How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence

Luke J. S. Green

Curtin University

Camilla C. Luck

Curtin University

Ottmar V. Lipp

Curtin University

Word count: 11906 (239 word abstract)

This is the peer reviewed version of the following article: Green, L.J.S. and Luck, C.C. and Lipp, O.V. 2020. How disappointing: Startle modulation reveals conditional stimuli presented after pleasant unconditional stimuli acquire negative valence. Psychophysiology. 57 (8). Article No. e13563 which has been published in final form at https://doi.org/10.1111/psyp.13563. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Abstract

Past research on backward conditioning in evaluative and fear conditioning yielded inconsistent results in that self-report measures suggest that conditional stimuli (CS) acquired the valence of the US in fear conditioning (assimilation effects), but the opposite valence in evaluative conditioning (contrast effects). Conversely, implicit measure of CS valence suggest assimilation effects in evaluative backward conditioning whereas startle modulation indicates contrast effects in backward fear conditioning. The current study investigated whether US intensity could account for the dissociation on implicit measures between fear and evaluative conditioning. Self-report measures of evaluative learning indicated assimilation effects for forward conditioning, whereas backward contrast effects were observed with intense USs only. Blink startle modulation indicated assimilation effects in forward conditioning and contrast effects in backward conditioning, regardless of US intensity. Experiment 2 included a neutral US in order to assess whether the offset of the positive US elicits an opponent emotional response that mirrors relief (disappointment), which is thought to mediate the reduction in startle seen during backward CSs in fear conditioning. This opponent emotional response was evident as startle magnitude during backward CSs increased linearly with increasing US pleasantness. Omission of the forward CSs led to an assimilation effect in self-report measures. The current results extend our understanding of emotional learning to stimuli encountered after salient emotional events. Startle reflects the emotion prevailing after US offset, relief or disappointment, whereas self-report measures seem more attuned to factors such as US predictability and intensity.

Keywords: Evaluative conditioning, associative learning, backward conditioning, startle modulation, propositional learning

1. Introduction

Evaluations, i.e. how positive or negative a stimulus or event is, can influence many aspects of our lives, including career choice, voting and consumer behaviour, and our relationships with other people (see Galdi, Arcuri, & Gawronski, 2008; Gibson, 2008; LeBel & Campbell, 2009). These evaluations can be manipulated through *evaluative conditioning* (De Houwer, Thomas, & Baeyens, 2001; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010), a process in which the valence of a neutral conditional stimulus (CS) can be changed by pairing it with a positively or negatively valenced unconditional stimulus (US; De Houwer, 2007). Advertising campaigns often exploit EC by, for instance, presenting a product (the CS) with a popular celebrity (the US), resulting in positive evaluations of the product. As we all continually encounter stimuli of differing valence that co-occur in different temporal and spatial arrangements, the study of evaluative conditioning has immense importance as it is relevant for many facets of psychology and life in general.

Evaluative conditioning can be studied in the laboratory using a picture-picture paradigm. In this paradigm, neutral pictures (CS) are paired with positive or negative pictures (USs), which results in CSs paired with positive USs becoming more pleasant, and CSs paired with negative USs becoming more unpleasant (Hofmann et al., 2010; Levey & Martin, 1975; Mallan, Lipp, & Libera, 2008). These changes in CS valence can be tracked with explicit, i.e., self-report, and implicit measures, i.e., affective priming or Implicit Association Tests (Fazio & Olson, 2003). A similar evaluative change occurs in differential fear conditioning, as pairing a neutral picture (CS+) with an aversive electro-tactile stimulus (US) results in explicit negative evaluations of the CS+ in comparison to a neutral picture that was not paired with the US (CS-; Lipp, 2006). Moreover, negative CS+ valence acquired during differential fear conditioning can be measured implicitly using the startle blink reflex, as startle blink magnitude is larger during negative stimuli and smaller during positive stimuli when compared to neutral stimuli (Bradley, Cuthbert, & Lang, 1990; Vrana, Spence, & Lang, 1988; but see Lipp, Siddle, & Dall, 2003).

Changes in stimulus evaluation such that the CS acquires the valence of the US are known as assimilation effects. Assimilation effects have been demonstrated in evaluative and fear conditioning

utilising forward conditioning procedures (CS-US), and in evaluative conditioning utilising simultaneous (CS+US) and backward conditioning procedures (US-CS; Mallan et al., 2008; Hoffman et al., 2010). Assimilation effects have been shown using explicit valence ratings, the startle blink reflex, and reaction time based implicit measures of CS valence (Mallan et al., 2008; Olson & Fazio, 2001). However, contrast effects, the CS acquiring valence that is opposite to that of the US, have also been observed. Contrast effects have been shown on explicit valence ratings in evaluative conditioning employing instructions that emphasise CS agency (start/stop) after both forward and backward CS and US pairings (Hu, Gawronski, & Balas, 2017; Moran & Bar-Anan, 2013; Moran, Bar-Anan, & Nosek, 2016; Unkelbach & Fiedler, 2016). They have also been observed for backward conditioning on the startle blink reflex and explicit valence ratings in fear conditioning (Andreatta, Mühlberger, Yarali, Gerber, & Pauli, 2010; Andreatta, Mühlberger, Glotzbach-Schoon, & Pauli, 2013; Luck & Lipp, 2017; see also Mühlberger et al., 2011 for evidence of contrast effects during videos of faces changing from neutral to happy and angry and vice versa).

1.1. Contrast Effects in Evaluative Conditioning

Moran and Bar-Anan (2013) demonstrated contrast effects for backward CSs in a concurrent forward and backward evaluative conditioning procedure. In these within-subjects experiments, a pleasant melody (positive US) and an unpleasant human scream (negative US) were paired with CSs drawn from four different families of alien creatures. On positive trials, one CS (forward CSpos; F-CSpos) was presented before the pleasant melody (positive US) and a second CS (backward CSpos; B-CSpos) was presented after the pleasant melody. On negative trials, a third CS (forward CSneg; F-CSneg) preceded the unpleasant human scream (negative US) and a fourth CS (backward CSneg; B-CSneg) followed the unpleasant human scream. Before conditioning, participants were informed that each CS family had a different role to play; that one would start the positive US, one would stop the positive US, one would start the negative US, and one would stop the negative US, and that they needed to learn the role of each family for a later memory test. When assessing CS valence with explicit (ratings) and implicit measures (response time based tasks), a dissociation emerged. On implicit measures, assimilation effects were demonstrated for forward and backward conditioning, as CSs paired with the positive US were evaluated more positively than CSs paired with the negative

US. On explicit measures, an assimilation effect was demonstrated for forward conditioning only; CSs preceding the positive US (F-CSpos) were evaluated as more pleasant than CSs preceding the negative US (F-CSneg). For backward conditioning, however, a contrast effect was demonstrated; CSs following the positive US (B-CSpos) were rated as more negative than CSs following the negative US (B-CSneg). This contrast effect has also been replicated using picture USs but more explicit instructions highlighting the agency of the CSs in starting and stopping 'happy' and 'sad' events were required to obtain this result (Green, Luck, Gawronski, & Lipp, 2019; Moran et al., 2016).

