast effect did not occur, we turn to findings from the pain relief literature and a study on elaborated encoding in EC by Fiedler and Unkelbach (2011). Studies on pain relief and relief learning suggest that the more intense or aversive the pain eliciting stimulus, the greater the amount of pain relief experienced at its offset (Andreatta et al., 2010; Bitar, Marchand, & Potvin, 2018). Moreover, Fiedler and Unkelbach (2011) showed that increasing the relevance of the relational qualifier to the participant resulted in more elaborate encoding of the propositional information about the relation between the CS and the US, which in turn led to contrast effects. Taken together, these findings suggest that USs of higher intensity may lead to deeper processing and more substantial encoding of CS-US relations, thus leading to larger backward CS contrast effects. While there was no relational qualifier present in the 'observe instructions' groups, it is plausible that the higher the intensity of the US, the higher the relevance and encoding of the US and its offset. This assumption suggests that if a threshold level of processing is not met (either due to a low US intensity or a lack of elaborate encoding of the US offset), contrast effects may not occur. By this reasoning, it stands that the USs employed in the main experiment may not have been intense enough to elicit affective relief/disappointment and/or may not have been relevant enough for participants to encode the offset of the US as an important event that would trigger an emotional response such as relief or disappointment.

The backward CS contrast effects observed on explicit valence ratings confirmed that the difference in results between Pilot Studies 1 and 2 were a result of including 'valence' and 'agency' components in the instructions. It is possible that the propositional information in Pilot Study 1 did not lend itself to sufficiently deep encoding to drive backward CS contrast effects, because the instructions lacked personal relevance to the participants. In the 'valence-agency' instructions, participants were told that the aliens controlled whether 'happy' or 'sad' events would happen specifically to them. This information is of greater personal relevance to participants, thereby increasing the salience of the propositional relation and, presumably, the depth at which this information is encoded. The instructions from Pilot Study 1 did not contain this level of specificity

toward the participant. Therefore, the combination of low intensity pictorial USs (as compared to the auditory USs used by Moran & Bar-Anan, 2013) and the instructional manipulation used may explain the lack of contrast effects in Pilot Study 1. Furthermore, combining these same USs with an instructional manipulation that highlights personal relevance and emphasizes 'valence' and 'agency' as in Pilot Study 2 did result in backward CS contrast effects.

The overlap of CSs with USs in the 'Moran paradigm' that made the CSs look like they had control over starting and stopping the USs was shown not to be a requirement for backward CS contrast effects. Moreover, varying the duration of the USs appeared to have no effect on backward CS learning. This further supports the notion that the results from Moran and Bar-Anan (2013) and Moran et al. (2016) are purely driven by the instructional manipulation, not by the appearance that CSs control US onset and offset, or because the offset of the US was unpredictable. Future research could investigate the importance of these parameters with more intense USs. It is possible that these parameters do not influence backward CS learning when a propositional mechanism is at play. However, it is possible that these parameters are important when affective relief/disappointment is engaged, as in the Andreatta et al. (2010; 2013) studies.

In summary, the current findings suggest that in a picture-picture paradigm affective relief/disappointment at the offset of an aversive/pleasant stimulus in the presence of a predictable US is not sufficient for backward CS contrast effects to occur. Rather, instructions determined whether backward CS valence ratings showed an assimilation or a contrast effect. These instructional manipulations are likely to interact with the properties of the US to influence CS evaluations during backward conditioning, possibly due to different levels of processing and encoding of propositional information about stimulus relations. The findings reported here clarify the effects of instructional manipulations and affective relief/disappointment in backward EC utilising picture-picture paradigms.

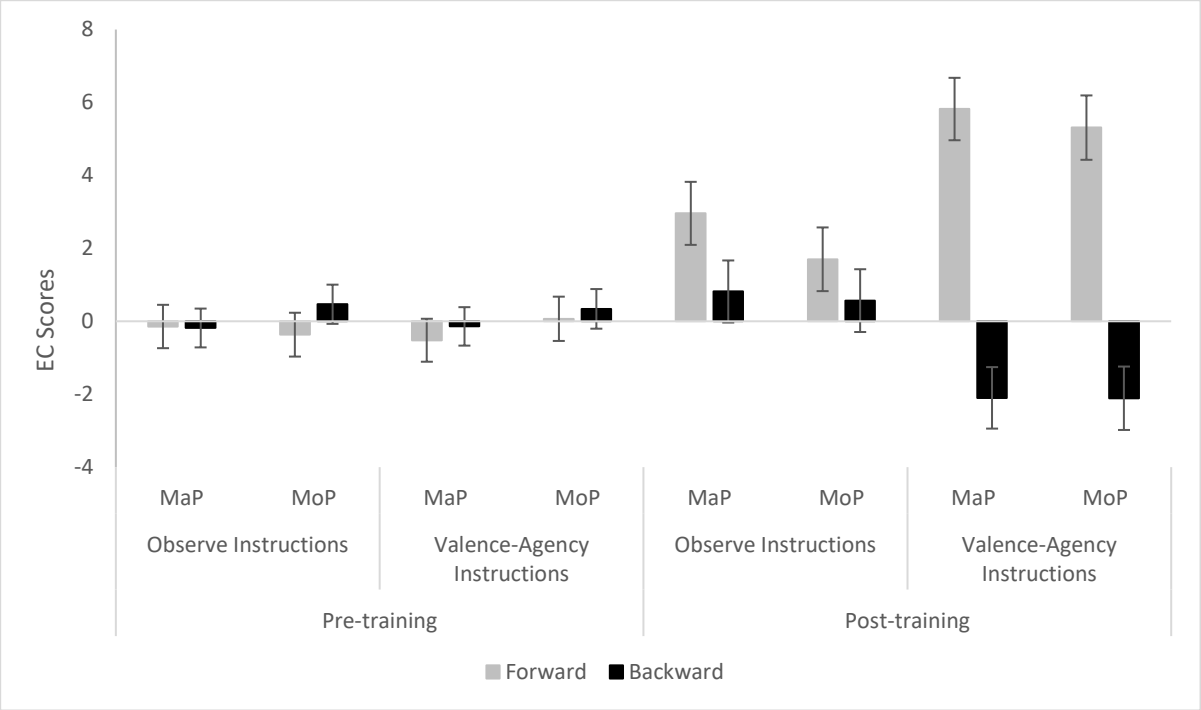

Figure 1. EC scores on explicit valence ratings for forward and backward conditioning measured pre-training and post-training as a function of instructions ('observe instructions' and 'valence-agency instructions') and paradigm ('Mallan paradigm (MaP)' and 'Moran paradigm (MoP)'). Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

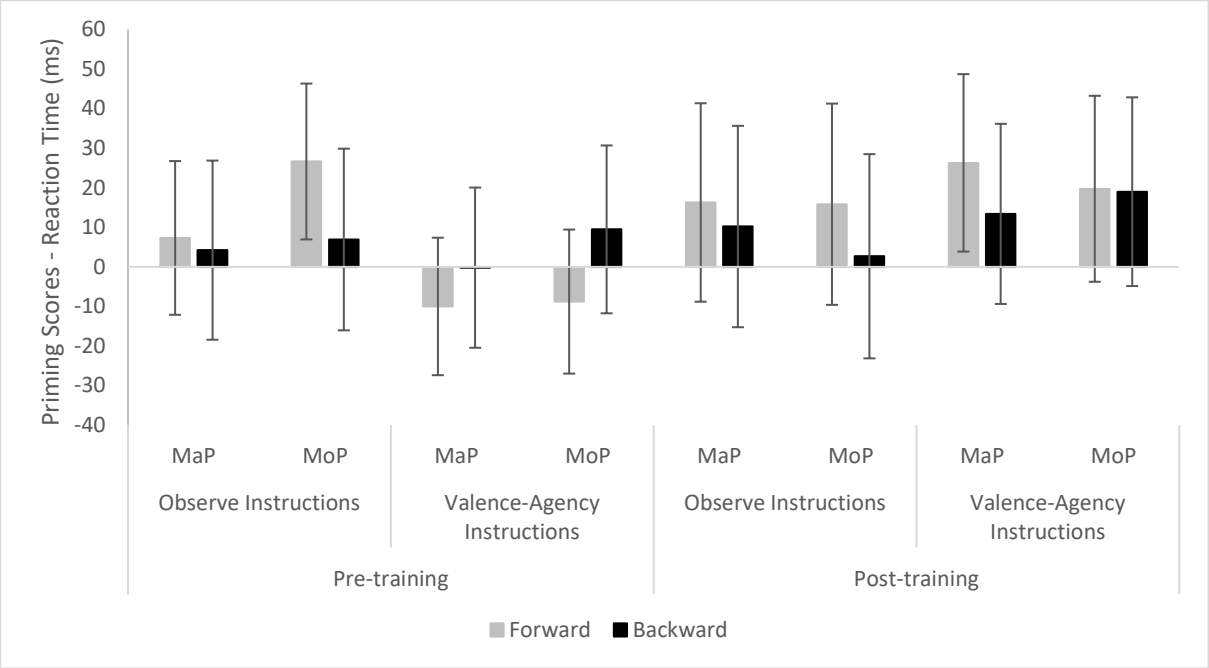

Figure 2. EC scores on affective priming for forward and backward conditioning measured pre-training and post-training as a function of instructions ('observe instructions' and 'valence-agency instructions') and paradigm ('Mallan paradigm (MaP)' and 'Moran paradigm (MoP)'). Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

References

Andreatta, M., Mühlberger, A., Glotzbach-Schoon, E., & Pauli, P. (2013). Pain predictability reverses valence ratings of a relief-associated stimulus. *Frontiers in Systems Neuroscience, 7,* 1-12.

Andreatta, M., Mühlberger, A., Yarali, A., Gerber, B., & Pauli, P. (2010). A rift between implicit and explicit conditioned valence in human pain relief learning. *Proceedings of the Royal Society of London B: Biological Sciences, 277,* 2411-2416.

Andreatta, M., & Pauli, P. (2017). Learning mechanisms underlying threat absence and threat relief: Influences of trait anxiety. *Neurobiology of Learning and Memory, 145,* 105-113.

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion.* doi:10.1080/02699931.2019.1604322.

Bitar, N., Marchand, S., & Potvin, S. (2018). Pleasant pain relief and inhibitory conditioned pain modulation: A psychophysical study. *Pain Research and Management, 1935056*.

Center for the Study of Emotion and Attention [CSEA – NIHM] (1999). *International affective picture system: Digitized photographs*. The Center for Research in Psychophysiology, University of Florida.

Corneille, O., & Stahl, C. (2019). Associative attitudes learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review, 23,* 161-189.

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

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135,* 347-368.

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127,* 853-869.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69,* 1013-1027.

Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition and Emotion, 25,* 639-656.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science, 321,* 1100-1102.

Gast, A, & Rothermund, K. (2011). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition and Emotion, 25,* 89-110.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132,* 692-731.

Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology, 44,* 59-127.

Gawronski, B., & Bodenhausen, G. V. (2018). Evaluative conditioning from the perspective of the associative-propositional evaluation model. *Social Psychological Bulletin, 13*(3), e28024.

Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (2nd edition, pp. 283-310). New York, NY: Cambridge University Press.

Gerber, B., Yarali, A., Diegelmann, S., Wotjak, C. T., Pauli, P., & Fendt, M. (2014). Pain relief learning in flies, rats, and man: Basic research and applied perspectives. *Learning and Memory, 21,* 232-252.

Gibson, B. (2008). Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research, 35,* 178-188.

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*, 1464-1480.

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

Hu, X., Gawronski, B., & Balas, R. (2017a). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43,* 17-32.

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8,* 858-866.

Inquisit 4 [Computer software]. (2016). Retrieved from https://www.millisecond.com.

Kim, J. C., Sweldens, S., & Hütter, M. (2016). The symmetric nature of evaluative memory associations: Equal effectiveness of forward versus backward evaluative conditioning. *Social Psychological and Personality Science, 7,* 61-68.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (in press). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology.*

LeBel, E. P., & Campbell, L. (2009). Implicit partner affect, relationship satisfaction, and the prediction of romantic breakup. *Journal of Experimental Social Psychology, 45,* 1291-1294.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49,* 433-442.

Luck, C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. *Psychophysiology, 54,* 673-683.

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. *International Journal of Psychophysiology, 69.* 9-17.

Mathôt, S. (2017) Bayes like a Baws: Interpreting Bayesian repeated measures in JASP. Retrieved from https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27,* 743-752.

Moran, T., & Bar-Anan, Y. (2019). The effect of co-occurrence and relational information on speeded evaluation. *Cognition and Emotion*. doi:10.1080/02699931.2019.1604321

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34.* 435-461.

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. *Psychological Science, 12,* 413-147.

Supplementary Materials: Pilot Study 1

To investigate whether affective relief/disappointment alone would elicit backward CS contrast effects in a within-subjects forward and backward conditioning procedure, we combined the concurrent forward and backward conditioning procedure and stimuli from Moran and Bar-Anan (2013) with the stimulus presentation parameters from Mallan et al. (2008). In addition, we manipulated between groups whether participants received the instructions used by Moran and Bar-Anan (2013; termed the *start-stop instructions* group) or the instructions used by Mallan et al. (2008; termed the *observe instructions* group). We expected to observe contrast effects for backward CSs and assimilation effects for forward CSs in both groups on explicit valence ratings. Based on the recent findings of Bading, Stahl, and Rothermund (2019), and Hu, Gawronski, and Balas (2017a), we expected to observe assimilation effects on implicit measures for forward and backward conditioning in both groups.

Method

Participants and design. Participants were recruited through Amazon Mechanical Turk (M-Turk) through TurkPrime (Litman, Robinson, & Abberbock, 2016). The sample comprised 94 participants (mean age = 35.67, *SD* = 9.84) after duplicates and participants failing to complete the experiment were removed (*n* = 19). The sample size was based on Moran and Bar-Anan (2013) and Moran, Bar-Anan, and Nosek (2016) who had sample sizes ranging from 32 to 68 participants. In these studies, the within-subjects interaction of interest yielded large effects sizes between $\eta_p^2 = .15$ and $\eta_p^2 = .60$. Based on these effect sizes, we anticipated that approximately 50 participants per group would provide sufficient power to detect the effects of interest. The start-stop instructions group consisted of 42 participants (17 female, 25 male); the observe instructions group included 52 participants (24 female, 28 male). Groups did not differ on gender, $\chi^2(1) = 0.31$, $p = .581$, ethnicity, $\chi^2(4) = 6.86$, $p = .144$, or age, $t(92) = 1.35$, $p = .179$, $d = 0.28$, $BF_{10} = 0.49$. A 2 (Instructions: start-stop vs. observe; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participants) × 2 (Time: pre-test vs post-test; within-participants) × 2 (US Valence: positive vs.

negative; within-participants) mixed design was used to determine the effect of instruction type on CS valence after a forward and backward conditioning procedure.

Explicit valence ratings. Each CS was presented one-by-one and participants were asked to rate how pleasant they found the stimulus on a 9-point scale ranging from 1 (*unpleasant*) to 9 (*pleasant*).

Affective priming task. Each of the four CSs was presented once with 10 positive target words and 10 negative target words for a total of 80 trials. A fixation cross was presented for 500ms, followed by the CS prime for 200ms, and then the target word until the participant provided their response. Participants were instructed to press the *I* key if the target word was positive and the *E* key if the target word was negative. Target words were taken from Hu et al. (2017a, 2017b). The positive words were *pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful,* and *nice*. The negative words were *unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly.*

Recollective Memory Test. For exploratory purposes, the current study also included measures of recollective memory. In the observe instructions group, participants were shown each CS and asked: "Circle the appropriate answer below. Was this picture presented: Together with pleasant pictures, together with unpleasant pictures, together with pleasant and unpleasant pictures, I did not see this picture, I could not tell?" In the start-stop instructions group, participants were shown each CS and asked: "Circle the appropriate answer below. What is the role of this creature: To start pleasant pictures, to stop pleasant pictures, to start unpleasant pictures, to stop unpleasant pictures?" Using the sum of correct responses on the memory test, accuracy scores on the test could range from zero to four. Both groups were also presented with each US and each CS, and asked to indicate which CS came before or after each US. This procedure resulted in an accuracy score ranging from 0 to 16. Participants were classified as remembering the CS-US contingencies only if they scored 100% on both memory tests. The analyses of the recollective memory data for Pilot Study 1 and all subsequent

experiments did not add substantially to the current report and are available in the additional analyses section below.

Demographics questionnaire. Participants were asked to report their age, gender, and ethnicity, and to provide information about the environment in which they completed the task, and if they had any comments.

Apparatus/stimuli. Four images of aliens, one from each of the four families of alien creatures created by Moran and Bar-Anan (2013), were used as CSs (see below; materials from Moran and Bar-Anan, 2013, available at https://osf.io/cqsnj/). Each alien differed in colour and head shape. Four positive and four negative pictures from the International Affective Picture System (IAPS; CSEA, 1999) were used as USs (1050, 1300, 1440, 1710, 5833, 6313, 6560, and 8190). Inquisit 4 Web by Millisecond Software ™ (2016) was used to run the experiment and to record responses in all tasks.

Procedure. Participants selected the HIT (human intelligence task) on M-Turk and read the description of the study. When participants began the study, they were presented with an information sheet outlining the tasks, informed that they could withdraw at any time by pressing 'ctrl + q', and then prompted to press 'continue' if they consented to participate. Informed consent was implied if participants pressed 'continue'. Next, the first explicit valence ratings and affective priming task was presented followed by the training phase. The training phase comprised 12 positive and 12 negative trials presented pseudo-randomly, with inter-trial intervals of 4, 6, and 8 seconds. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was adapted from Moran and Bar-Anan (2013), with some modifications based on Mallan et al. (2008). We used one CS from each of the four alien families, four positive and four negative pictures as USs, and each stimulus was presented for 4 seconds with onset and offsets coinciding. CSs were counter-balanced using a Latin square resulting in four CS orders, with each CS occurring in each role equally.

In the *start-stop instructions* group, participants received the following instructions adapted from Moran and Bar-Anan (2013) before the training phase:

> *In this task you are going to see two types of pictures. Pleasant pictures: Including pictures of scenery and animals. Unpleasant pictures: Including pictures of threat of violence and aggressive animals. These pictures will be shown multiple times. Before and after the pleasant and unpleasant pictures, four different types of creatures will appear on the screen. Each creature will have a fixed role out of four possible roles. Your task is to learn which role each creature plays. At the end of the experiment we will examine your memory of the role played by each creature. The four possible roles are: After the appearance of one creature, a pleasant picture appears. On the appearance of one creature, a pleasant picture ends. After the appearance of one creature, an unpleasant picture appears. On the appearance of one creature, an unpleasant picture ends. The four creatures that will start and stop the pictures are:*
>
> *The experiment is about to begin. Remember, you must learn the role of each of the creatures: Which creature starts the pleasant pictures? Which creature stops the pleasant pictures? Which creature starts the unpleasant pictures? Which creature stops the unpleasant pictures? At the end of the experiment we will examine what you have learned. To start the experiment, press the space bar.*

In the *observe instructions* group, participants received the following instructions adapted from Mallan et al. (2008):

> *In this task you will be presented with a series of pictures. Please pay attention to which pictures follow each other as you will be tested on this at the end of the experiment.*

After the training phase, the second explicit valence ratings and affective priming task was presented, followed by the memory test and demographics questionnaire. Participants then received a completion code to receive their compensation, and were thanked for their participation. The

experiment took approximately 20 minutes on average to complete, and participants were compensated US-$4.80.