1.2. Contrast Effects in Fear Conditioning (Relief Learning)

In fear conditioning, assimilation effects are found on explicit valence ratings and the startle blink reflex after forward conditioning, while backward conditioning leads to assimilation effects on explicit measures and contrast effects for the startle blink reflex (Andreatta et al., 2010; Andreatta, Mühlberger, & Pauli, 2016). This dissociation between explicit and implicit measures of CS valence in backward fear conditioning was initially demonstrated by Andreatta et al. (2010). Pictures of geometric shapes were presented either alone (CS-) or paired with an aversive electro-tactile stimulus (CS+) in a forward conditioning group (CS+-US/CS-), a backward conditioning group (US-6s gap-CS+/CS-), and a control group (CS+-6s gap-US/CS-). The CS+ was rated more negatively after forward and backward conditioning than before conditioning. Compared with the mean of all responses, startle blink magnitude elicited during CS+ was larger in the forward conditioning group (suggesting negative valence) and smaller during the CS+ in the backward conditioning group (suggesting positive valence). Moreover, startle blink magnitude during the CS+ in the backward conditioning group was smaller than during the CS- (suggesting the CS+ had acquired positive valence relative to the safety signal). These startle results are due to the 'relief' experienced at the offset of the aversive electro-tactile stimulus being conditioned to the backward CS (relief learning), and have been replicated using different timings between US offset and CS onset (Andreatta, et al., 2016; Luck & Lipp, 2017; see also Gerber et al., 2014 and Deutsch, Smith, Kordts-Freudinger, & Reichardt, 2015 for reviews on relief learning).

1.3. Explaining Opposite Patterns of Dissociations

Moran and Bar-Anan (2013) and Andreatta et al. (2010) found different patterns of dissociations between their explicit and respective implicit measures for backward conditioning. Moran and Bar-Anan (2013) found contrast effects on explicit valence ratings, while Andreatta et al. (2010) found an assimilation effect. On the other hand, Moran and Bar-Anan (2013) found assimilation effects on implicit measures of CS valence, while Andreatta et al. (2010) found a contrast effect on the startle blink reflex. There are several differences between these studies that could explain the contrasting patterns of dissociations, such as using different task instructions, presenting forward and backward conditioning within-subjects concurrently instead of comparing forward and backward conditioning between-subjects separately, presenting CSs and USs for different durations and with different inter-stimulus intervals, and using USs of differing intensities.

The different pattern of explicit valence ratings reported by Andreatta et al. (2010) and Moran and Bar-Anan (2013) can be explained by the task instructions. Green, et al. (2019) showed that presenting relational information highlighting the role of the CSs in a within-subjects concurrent forward and backward conditioning procedure produces contrast effects, and that without this information assimilation effects emerge. The different pattern on implicit measures, however, remains unexplained. The different procedures used (presenting forward and backward conditioning within-subjects as compared to between-subjects) cannot account for this as Andreatta and Pauli (2017) using a within-subjects procedure found the same results as Andreatta et al. (2010) using a between-subjects procedure. The differences in CS and US presentation duration and different inter-stimulus intervals cannot explain the difference either. Luck and Lipp (2017) found the same pattern as Andreatta et al. (2010) when using a 100ms gap instead of a 6s gap between US offset and backward CS onset. Moreover, Green et al. (2019) found that small paradigmatic differences such as CS and US duration, using multiple or single CSs and USs, and overlapping the CS and US presentations, do not influence backward evaluative conditioning on explicit or implicit measures in a within-subjects concurrent forward and backward conditioning procedure. Differences in US intensity on the other hand may explain the different patterns of results on implicit measures, as the shock US used by Andreatta et al.

(2010) is arguably more unpleasant than the auditory USs used by Moran and Bar-Anan (2013)[1]. US intensity is likely to affect relief learning as Bitar, Marchard, and Potvin (2018) found that pain relief was positively correlated with pain level, such that higher levels of pain led to greater pain relief (see also Leknes, Brooks, Wiech, & Tracey, 2008). Moreover, stronger learning tends to occur with stronger USs (Annau & Kamin, 1961; Pavlov, 1927). Therefore, assuming that implicit measures are less sensitive than explicit measures to contrast effects, then the lower intensity USs used in Moran and Bar-Anan (2013) may not have been sufficient to drive these effects, in contrast to the higher intensity USs used in Andreatta et al. (2010).

In Experiment 1, we aimed to use Moran and Bar-Anan's (2013) procedure to replicate their backward CS contrast effects on explicit ratings, while, measuring the startle blink reflex and manipulating US intensity between groups in a 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US Valence: positive vs negative; within-participants) mixed design. We hypothesised that contrast effects on explicit valence ratings would emerge for backward CSs in both groups, with a larger effect in the high intensity group. As Andreatta et al. (2010) demonstrated that startle blink magnitude was inhibited during a CS following a shock US (suggesting positive valence), we hypothesised that the startle blink reflex would show backward CS contrast effects for both groups, with a larger difference in the high intensity group. Moreover, we hypothesised that US intensity would influence the pattern of responding to backward CSs on an implicit behavioural measure, such that assimilation effects would occur in the low intensity group, and contrast effects would occur in the high intensity group. Finally, assimilation effects were expected on all measures for forward CSs[2].

Experiment 1

2. Method

2.1. Participants. Following ethical approval for this research protocol from the Curtin University Human Research Ethics Committee, 66 undergraduate students from the School of

[1] Moran and Bar-Anan (2013) did not report the intensity to which their sound USs were set, but attempts to recreate them following their description suggest that they were below 90dBA.
[2] All materials, data, analysis files, and supplementary materials, are available at https://osf.io/q46mp/.

Psychology at Curtin University participated in exchange for course credit. Two participants were excluded for having participated in an earlier study employing the same conditioning task. The final sample comprised 64 students (50 female), *M* age = 21.63, *SD* = 6.46, with 32 participants per group. The sample size was based on Andreatta et al. (2013), who employed 28 participants and found an effect size of $\eta p^2 = 0.316$ for the within-subjects comparison of conditioning type. Moreover, Andreatta et al. (2010) employed 33-34 participants per group and found a significant Conditioning Type × CS interaction. Based on this, we determined that 32 participants per group would provide sufficient power to detect the effects we were interested in. Five participants in the low intensity group and four in the high intensity group failed the recollective memory test. To pass this test participants needed to correctly identify the role of each of the four CSs with 100% accuracy. Analyses were run with and without these participants. Results are reported for the entire sample with those from participants who passed the recollective memory test added only if they provide additional clarification.