Statistical analyses. Frequentist analyses were performed using IBM SPSS Statistics 25 and Bayesian analyses were performed using JASP 0.10.0.0. Bayesian follow-up analyses were performed using the BayesFactor package in R. For explicit valence ratings, EC scores were calculated as the difference between ratings of CSs paired with positive USs and ratings of CSs paired with negative USs. EC scores were calculated separately for forward vs. backward conditioning and for pre-training vs. post-training. Positive EC scores represent an assimilation effect and negative EC scores represent a contrast effect. In the affective priming task, trials on which target words were categorised incorrectly were scored as error trials. Trials on which reaction times were shorter than 300ms and longer than 1000ms were categorised as outliers, as they were deemed to be outside the window of a valid response (see Koppehele-Gossel, Hoffmann, Banse, & Gawronski, in press). Participants with a percentage of invalid trials larger than 25% were removed from the analyses (n = 5 in start-stop instructions group, n = 8 in observe instructions group). In the final sample at pre-test, 5.43% of trials were errors from incorrect categorisation of target words and 4.77% of trials were outliers. At post-test, 5.58% of trials were errors from incorrect categorisation of target words and 6.33% of trials were outliers. Priming scores were calculated as the difference in response times between incongruent and congruent trials: (CSs paired with positive USs/negative target words + CSs paired with negative USs/positive target words) – (CSs paired with positive USs/positive target words + CSs paired with negative USs/negative target words). Priming scores were calculated separately for forward vs. backward conditioning and for pre-training and post-training. Positive priming scores suggest an assimilation effect, while negative scores suggest a contrast effect. EC scores and priming scores from reaction time data were subjected to 2 (Instructions: start-stop vs. observe; between-participants) × 2 (Conditioning Type: forward vs. backward; within-participants) × 2 (Time: pre-test vs post-test; within-participants) mixed ANOVAs, and significant interactions were followed-up with pairwise comparisons and one sample t-tests where appropriate. Pillai's trace values of the multivariate solution

are reported for main effects and interactions ($\alpha = .05$). The reliability of the priming task was $\alpha = -.62$ at pre-test, and $\alpha = .16$ at post-test. The analyses of the error data from the affective priming task for Experiment 1 and all subsequent experiments did not add substantially to the current report and are available in the additional analyses section below.

Results

Explicit valence ratings. Mean EC scores are depicted in Figure S1. The ANOVA revealed significant main effects of Conditioning Type, F (1, 92) = 24.73, $p < .001$, $\eta_p^2 = .212$, $BF_{10} = 618.85$, and Time, F (1, 92) = 56.80, $p < .001$, $\eta_p^2 = .382$, $BF_{10} = 5.41 \times 10^9$, which were qualified by a significant two-way interaction between Conditioning Type and Time, F (1, 92) = 31.21, $p < .001$, $\eta_p^2 = .253$, $BF_{10} = 16794.55$, and a significant two-way interaction between Conditioning Type and Instructions, F (1, 92) = 8.93, $p = .004$, $\eta_p^2 = .088$, $BF_{10} = 13.65$. The three-way interaction between Instructions, Conditioning Type, and Time was not significant, F (1, 92) = 2.14, $p = .146$, $\eta_p^2 = .023$, $BF_{incl} = 0.52$. Follow-up analyses for the Conditioning Type × Time interaction revealed no difference between the two conditioning type conditions at pre-training, F (1, 92) = 0.01, $p = .912$, $\eta_p^2 = .000$, $BF_{10} = 0.12$, and a significantly larger EC score for forward conditioning than backward at post-training, F (1, 92) = 37.35, $p < .001$, $\eta_p^2 = .289$, $BF_{10} = 85889.24$. One sample t-tests confirmed that EC scores were significantly larger than zero for both forward conditioning, t (93) = 10.24, $p < .001$, $d = 1.06$, $BF_{10} = 9.50 \times 10^{13}$, and backward conditioning, t (93) = 2.51, $p = .014$, $d = 0.26$, $BF_{10} = 2.19$, at post-training, but not pre-training, t (93) = 0.31, $p = .759$, $d = 0.03$, $BF_{10} = 0.12$, and, t (93) = 0.73, $p = .470$, $d = 0.07$, $BF_{10} = 0.15$, respectively. Follow-up analyses for the Conditioning Type × Instructions interaction revealed that, averaged across pre- and post-test, EC scores for forward conditioning were significantly larger than EC scores for backward conditioning in the start-stop instructions group, F (1, 92) = 28.65, $p < .001$, $\eta_p^2 = .237$, $BF_{10} = 860.85$, but not the observe instructions group, F (1, 92) = 2.20, $p = .141$, $\eta_p^2 = .023$, $BF_{10} = 0.44$.

Affective priming. Mean affective priming scores are depicted in Figure S2. The ANOVA revealed only a significant main effect of Time, F (1, 79) = 6.29, $p = .014$, $\eta_p^2 = .074$, $BF_{10} = 2.79$,

indicating that priming scores were significantly larger at post-training than pre-training. Follow-up analyses revealed that priming scores were significantly larger than zero at post-training, t (80) = 3.02, p = .003, d = 0.34, BF_{10} = 5.43, but not pre-training, t (80) = 0.33, p = .746, d = 0.04, BF_{10} = 0.09. The three-way interaction between Instructions, Conditioning Type, and Time, was not significant, F (1, 79) = 0.11, p = .739, η_p^2 = .001, BF_{incl} = 0.26.

Discussion

Explicit valence ratings and reaction time priming scores revealed assimilation effects for forward and backward conditioning. Bayesian analyses provided weak to moderate evidence for the null hypothesis, suggesting that no differences on EC scores between instruction groups were present. Thus, the hypothesis that both groups would show backward CS contrast effects was not supported. It is possible that the 'start-stop instructions' are not sufficient to produce backward CS contrast effects without additional procedural details of Moran and Bar-Anan's (2013) paradigm. The paradigm we employed was different from Moran and Bar-Anan's, because we used multiple USs, single CSs, and the same stimulus presentation timing as Mallan et al. (2008). For example, it is possible that removing the overlap between US offset and CS onset and removing the varying US durations that made US offset predictable rendered the instructions less effective. We also measured explicit valence ratings and presented the affective priming task before and after the training phase, whereas Moran and Bar-Anan (2013) only presented their measures of CS valence after the training phase. This may have resulted in an 'evaluative mindset' that led participants to evaluate stimuli differently than in Moran and Bar-Anan (2013; Gast & Rothermund, 2011). Furthermore, it could be that the 'start-stop instructions' produce backward CS contrast effects only when using acoustic USs, or only when the USs are more intense and/or more salient than the picture USs used here. Consistent with this possibility, Moran et al. (2016) found backward CS contrast effects in a picture-picture paradigm using instructions that emphasised the valence of the US (i.e., "getting gold bars is a happy event, whereas getting garbage piles is a sad event") and the agency of the families in starting and stopping the US (i.e., "creatures control whether happy or sad events happen to you").

Before we can draw conclusions about our hypothesis regarding US predictability and affective relief/disappointment eliciting backward CS contrast effects, we need to ensure that we can elicit backward CS contrast effects with instructions in a picture-picture paradigm. Moreover, we need to assess whether putting participants in an 'evaluative mindset' by presenting valence ratings and affective priming before the learning phase is affecting backward CS learning. To achieve this we decided to replicate Moran et al.'s (2016) findings using their exact paradigm and instructions to determine whether US and CS overlap, US variability or an 'evaluative mindset' may have contributed to not observing backward CS contrast effects in the 'start-stop instructions' group of Experiment 1.

Pilot Study 2

Pilot Study 2 had two aims: (1) to replicate Experiment 1 from Moran et al. (2016), as we did not find contrast effects in our first experiment using a slightly different paradigm and using instructions that were based on an earlier study employing sound USs (Moran & Bar-Anan, 2013), and (2) to assess whether having participants complete explicit valence ratings and an affective priming task before conditioning puts them in an 'evaluative mindset' that results in a failure to find contrast effects in backward conditioning. If contrast effects emerge for backwardly conditioned CSs in both groups in Pilot Study 2, the lack of a contrast effect in Pilot Study 1 may be due to the differences in the paradigm or instructions. If contrast effects emerge for backwardly conditioned CSs in the no pre-measure group only, the lack of contrast effects observed in Pilot Study 1 may be due to putting participants in an 'evaluative mindset' before the conditioning task.

Method

Participants and design. Participants were recruited through M-Turk. The sample comprised 95 participants, with a mean age of 36.57, *SD* = 10.504, after duplicates and those failing to complete the experiment were removed (n = 18). As in Pilot Study 1, the sample size was based on previous research (Moran & Bar-Anan, 2013; Moran et al., 2016). The 'no pre-measure' group consisted of 48 participants (19 female) and the 'pre-measure' group included 47 participants (23 female). Groups did

not differ on gender, χ^2 (2) = 2.06, p = .356, ethnicity, χ^2 (4) = 0.42, p = .981, or age, t (93) = 0.81, p = .418, d = 0.17, BF_{10} = 0.29. A 2 (Pre-measure: pre-measure vs no pre-measure; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US Valence: positive vs negative; within-participants) mixed design was used to replicate Moran et al. (2016) and to determine whether presenting explicit valence rating and affective priming tasks before conditioning affects the pattern of responding on these tasks after conditioning. In the pre-measure group, participants completed an explicit valence rating and affective priming task before and after conditioning. In the no pre-measure group, which was a direct replication of the first experiment in Moran et al. (2016), participants completed a conditioning task, followed by explicit valence ratings and an affective priming task. Both groups then completed a recollective memory test and a demographics questionnaire.