2.2. Apparatus/Stimuli. Four families of four aliens from Moran and Bar-Anan (2013) were used as CSs (examples shown in Figure 1). The CS families differed in head shape and colour. Also taken from Moran and Bar-Anan (2013), the positive US was a pleasant guitar melody (the start of 'The Shape of My Heart' by Sting), and the negative US was a human scream (all stimuli can be found at https://osf.io/cqsnj/). A pilot study (n = 20) was performed to match the perceived intensities of the USs, which were manipulated by altering the volume (see Table 1 for dBA values of USs used in the low and high intensity groups). US intensity ratings were submitted to a 2 × 2 repeated measures ANOVA revealing only main effects for valence, F (1, 17) = 15.02, $p < .001$, $\eta_p^2 = .442$, and intensity, F (3, 17) = 30.37, $p < .001$, $\eta_p^2 = .843$. The perceived intensity of the positive and negative USs used within each group was comparable, however, the dynamics of each US differed which led to different dBA readings (pilot study reported at https://osf.io/q46mp/).

Orbicularis Oculi electromyogram (EMG), skin conductance, and respiration were recorded using a Biopac MP150 system with AcqKnowledge Version 4.1 at a sampling rate of 1000Hz. Orbicularis Oculi EMG was measured using two 4 mm Ag/AgCl electrodes filled with electrode gel and attached using double-sided adhesive electrode collars. The first electrode was placed directly

under the pupil of the left eye, and the second under the corner of the left eye. Impedance was assessed to confirm electrode contact, though no threshold criterion was employed. A custom built noise-generator was used to present a 105dBA white noise burst lasting 50ms with a near instantaneous rise time as the startle eliciting stimulus. The EMG signal was amplified by a Biopac EMG100C amplifier at a gain of 5000 and high and low pass filtered at 10 and 500 Hz. Electrodermal responding was measured using two self-adhesive isotonic Biopac EL507 electrodes attached to the thenar and hypothenar eminences of the non-dominant hand. A Biopac EDA100C amplifier was used to DC amplify responses at a gain of 5 µSiemens per volt. A chest gauge was used to measure respiration to control for respiration or movement related artefacts in electrodermal responding. For the affective priming task, the CSs were used as primes and 10 positive words (*pleasant, good, outstanding, beautiful, magnificent, marvellous, excellent, appealing, delightful,* and *nice*) and 10 negative words (*unpleasant, bad, horrible, miserable, hideous, dreadful, painful, repulsive, awful,* and *ugly*) served as target stimuli. DMDX (Forster & Forster, 2003) was used to control stimulus presentations and markers and to present and record responses from the explicit valence ratings task, the affective priming task, and the memory test. Sennheiser HD-25-1 headphones were used to present auditory USs and startle probes.

INSERT FIGURE 1 ABOUT HERE

2.3. Procedure. Participants read the information sheet and were played each US from their condition for 30 seconds before providing informed consent. After signing the consent form, participants washed and dried the area under their left eye and their hands. The recording equipment was attached, and three habituation startles were presented (timing controlled manually by the experimenter to avoid coinciding with deliberate blinks, laughter, and fidgeting etc.) followed by a three minute baseline recording of skin conductance. The startles probes during habituation were not controlled by the software to be presented as specific time intervals. The experimenter pressed the shift key to present startle probes to the participant after they had recovered from the previous probe (i.e. no laughter, fidgeting, closing the eyes etc., and the EMG recording had returned to baseline).

The researcher then told participants to learn which family of aliens started and stopped the positive and negative sounds and that they would be tested on this at the end. The researcher started the script and the instructions were presented again on the screen. Participants were then presented with the CS-US-CS procedure comprising 10 positive US and 10 negative US trials (see Figure 1 for a depiction of a positive US trial). CSs were presented for 8s each, with a 2s overlap between CS-US and US-CS. Two CSs from each set of CSs were presented 3 times and 2 were presented twice, totalling 10 trials per set (10 F-CSpos, 10 B-CSpos, 10 F-CSneg, and 10 B-CSneg). USs were presented for 10, 15, 20, 25, or 30s. Startle probes were presented at 4.5 or 5.5s after forward CS onset, and 6.5 and 7.5s after backward CS onset (i.e., 4.5 or 5.5s after US offset; see Figure 2 for a depiction of startle probe timing). Startle probes were presented on six trials for each CS set, for a total of 24 probes. These probes were assigned randomly within forward and backward CSs separately. The inter-trial intervals were 12, 14, or 16s (randomly dispersed) and startle probes were presented half-way through half of the inter-trial intervals for a total of 10 startle probes. After conditioning, the experimenter informed participants they would now be asked to rate how much they liked each family of aliens. Participants were shown each set of CSs (four CSs per set) separately and asked to provide a rating of how much they liked each family on a scale from 1 = don't like at all, to 9 = like a lot. After providing ratings, the experimenter explained the affective priming task to the participants. The affective priming task comprised 80 trials where each set of CSs were presented with 10 positive words and 10 negative words. Two CSs from each set were each presented with all positive words once, while the other two CSs were presented with all negative words once. During a trial, a fixation cross was presented for 500ms, followed by the CS prime for 200ms, and then the target word until the participant responded by pressing the right 'SHIFT' key if the target word was positive, and the left 'SHIFT' key if the target word was negative. After a 20 trial affective priming practice task, the experimenter told participants they would now do the main affective priming task. Following the main affective priming task, the experimenter told participants it was time for the memory test. Participants were shown each family separately, and asked:

> *What was the role of the creatures in this picture? 1. Started the human sound. 2. Stopped the human sound. 3. Started the musical sound. 4. Stopped the musical sound.*

After the memory test, participants were told to pay attention to the screen and follow any instructions that appeared. Participants were told 'the experiment will now continue', and an extinction phase was presented. During extinction, each member of each CS set was presented twice for a total of 32 trials. Startles were presented at 4.5 or 5.5s after CS onset on six of the eight presentations of each CS set, totalling 24 startle probes. The inter-trial intervals were 12, 14, or 16s, and startle probes were presented half-way through the interval on half of the trials for a total of 16 startle probes. After this, participants were disconnected from the recording equipment and asked to fill out a demographics questionnaire. Participants were asked their age, gender, and ethnicity, as well as how pleasant/unpleasant the human sound, musical sound, and loud noises were on a 7-point scale ranging from -3 = very unpleasant to 3 = very pleasant. Participants were also asked how intense the human and musical sounds were on a 7-point scale from 0 = not at all to 6 = very intense, and how startling the loud noises (startle probes) were on a 7-point scale from 0 = not at all to 6 = very startling. Participants were then debriefed, and thanked for their time. The entire experiment took approximately 1 hour.