Explicit valence ratings. Each CS family was presented alone and participants were asked "Based on your very first emotional response, how much do you like the creatures in the picture? Click the appropriate answer below: dislike strongly, dislike moderately, dislike slightly, like slightly, like moderately, like strongly".

Affective priming task. Two creatures from each family were presented with positive and negative words twice, and two creatures from each family were presented with positive and negative words three times, for a total of 10 positive and 10 negative word pairings per family. This resulted in 80 trials. All other details were the same as in Pilot Study 1.

Recollective memory test. Each CS family was presented alone and participants were asked "In the game, what was the role of the creatures in the picture? Click the appropriate answer below: Starting gold, starting garbage, stopping gold, stopping garbage?"

Demographics questionnaire. The demographics questionnaire was identical to Experiment 1.

Apparatus/stimuli. CSs and USs were those used by Moran et al. (2016; available at https://osf.io/v2trw/). CSs were four families of alien creatures, with each family comprising four

creatures for a total of 16 CSs. The positive US was a picture of puppies, gold bars, and a baby, presented next to each other as a single image, and the negative US was a picture of an aggressive dog, garbage, and a crying child presented next to each other as a single image. Inquisit 4 by Millisecond Software ™ (2016) was used to run the experiment and to record responses in all tasks.

Procedure. The 'pre-measure' group completed explicit valence ratings and an affective priming task before the training phase, whilst the 'no pre-measure' group went straight to the training phase. The training phase comprised 12 positive and 12 negative trials randomly presented with inter-trial intervals of 2s. Each trial consisted of a forward CS, followed by a positive or negative US, followed by a backward CS. This CS-US-CS paradigm was an exact replication of Moran et al. (2016). CSs were presented for 1.5 seconds and USs were presented in blocks of 1s flashes with a 200ms break between each flash for a total of 3 or 5s of total US presentation time. Onset of the US coincided with offset of the forward CS, and onset of the backward CS occurred 200ms after the last US appearance.

Prior to the training phase, we presented participants with the exact instructions used by Moran et al. (2016; termed 'valence-agency instructions'). These instructions differ slightly from those used by Moran and Bar-Anan (2013; termed 'start-stop instructions'), as they highlight valence and agency. The instructions read as follows:

> *In the next game, you will get piles of shiny gold bars, but also some stinky garbage piles. Getting gold bars is a happy event, whereas getting garbage piles is a sad event. In the game, four families of creatures control whether happy or sad events happen to you. These are the four families. One family of creatures will always start the gold bars coming your way. A second family of creatures will always stop the gold bars. A third family of creatures will always start garbage piles coming your way. A fourth family of creatures will always stop the garbage piles. Your goal in this game is to learn which family of creatures starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage. We will test your learning later in the game, so please pay close attention. If you*

read and understood the instructions, hit the spare bar to continue. Please pay close attention to the images on the screen. Make sure you learn and remember which family does each of the four actions (start gold, stop gold, start garbage, stop garbage). Press space to start the game.

After 12 trials, the following instructions were presented:

Do you know by now which family starts the gold, which family stops the gold, which family starts the garbage, and which family stops the garbage? Try to memorize what each family does for a later test. Press space for a few more rounds to help you remember the roles of the families better.

After the training phase, the explicit valence ratings and affective priming tasks were presented, followed by the recollective memory test and demographics questionnaire. Participants then received a completion code to enter to receive their payment, and thanked for participating. The experiment took 13 minutes on average to complete, and participants were compensated US-$5.

Statistical analyses. Responses following CSs within the same family in the affective priming task were averaged to provide overall means for each family. EC scores were calculated following the procedures in Pilot Study 1. EC scores and priming scores were subjected to separate 2 (Pre-measure: pre-measure vs no pre-measure; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) mixed-model ANOVAs. Participants with a percentage of invalid trials larger than 25% were removed from the priming analyses (n = 4 in the ‘pre-measure’ group, n = 6 in the ‘no pre-measure’ group). In the final sample at pre-test, 6.62% of trials were incorrectly categorised and 7.47% of trials were outliers. At post-test, 7.50% of trials were incorrectly categorised and 6.33% of trials were outliers. The reliability of the priming task was $\alpha = .13$ at pre-test, and $\alpha = .45$ at post-test. All other details were the same as in Experiment 1.

Results

Explicit valence ratings. Figure S3 shows mean EC scores at post-test as a function of Conditioning Type and Pre-measure. The figure suggests assimilation effects for forward conditioning and contrast effects for backward conditioning in both the ‘pre-measure’ and the ‘no-pre-measure’

groups. A main effect of Conditioning Type revealed that forward conditioning EC scores were significantly larger than backward conditioning EC scores, F (1, 93) = 223.65, $p < .001$, $\eta_p^2 = .706$, $BF_{10} = 1.81 \times 10^{40}$. One-sample t-tests further showed that forward conditioning EC scores were significantly larger than zero, t (94) = 18.67, $p < .001$, $d = 1.92$, $BF_{10} = 1.30 \times 10^{30}$, and backward conditioning EC scores were significantly smaller than zero, t (94) = 7.49, $p < .001$, $d = 0.77$, 2.45×10^8. The two-way interaction between Instructions and Conditioning Type was not significant, F (1, 93) = 1.21, $p = .274$, $\eta_p^2 = .013$, $BF_{incl} = 0.49$.

Affective priming. Figure S4 shows mean EC scores on affective priming at post-test as a function of Conditioning Type and Pre-measure. A marginal main effect of Conditioning Type suggests a larger priming score for forward conditioning than backward conditioning, F (1, 83) = 3.42, $p = .068$, $\eta_p^2 = .040$, $BF_{incl} = 0.99$. The two-way interaction between Instructions and Conditioning Type was not significant, F (1, 83) = 1.14, $p = .289$, $\eta_p^2 = .014$, $BF_{incl} = 0.41$. One sample t-tests showed that only priming scores for forward conditioning were significantly larger than zero, t (84) = 2.53, $p = .013$, $d = 0.27$, $BF_{10} = 2.39$. Priming scores for backward conditioning were not significantly different from zero, t (84) = 0.04, $p = .972$, $d < 0.01$, $BF_{10} = 0.12$.

Discussion

In the current study, explicit valence ratings showed assimilation effects for forward CSs and contrast effects for backward CSs in both groups, and this pattern emerged regardless of whether participants did or did not complete measures of CS valence before the training phase. In addition to successfully replicating the backward CS contrast effects observed by Moran et al. (2016), these findings suggest that the lack of a contrast effect for backwardly conditioned CSs in Pilot Study 1 was not due to participants being in a ‘evaluative mindset’. However, differing from Moran and Bar-Anan’s (2013) findings, affective priming scores revealed a significant but weak assimilation effect only for forward, but not for backward, CSs.

Although Pilot Study 2 rules out an ‘evaluative mindset’ as a potential explanation for the lack of a backward contrast effects on explicit valence ratings in Pilot Study 1, it is possible that the

difference between the findings in Pilot Studies 1 and 2 is due to the procedural differences between the two studies mentioned in the Pilot Study 1 discussion. Whereas the conditioning procedure in Pilot Study 1 aligned more closely with Mallan et al.'s (2008) paradigm, the conditioning procedure in Pilot Study 2 directly replicated Moran et al.'s (2016) paradigm which included US/CS overlap and US variability. These parameters may increase the influence of the instructions and may also increase the amount of affective relief/disappointment that occurs at the offset of the US. On the other hand, the different results may have been due to differences between the instructions used in Pilot Study 1 (taken from Moran & Bar-Anan, 2013) and those used in Pilot Study 2 (taken from Moran et al., 2016), which different in the strength and perception of control the CSs possess ('valence' and 'agency' components). Finally, the observed backward CS contrast effect in Pilot Study 2 may have been artificially amplified by using a 6-point explicit valence rating scale instead of the 9-point scale used in Pilot Study 1. The lack of a midpoint on the scale forces participants to choose either 'dislike slightly', or 'like slightly', which may have pushed those who would have rated CSs as neutral to select a specific valence. This aspect of the rating scale may have resulted in a significant backward CS contrast effect in Pilot Study 2 that was not observed in Pilot Study 1, because participants had the option to rate stimuli as neutral.

No priming effects were found for backward CSs which deviates from previous research and a-priori predictions. It is possible that presenting multiple exemplars of the same set of CSs reduced an already small effect to a null effect as supported by the Bayesian analyses. It is also possible that competing assimilation and contrast effects cancelled each other out leading to a null effect. This may be plausible given the speeded nature of the task. However, results regarding the effects of speeded tasks (i.e., valence rating tasks that impose time limits on responding, as well as other implicit measures) on backward CS evaluations in this paradigm are currently inconclusive (see Moran & Bar-Anan, 2019).

The concerns regarding explicit valence ratings mentioned above, and potential moderators of the backward CS contrast effects present in Pilot Study 2, but not in Pilot Study 1, were addressed in

the main experiment by contrasting paradigm (‘Mallan paradigm’ from Pilot Study 1 vs. ‘Moran paradigm’ from Pilot Study 2) and instructions (‘observe instructions’ from Pilot Study 1 vs. ‘valence-agency instructions’ from Pilot Study 2) in a mixed factorial design.

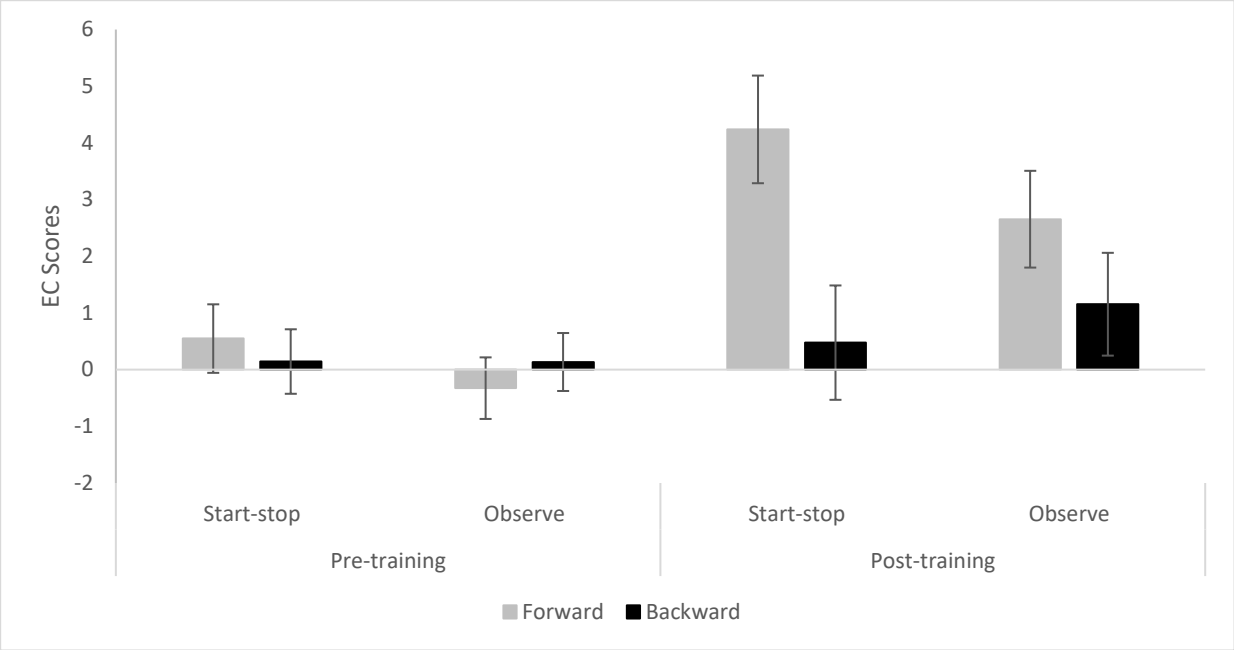

Figure S1. EC scores on explicit valence ratings for forward and backward conditioning measured pre-training and post-training for the start-stop instructions and observe instructions groups, Pilot Study 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

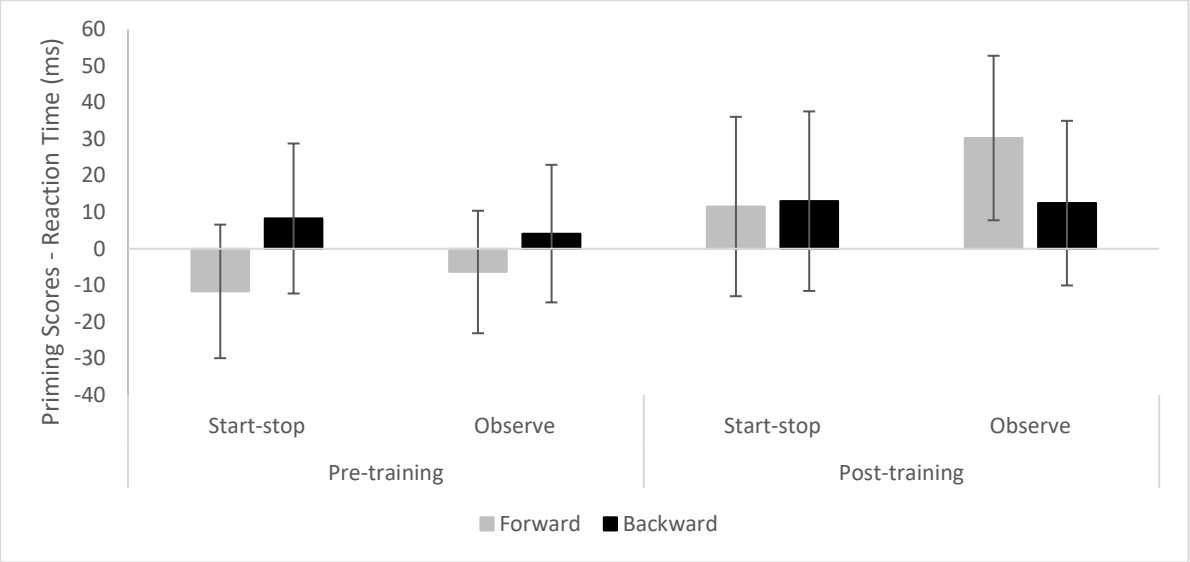

Figure S2. EC scores on affective priming for forward and backward conditioning, measured pre-training and post-training for the start-stop instructions and observe instructions groups, Pilot Study 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

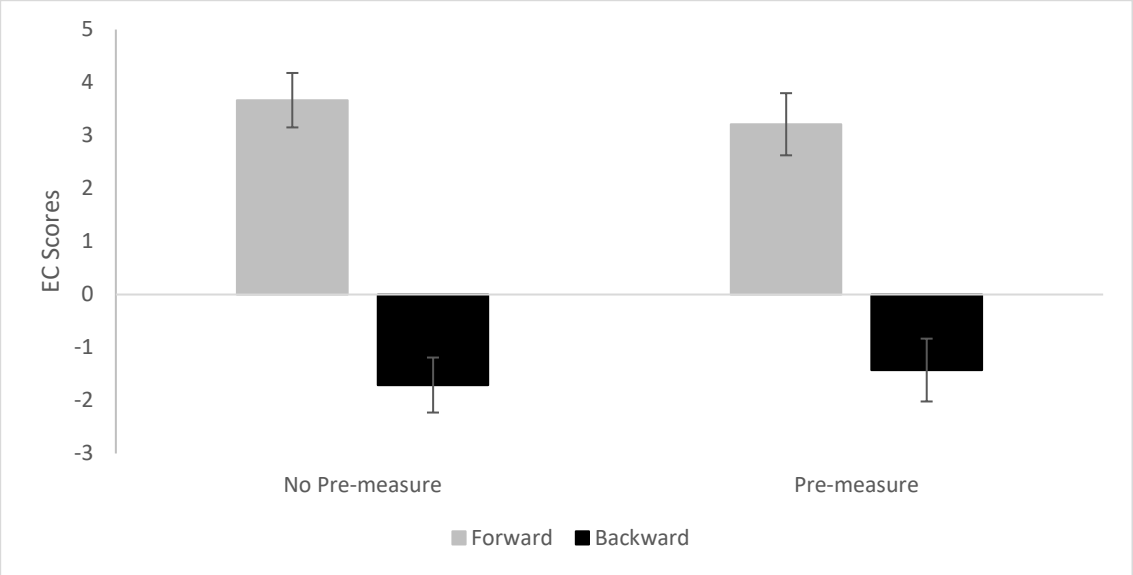

Figure S3. Post-test EC scores on explicit valence ratings for forward and backward conditioning for the 'no pre-measure' and 'pre-measure' groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

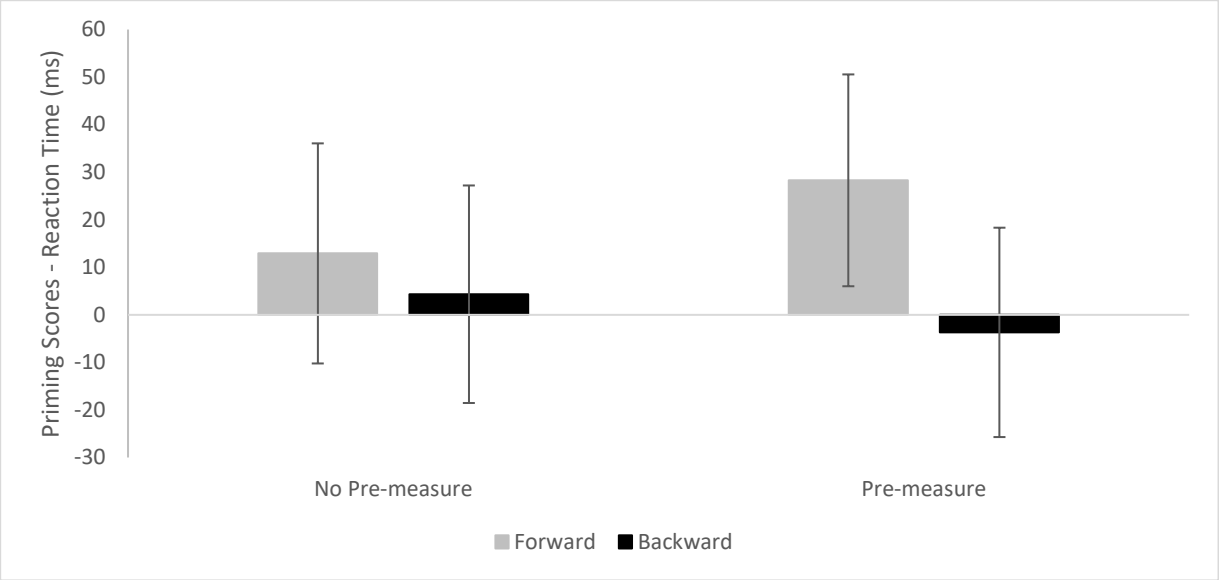

Figure S4. Post-test EC scores on affective priming for forward and backward conditioning for the 'no pre-measure' and 'pre-measure' groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

References

Bading, K., Stahl, C., & Rothermund, K. (2019). Why a standard IAT effect cannot provide evidence for association formation: The role of similarity construction. *Cognition and Emotion.* doi:10.1080/02699931.2019.1604322.