INSERT FIGURE 2 ABOUT HERE

2.4. Scoring, response definition, and statistical analyses. Data were analysed using mixed-model ANOVAs in IBM SPSS Statistics 25. Significant interactions ($\alpha = .05$, Pillai's trace statistics of the multivariate solution reported) were followed-up with pairwise comparisons. The current report is focussed on explicit valence ratings, startle magnitude during acquisition, and reaction times from the affective priming task. Affective priming error data, startle blink latency data from acquisition and extinction, and startle blink magnitude data from extinction and electrodermal responses to forward CS and US onset were analysed, and results are included in the supplementary materials at https://osf.io/q46mp/. SCRs to backward CSs and SCRs during extinction were not analysed.

2.4.1. *Startle blink magnitude.* The raw EMG signal was notched at 50 Hz, high and low pass filtered at 30 and 500 Hz, and rectified and smoothed by using 5 consecutive measurement points to calculate a moving average. Startle blink magnitude was defined as the largest response that occurred

within 120ms of the startle probe, beginning 20-60ms after startle probe onset. A non-response trial was defined as a trial where response onset could not be visually identified within this window. Non-response trials were scored as zeros and included in the analysis. A trial was defined as missing if the response could not be visually differentiated from background EMG activity, or if a blink occurred between the startle probe onset and the response window onset (Experiment 1 = 2.73% and Experiment 2 = 1.69%). Individual differences and variation across individual trials were controlled for by blocking trials and transforming raw data into *T*-scores. Blocks of two trials were created for each US valence for forward and backward conditioning separately, resulting in three blocks per condition. *T*-scores were then subjected to a 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US valence: positive vs negative; within-participants) × 3 (Block: 1, 2, 3; within-participants) mixed model ANOVA. One participant from each group was excluded for failing to respond to more than 50% of the startle probes resulting in 62 participants being included in the startle blink magnitude analyses.

2.4.2. *Explicit valence ratings.* Participants rated each family of CSs on how much they liked them. The higher the rating, the more positive the valence of the family. These data were subjected to a 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) × 2 (US Valence: positive vs negative; within-participants) mixed model ANOVA.

2.4.3. *Affective priming.* Participants categorised positive or negative target words following the presentation of the CS primes. Incorrect categorisation of target words were scored as errors. Responses faster than 300ms and slower than 1000ms were also scored as errors, as they were deemed to be outside the window of a response suggestive of task adherence. Participants who made more than 25% errors were removed from the analyses, leaving a total of 59 participants (low intensity group, $n = 29$). Reaction times to each target word following CSs from the same set were averaged to provide mean reaction times, resulting in means for each CS set for positive target words and each CS set for negative target words. Percentage of errors was also calculated for each set for each target word. These means were then used to calculate priming scores (incongruent trials [CSs paired with positive USs/negative target words + CSs paired with negative USs/positive target words] – congruent

trials [CSs paired with positive USs/positive target words + CSs paired with negative USs/negative target words]), which were subjected to separate 2 (Group: low vs high intensity; between-participants) × 2 (Conditioning Type: forward vs backward; within-participants) mixed model ANOVAs.

2.5. Manipulation checks. Groups did not differ on gender, $\chi 2$ (1, N = 64) = < .001, p > .999, ethnicity, $\chi 2$ (4, N = 64) = 2.1, p = .718, or age, t (62) = 0.88, p = .378, d = 0.23. Post-experimental valence and intensity ratings of the USs and startle probe were subjected to separate 2 (Group: low vs high intensity; between-participants) × 3 (Valence: positive US vs negative US vs startle probe; within-participants) mixed model ANOVAs. For the valence ratings, a main effect of US valence, F (2, 61) = 586.63, p < .001, η_p^2 = .95, showed that the positive US was rated more positively than the negative US, (M = 2.30, SD = 0.63 vs M = -2.20, SD = 0.80), t (62) = 33.83, p < .001, d = 4.23, and the startle probe, (M = 2.30, SD = 0.63 vs M = -1.42, SD = 1.11), t (62) = 22.40, p < .001, d = 2.80, and that the startle probe was rated more positively than the negative US, (M = -1.42, SD = 1.11 vs M = -2.20, SD = 0.80), t (62) = 5.10, p < .001, d = 0.64. For the intensity ratings, a main effect of US valence, F (2, 61) = 53.14, p < .001, η_p^2 = .635, was qualified by a Group × Valence interaction, F (2, 61) = 3.27, p = .045, η_p^2 = .097, suggesting that the negative US was more intense in the high intensity group than the low intensity group, (M = 4.72, SD = 1.02 vs M = 3.72, SD = 1.37), t (62) = 3.30, p = .002, d = 0.84. In the low intensity group, the positive US, (M = 1.62, SD = 1.52), was rated as less intense than the negative US, (M = 3.72, SD = 1.37), t (61) = 6.20, p < .001, d = 1.10, and the startle probe, (M = 4.09, SD = 1.28), t (61) = 6.75, p < .001, d = 1.19, while, the negative US and startle probe did not differ, (M = 3.72, SD = 1.37 vs M = 4.09, SD = 1.28), t (61) = 1.49, p = .141, d = 0.26. In the high intensity group, the negative US, (M = 4.72, SD = 1.02), was rated as more intense than the startle probe, (M = 4.22, SD = 1.43), t (61) = 1.98, p = .051, d = 0.35, and the positive US, (M = 1.97, SD = 1.45), t (61) = 8.14, p < .001, d = 1.44, and the startle probe was rated as more intense than the positive US, (M = 4.22, SD = 1.43 vs M = 1.97, SD = 1.45), t (61) = 6.15, p < .001, d = 1.09. The startle magnitude during the inter-trial intervals did not differ between groups for acquisition, t (60) = 1.27, p = .209, d = 0.32.

3. Results

3.1. Startle blink magnitude – Acquisition. Figure 3 suggests larger startle responses during CSs presented before negative USs compared to CSs presented before positive USs in both groups, suggesting an assimilation effect. Startle responses during backward CSs following positive USs appear larger compared to responses during backward CSs following negative USs, suggesting a contrast effect. Main effects of conditioning type, F (1, 60) = 158.93, $p < .001$, $\eta_p^2 = .726$, and block, F (2, 59) = 120.33, $p < .001$, $\eta_p^2 = .803$, were qualified by a Conditioning Type × Block interaction, F (2, 59) = 14.59, $p < .001$, $\eta_p^2 = .331$, a Group × Conditioning Type interaction, F (1, 60) = 22.29, $p < .001$, $\eta_p^2 = .271$, and a Conditioning Type × US Valence interaction, F (1, 60) = 10.96, $p = .002$, $\eta_p^2 = .154$. As all follow-up analyses for the Conditioning Type × Block interaction were significant, difference scores between blocks 1 and 2, 2 and 3, and 1 and 3, were calculated for forward and backward conditioning, and subjected to paired sample *t*-tests comparing forward and backward conditioning for each difference score. Forward conditioning showed a larger difference than backward conditioning between blocks 1 and 2, t (61) = 5.20, $p < .001$, $d = 0.66$, and blocks 1 and 3, t (61) = 4.91, $p < .001$, $d = 0.62$, with no difference appearing between blocks 2 and 3, t (61) = 0.53, $p = .60$, $d = 0.07$. The Group × Conditioning Type interaction showed there was no difference in startle blink magnitude between the low intensity and high intensity groups during forward CSs, F (1, 60) = 0.88, $p = .352$, $\eta_p^2 = .014$, and that startle blink magnitude was smaller in the high intensity group, than the low intensity group, F (1, 60) = 37.73, $p < .001$, $\eta_p^2 = .386$, during backward CSs, indicative of greater relief at US offset following high intensity stimuli. The Conditioning Type × US Valence interaction showed blink magnitude was larger during positive backward CSs than negative backward CSs, F (1, 60) = 10.56, $p = .002$, $\eta_p^2 = .150$, which is indicative of a contrast effect. No differences between valence for forward conditioning was found, F (1, 60) = 2.36, $p = .129$, $\eta_p^2 = .038$. However, when only participants who passed the memory test were included in the analysis, startle blink magnitude during negative forward CSs was larger than during positive forward CSs, F (1, 51) = 4.81, $p = .033$, $\eta_p^2 = .086$.

INSERT FIGURE 3 ABOUT HERE

3.2. Explicit valence ratings. Figure 4 suggests assimilation effects for forward conditioning as CSs paired with positive USs are rated as more pleasant than CSs paired with negative USs. Contrast effects seem to be present for backward CSs for both groups, as CSs paired with positive USs are rated as less pleasant than CSs paired with negative USs. Main effects of conditioning type, F (1, 62) = 27.13, $p < .001$, $\eta_p^2 = .304$, and US valence, F (1, 62) = 130.41, $p < .001$, $\eta_p^2 = .678$, and a Conditioning Type × US Valence interaction, F (1, 62) = 212.47, $p < .001$, $\eta_p^2 = .774$, were qualified by a Group × Conditioning Type × US Valence interaction, F (1, 62) = 8.17, $p = .006$, $\eta_p^2 = .116$. Follow up analyses revealed a contrast effect in the high intensity group as backward CSs paired with positive USs were rated as less pleasant than backward CSs paired with negative USs, F (1, 62) = 29.59, $p < .001$, $\eta_p^2 = .323$, but not in the low intensity group, F (1, 62) = 0.80, $p = .375$, $\eta_p^2 = .013$. Moreover, assimilation effects were found for forward conditioning in both groups, as CSs paired with positive USs were rated as more positive than CSs paired with negative USs; low intensity: F (1, 62) = 141.10, $p < .001$, $\eta_p^2 = .695$, high intensity: F (1, 62) = 188.84, $p < .001$, $\eta_p^2 = .753$.

INSERT FIGURE 4 ABOUT HERE

3.3. Affective priming – Reaction times. As shown in Figure 5, assimilation effects are suggested for forward conditioning, regardless of group. A main effect of conditioning type showed that the priming score for forward conditioning was significantly larger than that for backward conditioning, F (1, 59) = 8.48, $p = .005$, $\eta_p^2 = .126$. Moreover, the forward conditioning priming score was significantly larger than 0, t (60) = 5.10, $p < .001$, $d = 0.64$, while the backward conditioning priming score was not, t (60) = 1.84, $p = .070$, $d = 0.24$.

INSERT FIGURE 5 ABOUT HERE

4. Discussion

Experiment 1 aimed to replicate the backward CS contrast effects shown in Moran and Bar-Anan (2013) on explicit valence ratings while measuring the startle blink reflex and to determine whether US intensity could account for the differing dissociations between implicit and explicit measures reported by Moran and Bar-Anan (2013) and Andreatta et al. (2010). Blink magnitude data revealed a contrast effect for backward CSs in both groups, and an assimilation effect for forward CSs which was significant only in participants who recalled the contingencies. Startle blink responses were smaller during backward CSs in the high intensity group than the low intensity group. This suggests greater relief at the offset of the high intensity USs compared to the low intensity USs, which was expected only for negative USs. Increasing the volume of the USs may have resulted in both USs becoming less pleasant, rendering the positive US slightly unpleasant. Furthermore, this would explain the lack of differentiation between forward CSs paired with positive and negative USs, as the positive US becoming slightly more negative in the high intensity group would wash out any effects of differential US valence.

We found backward CS contrast effects on explicit valence ratings for the high intensity group only and assimilation effects for forward CSs in both groups. Thus, the intensity manipulation functioned as expected, resulting in larger backward CS contrast effects in the high intensity group than the low intensity group. However, unexpectedly, the backward CS contrast effect in the low intensity group was not significant. This may be due to the fact that our low intensity USs were less intense as those used in Moran and Bar-Anan (2013), which supports the idea that US intensity does in fact moderate backward CS contrast effects. It is also possible that presenting startle probes during acquisition made the USs seem less intense, therefore requiring more intense USs to produce a contrast effect. This explanation is supported by the fact that a backward CS contrast effect occurred in the high intensity group, as startle probes were also present during acquisition in this group.

Priming scores provide support for assimilation effects for forward CSs regardless of group, indicating a more negative evaluation of CSs preceding negative USs. Priming scores for backward CSs did not reveal any acquisition of differential valence as a function of US valence. Thus, there was no support for the hypothesis that for backward conditioned CSs contrast effects would appear in the

high intensity condition and assimilation effects in the low intensity condition in the affective priming task. This pattern of results, which is in contrast to that seen for explicit evaluations, may reflect a difference in sensitivity between explicit and implicit measures.

In summary, we were able to partially replicate the findings from Moran and Bar-Anan (2013) while measuring physiology and manipulating US intensity. We showed that startle blinks elicited during backward CSs largely follow the pattern of explicit valence ratings in this paradigm. Our findings also demonstrate that US intensity cannot account for the differing dissociations between implicit and explicit measures reported by Moran and Bar-Anan (2013) and Andreatta et al. (2010).