Center for the Study of Emotion and Attention [CSEA – NIHM] (1999). *International affective picture system: Digitized photographs*. The Center for Research in Psychophysiology, University of Florida.

Gast, A, & Rothermund, K. (2011). What you see is what will change: Evaluative conditioning effects depend on a focus on valence. *Cognition and Emotion, 25,* 89-110.

Hu, X., Gawronski, B., & Balas, R. (2017a). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin, 43,* 17-32.

Hu, X., Gawronski, B., & Balas, R. (2017b). Propositional versus dual-process accounts of evaluative conditioning: II. The effectiveness of counter-conditioning and counter-instructions in changing implicit and explicit evaluations. *Social Psychological and Personality Science, 8,* 858-866.

Inquisit 4 [Computer software]. (2016). Retrieved from https://www.millisecond.com.

Koppehele-Gossel, J., Hoffmann, L., Banse, R., & Gawronski, B. (in press). Evaluative priming as an implicit measure of evaluation: An examination of outlier-treatments for evaluative priming scores. *Journal of Experimental Social Psychology.*

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behavior Research Methods, 1-10.

Mallan, K. M., Lipp, O. V., & Libera, M. (2008). Affect, attention, or anticipatory arousal? Human blink startle modulation in forward and backward affective conditioning. *International Journal of Psychophysiology, 69.* 9-17.

Moran, T., & Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion, 27,* 743-752.

Moran, T., & Bar-Anan, Y. (2019). The effect of co-occurrence and relational information on speeded evaluation. *Cognition and Emotion*. doi:10.1080/02699931.2019.1604321

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition, 34.* 435-461.

Additional Analyses: Pilot Study 1

Recollective Memory Test

Seven participants in the start-stop instructions group, and 43 participants in the observe instructions group failed to correctly verbalise the contingency. This difference between groups seems due to the very strict criterion (100% correct report) applied and is not indicative of learning without awareness in the observe instructions group.

Explicit valence ratings

The pattern of EC scores for participants who passed the recollective memory test was similar to the entire sample. The only difference was that the significant Conditioning Type × Time interaction from the entire sample was marginal, F (1, 42) = 4.015, p = .052, η_p^2 = .087. Follow-up analyses revealed the same pattern as the entire sample.

Affective priming – Reaction Time

When analysing only those who passed the recollective memory test, the significant main effect of time became marginal, F (1, 39) = 3.511, p = .068, η_p^2 = .083. As per the entire sample, EC scores were larger at post-training than pre-training.

Affective priming – Errors

Mean EC scores for errors on the affective priming measure are depicted in Figure S5. The ANOVA revealed only a significant main effect of Time, F (1, 79) = 3.991, p = .049, η_p^2 = .048, indicating that EC scores were significantly larger at post-training than pre-training. Follow-up analyses revealed that priming scores were not different zero at pre-training, t (80) = 1.218, p = .227, d = 0.14, or at post-training, t (80) = 1.832, p = .071, d = 0.20. When analysing data from participants who passed the recollective memory test, the main effect of time was not significant, F (1, 39) = 0.262, p = .612, η_p^2 = .007.

Additional Analyses: Pilot Study 2

Recollective Memory Test

Nine participants in each group failed to correctly verbalise the contingency.

Explicit valence ratings

The pattern of results was the same as the entire sample when analysing those who passed the recollective memory test.

Affective priming – Reaction Time

When analysing only those who passed the recollective memory test, the main effect of conditioning type became significant, F (1, 71) = 8.784, p = .004, η_p^2 = .110. Follow-up analyses revealed the same pattern as the entire sample.

Affective priming – Errors

Figure S6 shows mean EC scores for errors on affective priming at post-test as a function of Conditioning Type and Pre-measure. A main effect of Conditioning Type suggests a greater priming score for forward conditioning than backward conditioning, F (1, 83) = 4.278, p = .042, η_p^2 = .049. However, one sample t-tests showed that priming scores for forward conditioning, t (84) = 1.302, p = .197, d = 0.14, and backward conditioning were not significantly different from zero, t (84) = 1.565, p = .121, d = 0.17.

Figure S7 suggests a contrast effect for backward CSs in the pre-measure group only when analysing participants who passed the recollective memory test. The Conditioning Type × Group interaction was significant, F (1, 71) = 5.513, p = .022, η_p^2 = .072. Follow-up analyses revealed that EC error scores were smaller for backward conditioning than forward conditioning for the pre-measure group only, t (36) = 2.842, p = .006, d = 0.47, with no differences for the no pre-measure group t (35) = 0.494, p = .623, d = 0.08. One-sample t-tests showed that backward conditioning EC error scores in the pre-measure group were significantly below zero, t (36) = 2.412, p = .021, d = 0.40. No other comparisons differed from zero, all ts < 1.346, all ps > .181, all ds < 0.22.

Analyses of pre-measure only group across Time

EC and priming scores from the 'Pre-measure' group were also subjected to a 2 (Conditioning Type: forward vs backward; within-participants) × 2 (Time: pre-test vs post-test; within-participants) repeated measures ANOVA.

Explicit valence ratings. Figure S8 shows an assimilation effect for forward conditioning and a contrast effect for backward conditioning at post-training. Main effects of Conditioning Type, F (1, 46) = 97.713, $p < .001$, $\eta p^2 = .680$, and Time, F (1, 46) = 21.192, $p < .001$, $\eta p^2 = .315$, were qualified by a Conditioning Type × Time interaction, F (1, 46) = 69.436, $p < .001$, $\eta p^2 = .602$. Follow-up analyses revealed that EC scores for forward conditioning were significantly larger than EC scores for backward conditioning at post-training, F (1, 46) = 106.963, $p < .001$, $\eta p^2 = .699$, but no difference occurred at pre-training, F (1, 46) = 0.028, $p = .868$, $\eta p^2 = .001$. One-sample t-tests showed forward conditioning EC scores were significantly larger than 0, t (46) = 11.965, $p < .001$, $d = 1.75$, and backward conditioning EC scores were significantly less than 0, t (46) = 4.893, $p < .001$, $d = 0.71$, at post-training. Forward and backward conditioning EC scores did not differ from 0 at pre-training, t (46) = 0.133, $p = .894$, $d = 0.02$, and t (46) = 0.116, $p = .908$, $d = 0.02$, respectively. The pattern of results was the same when only analysing those who passed the recollective memory test.

Affective priming – Reaction time. Figure S9 shows an assimilation effect for forward conditioning at post-training. The Conditioning Type × Time interaction was significant, F (1, 39) = 9.895, $p = .003$, $\eta p^2 = .202$. Forward conditioning priming scores were larger than backward conditioning at post-training, F (1, 39) = 6.844, $p = .013$, $\eta p^2 = .149$, but not pre-training, F (1, 39) = 0.902, $p = .348$, $\eta p^2 = .023$. One sample t-tests showed that forward conditioning priming scores at post-training were significantly larger than 0, t (39) = 2.908, $p = .006$, d's < 0.46, and that no other scores differed from 0, t's < 1.562, p's > .126, d's < 0.25. The pattern of results from those who passed the recollective memory did not differ from the entire sample.

Affective priming – Errors. Figure S10 shows mean EC scores for errors on affective priming at post-test as a function of conditioning type and Pre-measure. A main effect of Conditioning

Type suggests a larger priming score for forward conditioning than backward conditioning, F (1, 39) = 9.823, p = .003, η_p^2 = .201. A one sample t-test showed that the priming score for forward conditioning did not differ from zero, t (39) = 1.657, p = .105, d = 0.26. For backward conditioning, the priming score was significantly less than zero, t (39) = 2.726, p = .010, d = 0.43. The pattern of results from those who passed the recollective memory did not differ from the entire sample.

Additional Analyses: Main Experiment

Recollective Memory Test

Thirty two participants in the 'Start-Stop instruction, Mallan paradigm' group, 33 participants in the 'Start-Stop instruction, Moran paradigm' group, eight participants in the 'Valence-Agency instruction, Mallan paradigm' group, and seven participants in the 'Valence-Agency instruction, Moran paradigm' group failed to verbalise the contingency.

Explicit valence ratings

When analysing participants who passed the recollective memory test, the Instructions × Time interaction became significant, F (1, 106) = 5.451, p = .021, η_p^2 = .049, and the Time × Paradigm interaction became marginal, F (1, 106) = 3.538, p = .063, η_p^2 = .032. Overall, the pattern of results did not change as the Instructions × Conditioning Type × Time interaction remained significant, F (1, 106) = 20.693, p < .001, η_p^2 = .163.

Affective priming – Reaction time

When analysing participants who passed the recollective memory test the main effect of Time became significant, F (1, 95) = 4.142, p = .045, η_p^2 = .042, revealing the same pattern of results as the entire sample.

Affective priming – Errors

Figure S11 shows mean EC scores on affective priming for forward and backward conditioning measured pre-training and post-training as a function of Instructions and Paradigm. The figure suggests a contrast effect at post-training in the valence-agency instructions/Mallan paradigm

group for backward conditioning. An Instructions × Time interaction, F (1, 153) = 6.124, p = .014, η_p^2 = .038, and a Paradigm × Conditioning Type × Time interaction were found. Following up the Instruction × Time interaction revealed that priming scores in the observe instructions group were marginally larger at post-training than pre-training, t (39) = 1.696, p = .092, d = 0.27. In the valence-agency instructions group, priming scores did not differ between pre- and post-training, t (153) = 0.321, p = .749, d = 0.05. Following up the Paradigm × Conditioning Type × Time interaction revealed larger priming scores in the Moran paradigm group at post-training when compared with pre-training for backward conditioning, t (153) = 2.402, p = .018, d = 0.38. No other comparisons were significant, ts < 1.552, ps > .123, ds < 0.25. When analysing participants who passed the recollective memory test, no main effects or interactions were significant, Fs < 2.959, ps > .089, η_p^2s < .030.