5. Experiment 2

The findings from Experiment 1 and work by Andreatta and colleagues indicate that backward CSs following aversive USs have acquired positive valence, as startles elicited during CSs presented after aversive stimuli (CS+/CSneg) are smaller than startles elicited during CSs presented alone (CS-) or CSs presented after positive USs (CSpos; Andreatta et al., 2010; Andreatta et al., 2013). This effect is known as 'relief learning', as the positive effect that occurs after the offset of an aversive stimulus elicits feelings of relief (Deutsch et al., 2015; Gerber et al., 2014). It has been proposed that a similar valence reversal would be observed after positive stimuli, i.e., that the offset of a positive stimulus would elicit negative feelings, which would result in stimuli presented after a positive US acquiring negative valence (B-CSpos; see Felsenberg et al., 2014 for demonstration in honeybees; Gerber et al., 2014). However, in absence of a neutral baseline condition, it is difficult to determine whether the relative difference in backward CS valence observed in Experiment 1 reflects positive valence for B-CSneg and negative valence for B-CSpos. If this were the case, we would expect startle modulation during backward CSs paired with negative, neutral, and positive USs to follow a linear trend. This would be shown by smaller startles during the B-CSneg (suggesting positive valence) compared to during a CS paired with a neutral US (B-CSneut), and larger startles during the B-CSpos (suggestive of negative valence) compared to during the B-CSneut. On the other hand, if valence acquisition during pleasant and aversive backward conditioning are qualitatively different, and negative valence does not occur at the offset of a positive stimulus, a quadratic trend would be expected. In this case, startle responses during the B-CSneut would be larger than during

both the B-CSpos and B-CSneg. This would suggest positive valence for both the B-CSpos and B-CSneg, regardless of any observed difference between them (such as larger startle inhibition during the B-CSneg relative to the B-CSpos). This would mean that the backward CS contrast effects observed on the startle blink reflex in Experiment 1 and in Andreatta and colleagues' work would be the sole result of startle inhibition during the B-CSneg, with no opponent process occurring for backward conditioning with the positive US (Andreatta et al., 2010; Andreatta et al., 2013).

To investigate whether a linear or quadratic trend best represents startle blink magnitude during backward conditioning, we added trials with a neutral US to the paradigm used in Experiment 1. The addition of these neutral US trials would have required participants to learn six contingencies; start positive US, stop positive US, start neutral US, stop neutral US, start negative US, and stop negative US. In order to make the task less challenging, we decided to remove the forward CSs from the procedure. This meant that participants only had to learn three contingencies (stop positive US, stop neutral US, and stop negative US), which increased the likelihood of correct contingency recall. Removing the forward CS from a CS-US-CS paradigm has been shown to have no effect on the startle blink reflex during backward conditioning (Andreatta et al., 2010; Andreatta et al., 2013). For explicit valence ratings, however, Andreatta et al. (2013) found that a concurrent forward and backward conditioning design (CS-US-CS) led to backward CS contrast effects, while simple backward conditioning (US-CS) led to backward CS assimilation effects. As we retained the instructions from Experiment 1 and presented them in a backward conditioning paradigm (US-CS), we were afforded the opportunity to test whether a backward CS contrast effect as predicted by the instructions, or a backward CS assimilation effect as predicted by the paradigm, would occur. No explicit hypothesis was proposed.

Experiment 2 was designed to assess whether startle modulation during backward conditioning with positive, neutral and aversive USs would reveal a linear or quadratic pattern. We also wanted to assess whether backward CS contrast effects would still be observed on explicit valence ratings when no forward CSs were presented due to the instructions highlighting the role of the backward CSs. To determine this, participants were told to learn which CSs stopped the pleasant, neutral, and aversive USs (backward conditioning: US-CS). It was hypothesised that startle blink

modulation would follow a significant linear trend indicative of backward CS contrast effects. Largest responses were expected during CSs following positive USs, and smallest responses during CSs following negative USs.

6. Method

6.1. Participants. Thirty-eight undergraduate students (25 female) from the School of Psychology at Curtin University participated in this experiment for course credit, *M* age = 22, *SD* = 6.89. As in Experiment 1, sample size was based on previous research (Andreatta et al., 2013; Andreatta et al., 2010). Two participants failed the recollective memory test. Analyses were run with and without these participants and the pattern of results do not differ, hence results from the full sample are reported.

6.2. Apparatus/Stimuli. The apparatus and stimuli were the same as for Experiment 1, except only three of the alien families were used as CS sets (yellow, purple, and red), USs from the low intensity group were used (positive US: 47 dBA; negative US: 72 dBA), and a neutrally valenced auditory US was added. Low intensity USs were used as they showed a clearer pattern of startle modulation in Experiment 1 (despite resulting in non-significant backward CS contrast effects on explicit valence ratings). The neutral US was selected by asking participants in a pilot study (n = 20) to provide valence and intensity ratings for 9 neutral stimuli chosen from the International Affective Digitized Sounds (2nd Edition; IADS-2) database. These stimuli were matched in volume to the positive and negative USs from the low intensity group in Experiment 1. The stimulus that was rated as the most neutral on valence and that participants could accurately describe was chosen as the neutral US. This stimulus was the sound of a train passing a train station (sound #425 from IADS-2, 56 dBA; pilot study reported at https://osf.io/q46mp/).

6.3. Procedure. The procedure was the same as in Experiment 1, except that during acquisition, only a backward conditioning procedure (US-CS) was used. Participants were presented with eight negative, eight positive, and eight neutral trials. Trials were presented in a pseudo random order, with no more than two consecutive trials of the same valence. On each trial, USs were presented for 10, 15, 20, 25, or 30 seconds, and CSs were presented for 8 seconds, beginning 2

seconds before US offset (see Figure 6 for a depiction of a positive US trial) . This resulted in USs and CSs overlapping for 2 seconds. Startle probes were assigned randomly and presented on 18 of the 24 trials, with six probes occurring in each valence condition (see Figure 7 for a depiction of startle probe timing). Half of the ITIs were probed, totalling 12 startle probes. During extinction, each CS from each of the three CS sets was presented once, and two CSs from each set were presented twice for a total of six presentations per set. CSs were shown for 8s each, for a total of 18 presentations. Four of the six presentations of each CS set were probed, totalling 18 startle probes. Half of the ITIs were probed, totalling nine startle probes. The instructions used were the same as in Experiment 1, except participants were told each of the three families would stop one of the sound USs[3]. In the affective priming task sixty trials were presented as only three CS sets were used in the conditioning task. Each of the three sets was presented once with positive and negative words. All other details of the affective priming task were the same as in Experiment 1. In the memory test, participants were shown each family separately, and asked:

What was the role of the creatures in this picture? 1. Stopped the human sound. 2. Stopped the musical sound. 3. Stopped the metropolitan sound.

The post-experimental questionnaire was the same as in Experiment 1, with the addition of asking for valence and intensity ratings of the metropolitan sound.