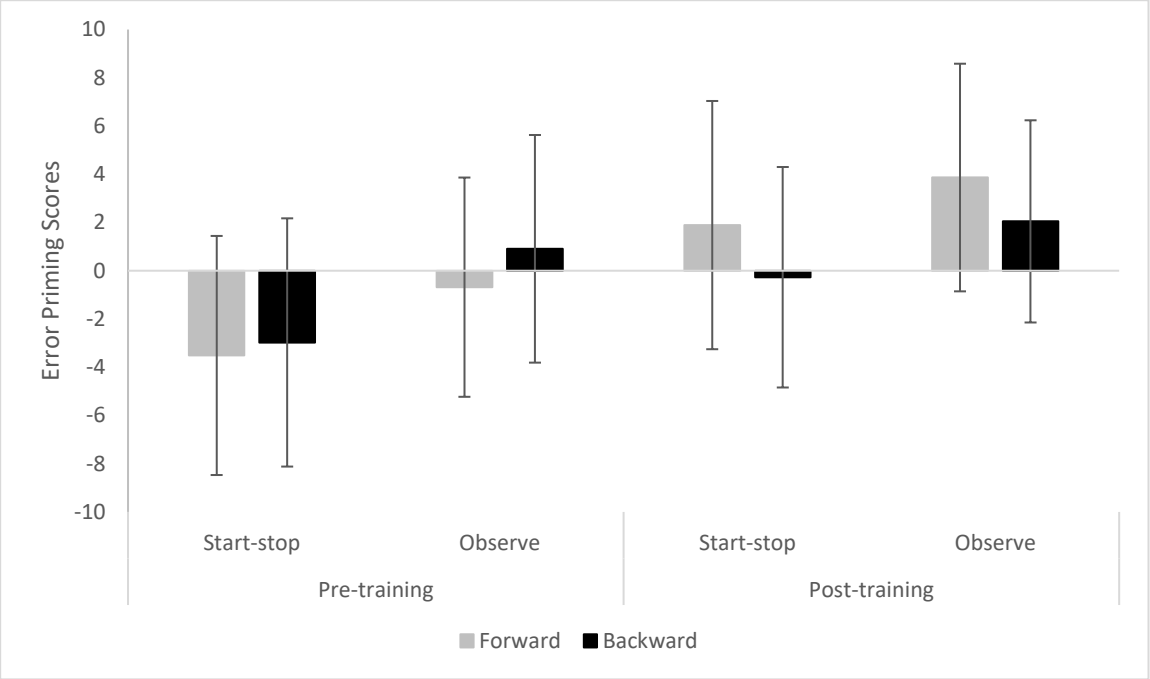

Figure S5. EC scores for errors on affective priming for forward and backward conditioning, measured pre-training and post-training for start-stop instructions and observe instructions, Pilot Study 1. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

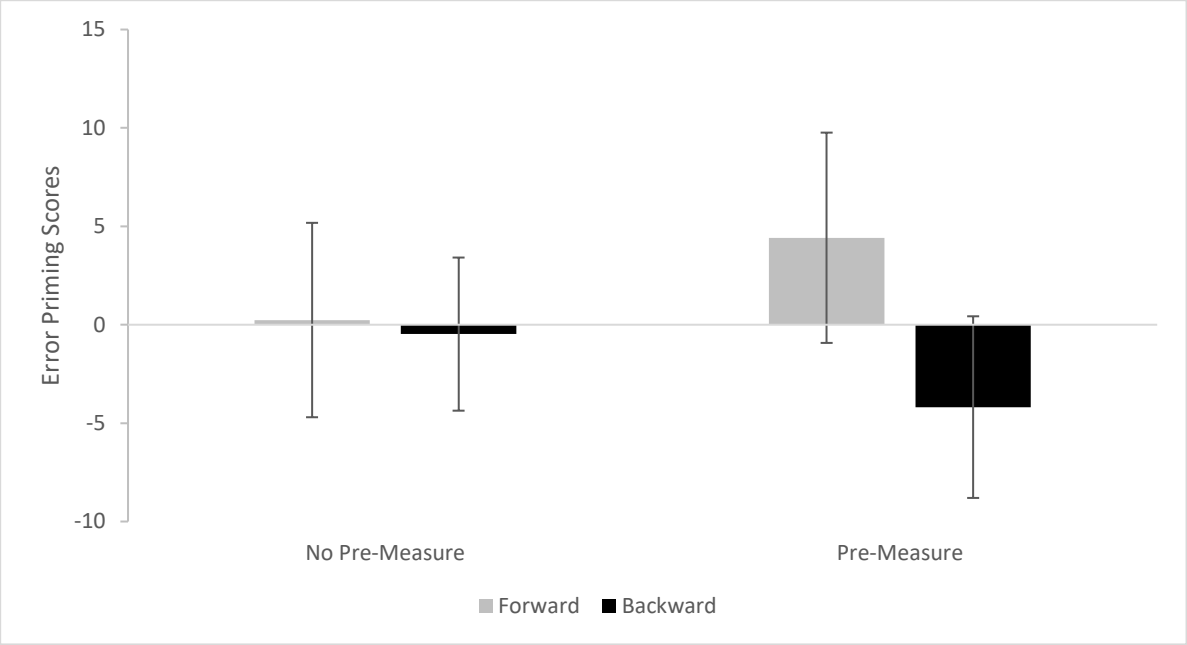

Figure S6. EC scores for errors on affective priming for forward and backward conditioning, measured post-training for no pre-measure and pre-measure groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

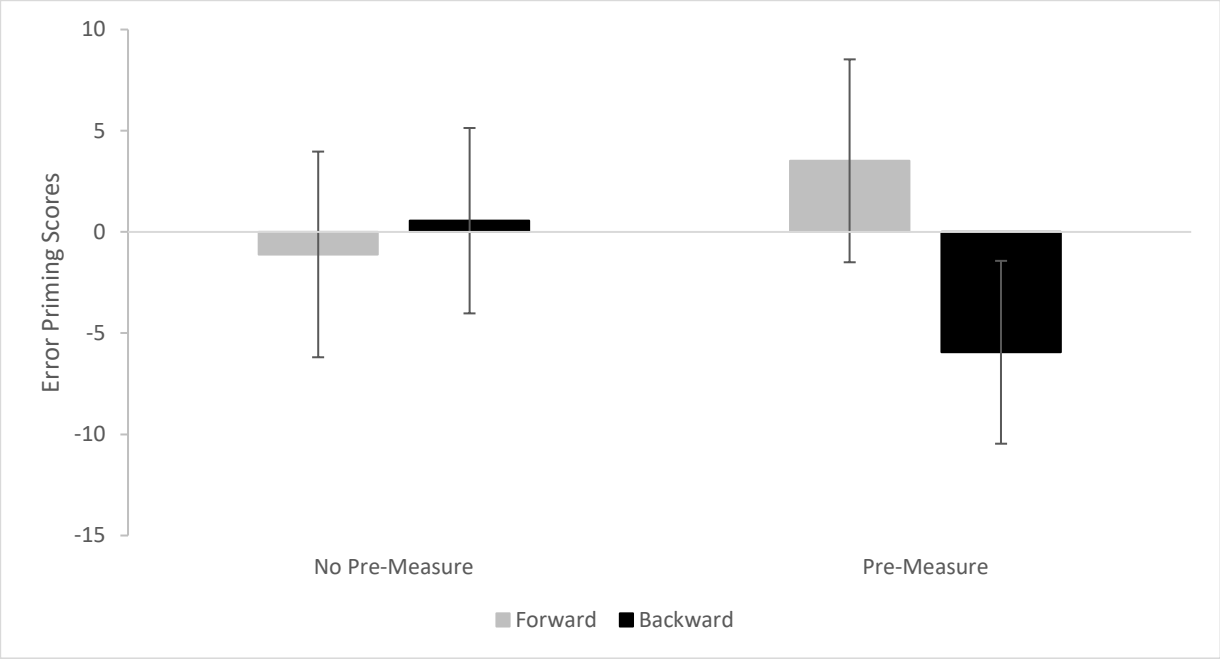

Figure S7. EC scores from participants who passed the recollective memory test for errors on affective priming for forward and backward conditioning, measured post-training for no pre-measure and premeasure groups, Pilot Study 2. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.

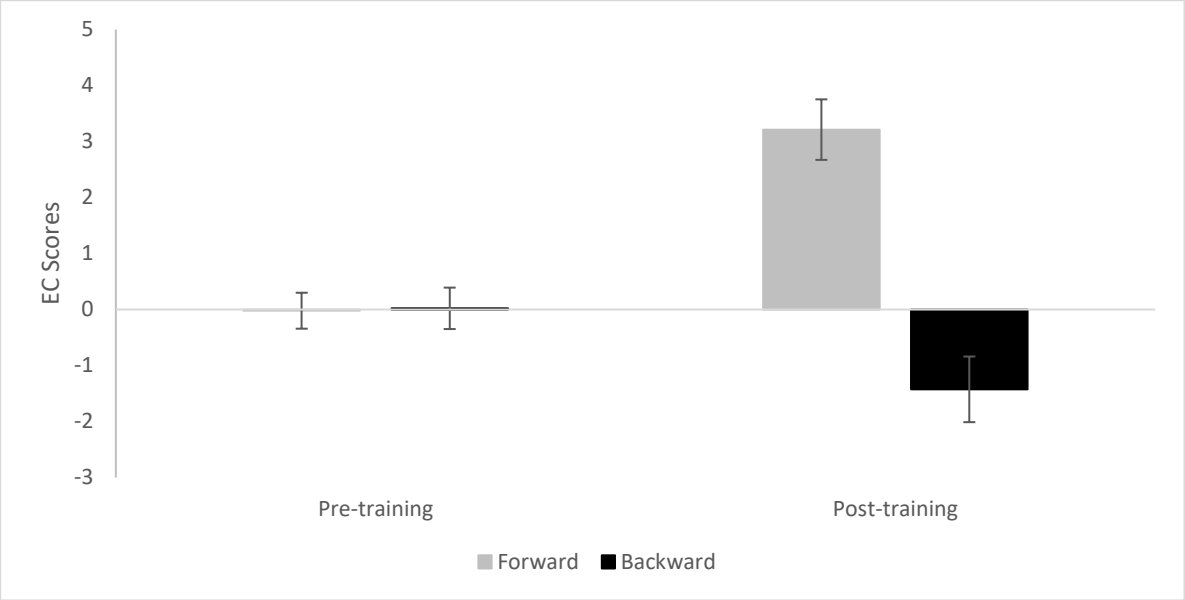

Figure S8. EC scores for forward and backward conditioning measured pre and post-training for the pre-measure group only, Pilot Study 2. Positive scores suggest assimilation effects, negative scores suggest contrast effects. Error bars show 95% confidence intervals of the mean.