INSERT FIGURE 6 ABOUT HERE

INSERT FIGURE 7 ABOUT HERE

6.4. Scoring, response definition, and statistical analyses. Data were analysed using repeated measures ANOVAs. All other details are the same as in Experiment 1 unless noted below.

[3] An error in the instructions was spotted by the 27th participant. Instead of saying "The three families of creatures:" it said "The four families of creatures:", and then showed only three families. Participants after this were asked if they noticed anything about the instructions, and then if they noticed if it said "four families" at any point upon completion of the experiment. Of the 11 participants asked, four of them noticed. Three participants said they thought it was a typo, and one of them thought it was referring to the fact that there were four aliens in each family.

6.4.1. *Startle.* *T*-scores were subjected to a 3 (US Valence: Positive vs neutral vs negative) × 3 (Block: 1, 2, 3) repeated measures ANOVA with a subsequent trend analysis. No participants were excluded.

6.4.2. *Explicit valence ratings.* Ratings were subjected to a repeated measures ANOVA and subsequent trend analysis comparing (US valence: Positive vs neutral vs negative). No participants were excluded.

6.4.3. *Affective priming.* As neutrally valenced USs were presented, scores based on the difference between positive and negative target words for each prime valence were calculated for reaction times and errors. Assimilation effects are represented by negative scores for positive CS primes and by positive scores for negative CS primes. These scores were subjected to a repeated measures ANOVA trend analysis (US valence: Positive vs neutral vs negative). No participants were excluded. Analysis of the error data did not add substantially to the current report and is reported in the supplementary materials at https://osf.io/q46mp/.

6.5. Manipulation checks. Post-experimental valence and intensity ratings of the USs and startle probe were subjected to separate repeated measures ANOVAs (positive US vs neutral US vs negative US vs startle probe; within-participants). A linear relationship was found for US valence ratings, F (3, 35) = 146.87, $p < .001$, $\eta_p^2 = .926$. The positive US (M = 2.40, SD = 0.72) was rated significantly more positively than the neutral US (M = 0.03, SD = 1.05), t (37) = 11.20, $p < .001$, d = 1.82, the negative US (M = -1.82, SD = 0.80), t (37) = 19.10, $p < .001$, d = 3.10, and the startle probe (M = -1.84, SD = 0.97), t (37) = 19.14, $p < .001$, d = 3.10. The neutral US (M = 0.03, SD = 1.05) was rated significantly more positively than the negative US (M = -1.82, SD = 0.80), t (37) = 8.19, $p < .001$, d = 1.33, and the startle probe (M = -1.84, SD = 0.97), t (37) = 8.35, $p < .001$, d = 1.35. There were no differences between the negative US (M = -1.82, SD = 0.80) and the startle probe (M = -1.84, SD = 0.97), t (37) = 0.13, p = .898, d = 0.02. A linear relationship was also observed for US intensity ratings, F (3, 35) = 72.74, $p < .001$, $\eta_p^2 = .862$. The startle probe (M = 4.55, SD = 1.06) was rated as significantly more intense than the negative US (M = 3.76, SD = 1.34), t (37) = 3.53, p = .001, d = 0.57, the neutral US (M = 2.18, SD = 1.33), t (37) = 10.26, $p < .001$, d = 1.66, and the positive US (M = 1.34, SD = 1.36), t (37) = 14.79, $p < .001$, d = 2.40. The negative US (M = 3.76, SD = 1.34) was

rated as significantly more intense than the neutral US (M = 2.18, SD = 1.33), t (37) = 5.29, p < .001, d = 0.86, and the positive US (M = 1.34, SD = 1.36), t (37) = 9.30, p < .001, d = 1.51. The neutral US (M = 2.18, SD = 1.33) was rated as significantly more intense than the positive US (M = 1.34, SD = 1.36), t (37) = 3.75, p = .001, d = 0.61.

7. Results

7.1. Startle blink magnitude – Acquisition. Figure 8 shows larger responses during CSs following positive USs than during CSs following neutral and negative USs which decreased across blocks. This was confirmed by main effects of US valence, F (2, 36) = 8.77, p = .001, η_p^2 = .328, and block, F (2, 36) = 23.58, p < .001, η_p^2 = .567. Responses during CSs following the positive US were marginally larger than responses during CSs following the neutral US, t (36) = 1.97, p = .057, d = 0.32, and significantly larger than responses during CSs following the negative US, t (36) = 4.23, p < .001, d = 0.69. Responses during CSs paired with the neutral US were significantly larger than responses during CSs paired with the negative US, t (36) = 2.19, p = .034, d = 0.36. Responses at block 1 were larger than blocks 2, t (36) = 3.40, p = .002, d = 0.55, and 3, t (36) = 6.80, p < .001, d = 1.10, and responses at block 2 were larger than block 3, t (36) = 3.89, p < .001, d = 0.63. Tests of within-subject contrasts showed only linear trends for US valence, F (1, 37) = 17.90, p < .001, η_p^2 = .326, and block, F (1, 37) = 46.29, p < .001, η_p^2 = .556.

INSERT FIGURE 8 ABOUT HERE

7.2. Explicit valence ratings. Figure 9 shows a linear trend for US valence suggestive of an assimilation effect, as CSs following the positive US were rated more positively than CSs following the neutral US, and CSs following the neutral US were rated as more positive than CSs following the negative US. This was confirmed by a significant one-way repeated measures ANOVA, F (2, 36) = 9.16, p = .001, η_p^2 = .337. CSs paired with the positive US were rated as significantly more pleasant than CSs paired with the neutral US, t (36) = 2.40, p = .022, d = 0.39, and the negative US, t (36) = 4.31, p < .001, d = 0.70, and CSs paired with the neutral US were rated as significantly more pleasant

than CSs paired with the negative US, t (36) = 3.46, p = .001, d = 0.56. Tests of within-subject contrasts showed only a significant linear trend for US valence, F (1, 37) = 18.61, p < .001, η_p^2 = .335.

INSERT FIGURE 9 ABOUT HERE

7.3. Affective priming – Reaction times. Figure 10 shows a linear trend suggestive of an assimilation effect, although the main effect for US valence was not significant, F (2, 36) = 1.65, p = .206, η_p^2 = .084, and the trend for US valence was only marginal, F (1, 37) = 3.14, p = .085, η_p^2 = .078.