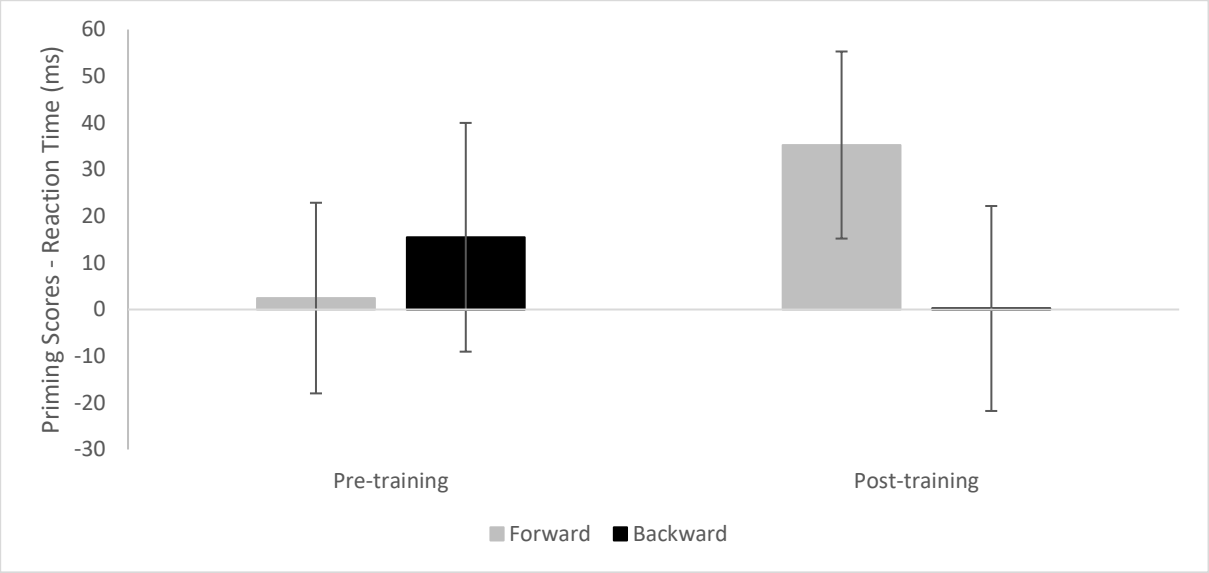

Figure S9. Priming scores from reaction times for forward and backward conditioning, measured pre and post-training for the pre-measure group only, Pilot Study 2. Positive scores suggest assimilation effects, negative scores suggest contrast effects. Error bars show 95% confidence intervals of the mean.

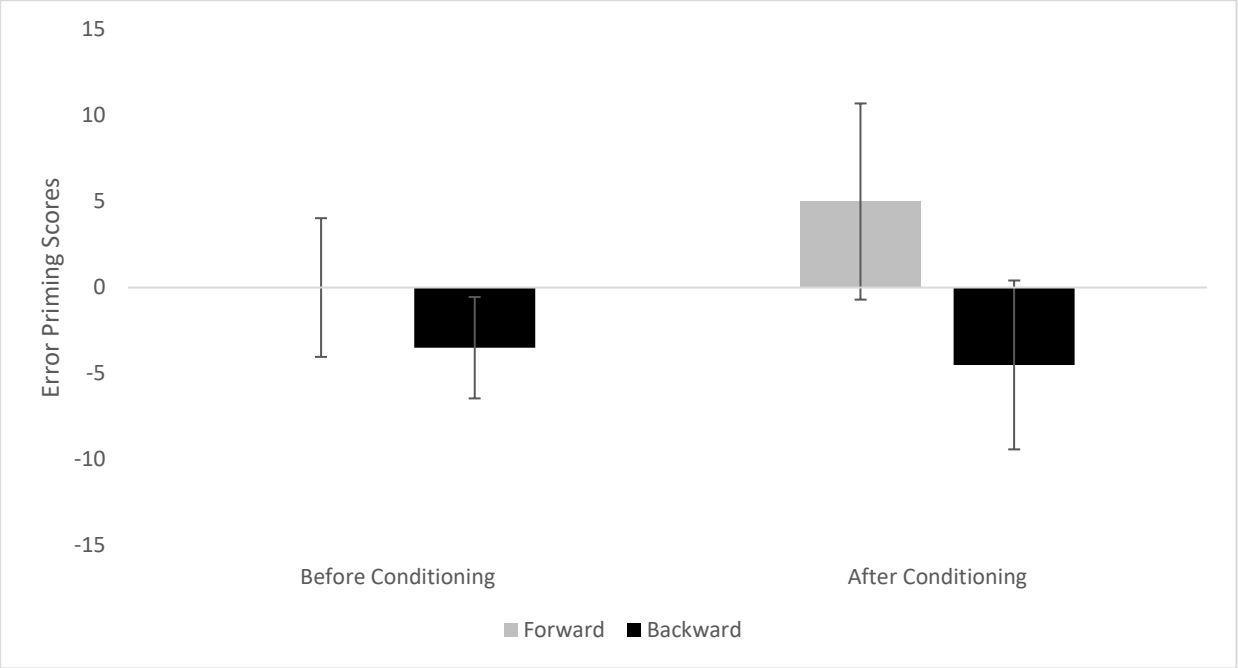

Figure S10. EC scores EC scores for errors on affective priming for forward and backward conditioning measured pre and post-training for the pre-measure group only, Pilot Study 2. Positive scores suggest assimilation effects, negative scores suggest contrast effects. Error bars show 95% confidence intervals of the mean.

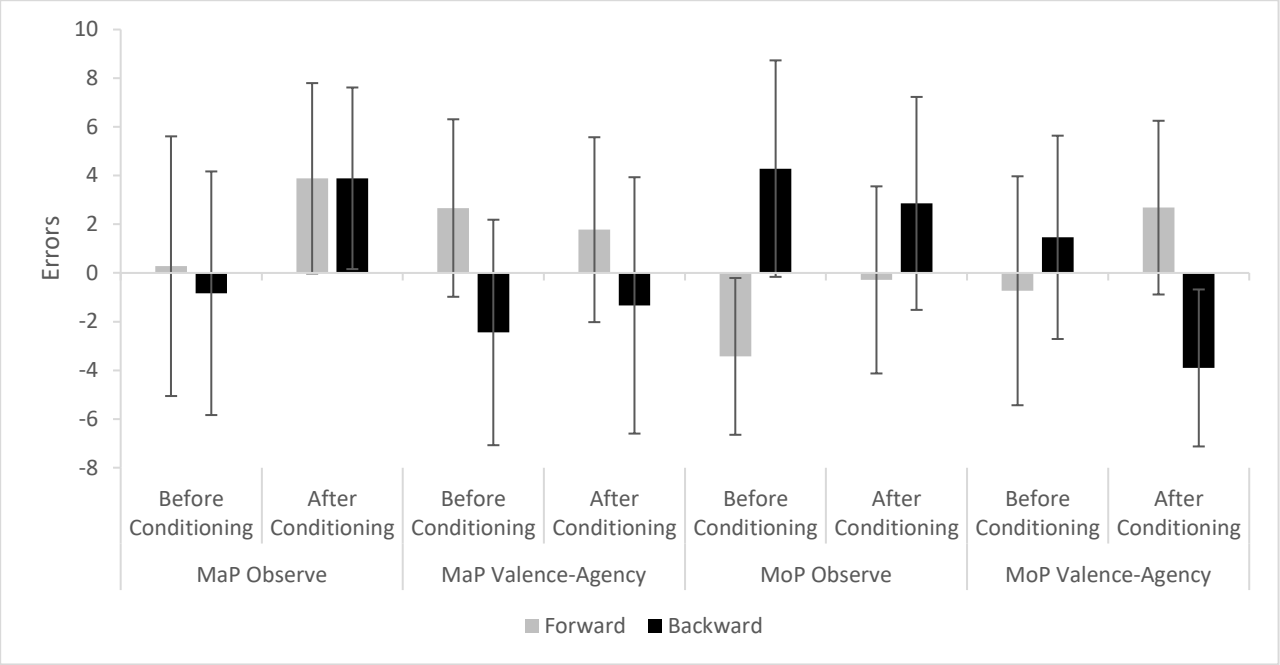

Figure S11. EC scores for errors on affective priming for forward and backward conditioning measured pre-training and post-training as a function of instructions ('observe instructions' and 'valence-agency instructions') and paradigm ('Mallan paradigm (MaP)' and 'Moran paradigm (MoP)'), Main Experiment. Positive scores indicate assimilation effects; negative scores indicate contrast effects. Error bars show 95% confidence intervals of the mean.