INSERT FIGURE 10 ABOUT HERE

8. Discussion

Experiment 2 aimed to determine whether both positive and negative USs lead to opposing emotional responses at their offset (shown by linear startle modulation), and whether instructions highlighting the role of the backward CSs exert their effect in a backward conditioning only paradigm. Linear trends for explicit valence ratings and affective priming (only trending) suggest that backward CS assimilation effects occurred, despite presenting instructions that should support backward CS contrast effects. This provides evidence that the backward CS contrast effects driven by instructional manipulations, as reported for instance in Experiment 1, were not due to demand characteristics, as the same pattern of results should have emerged here. Removing the forward CS appeared to have no impact on startle blink magnitude and inclusion of the neutral US pairing showed that startle blink modulation also followed a linear trend. This trend was suggestive of a contrast effect as startle blink magnitude was larger during CSs following the positive US than during CSs following neutral and negative USs. This confirmed that an opponent process mirroring relief occurs at the offset of positive stimuli, which to our knowledge is the first demonstration that the offset of both positive and negative stimuli elicits an opposing emotional reaction in humans which can be indexed by startle blink reflexes.

While emotional responses elicited at the offset of valenced stimuli seem the most plausible explanation for the pattern of startle modulation, it is also possible that the instructional manipulation highlights the role of the CSs and therefore affects startle modulation. This is because the same pattern of startle modulation is expected from the instructions, i.e. CSs stopping the negative US should become positive, and CSs stopping the positive US should become negative. Even though the instructions did not affect explicit valence ratings, we cannot rule out this conclusion because explicit valence ratings and the startle blink reflex have been shown to dissociate in backward conditioning only designs (US-CS; Andreatta et al., 2013; Andreatta et al., 2010). Future research should confirm that the inverse 'relief learning' process occurring at the offset of the positive US was not due to the instructions.

9. General Discussion

The current experiments assessed whether US intensity could explain the different patterns of dissociations between explicit and implicit measures of backward CS valence reported in studies of evaluative and fear conditioning (Experiment 1), and whether a linear pattern of startle modulation suggestive of opposing emotional responses after the offset of positive and negative USs would emerge during backward conditioning (Experiment 2). Moreover, Experiment 1 assessed whether due to the instructional manipulation used, startle modulation and explicit valence ratings would reveal backward CS contrast effects, and Experiment 2 tested whether presenting similar instructions in a backward conditioning only procedure would lead to assimilation or contrast effects. The current results suggest that backward conditioning leads to the same pattern of startle modulation regardless of whether forward and backward conditioning are trained concurrently or only backward conditioning is assessed, and whether a neutral US is included in the backward conditioning procedure. The intensity effect observed on the startle response in Experiment 1 shows that responses overall were smaller in the high intensity group, not that the pattern of startle modulation differs as a function of US valence or conditioning type. Hence, US intensity does not moderate backward CS contrast effects on the startle response. On the other hand, explicit valence ratings appear sensitive to US intensity and the conditioning procedure. Explicit valence ratings revealed backward CS contrast effects during concurrent forward and backward conditioning in Experiment 1 in the high intensity

condition only, and a backward CS assimilation effect during backward conditioning in Experiment 2. Moreover, the instructional manipulation highlighting the role of the CSs appeared to have no effect on explicit valence ratings without concurrent forward conditioning in Experiment 2.

The backward conditioning results on explicit valence ratings in Experiments 1 and 2 replicate earlier work on relief learning that yielded assimilation effects when the US was unpredictable (US-CS), and contrast effects when the US was predictable (CS-US-CS; Andreatta et al., 2013). However, unlike Andreatta and colleagues, we informed participants that the backward CS would stop the US which was expected to support backward CS contrast effects on explicit valence ratings, as observed in Experiment 1 and in other studies involving relational instructions (Moran & Bar-Anan, 2013; Moran et al., 2016). Below we offer three explanations as to why contrast effects did not occur in Experiment 2.

Firstly, it is possible that the instructions were less salient in the US-CS design than in the CS-US-CS design, as participants only had to learn one relation for each of the USs. Less focus on the instructions may have resulted in weaker encoding of the proposition that families stop the USs, therefore rendering the instructions ineffective. However, the fact that only two participants failed the recollective memory test suggests this was not the case, as poor encoding of the instructions should also result in poor performance on the recollective memory test.

Second, it may be that the onset of the US is more salient in the US-CS design because there is no forward CS. This may render the US more aversive, which may then overpower the effect of the instructions leading to an assimilation effect. If this was to occur, then more intense USs should also lead to assimilation effects. Experiment 1 shows this was not the case, as the high intensity USs led to larger contrast effects than the low intensity USs.

Finally, the lack of a backward CS contrast effect may be due to temporal overshadowing. In the CS-US-CS trials the forward CS could be considered the most salient CS as it predicts the onset of the US. If overshadowing can occur across stimuli in a temporal arrangement, then the forward CS may have overshadowed the association between the backward CS and the valence of the US. This may permit the association between the backward CS and the emotional response elicited by US offset to become apparent, and/or for the instructional manipulation to take effect. In absence of the

forward CS no overshadowing occurred which permitted the development of an association between US valence and the backward CS. This explanation can also account for the findings of Andreatta et al. (2013), who showed backward CS contrast effects in the CS-US-CS design and assimilation effects in the US-CS design, in absence of any instructional manipulation. If we consider that the forward CS overshadows valence transfer from the US to the backward CS, then backward CS valence may be influenced by feelings of relief after US offset, resulting in positive ratings of a backward CS following an aversive shock. While intriguing, the temporal overshadowing account holds only if we assume that the startle reflex and valence ratings reflect different learning mechanisms, as startle modulation showed the same pattern of results in both the CS-US-CS and US-CS designs.

The studies we have presented confirm that the offset of a positive US leads to negative emotion in humans (disappointment). Startle responses were larger during CSs presented at the offset of the positive US in comparison to CSs presented after neutral and negative USs. This process of disappointment learning mirrors that of relief learning which occurs during backward conditioning at the offset of a positive US. While it is early to speculate on potential clinical implications of disappointment learning, this phenomenon may hold explanatory value for those with affective disorders who tend to avoid situations in which they may experience pleasure.

In the case of avoiding pleasure, the pleasurable experience could be considered a positive US, and the offset of this pleasurable experience may result in disappointment. In terms of disappointment learning, any stimulus that is present at the offset of this pleasurable experience may acquire negative valence. This means that if any component of the pleasurable experience persists once pleasure is no longer being experienced, this component may become negative. The result would be a pleasurable experience that is now remembered as disappointing. In addition to this, the disappointment experienced at the offset of the pleasurable experience may also serve as a punisher that reduces the likelihood that an individual will partake in the pleasurable experience again. It is also possible that the disappointment experienced at the offset of the pleasurable experience is more intense and/or more salient than the pleasure itself. If so, individuals may avoid pleasant experiences all together in order to avoid the possibility of having to experience disappointment.

Another situation in which our findings may provide insight is substance abuse. Substance abuse may be motivated by the reduction of negative experiences through substance use which results in a pleasant experience. The offset of this pleasant experience may then result in disappointment, which instead of resulting in disappointment learning that may dissuade future substance abuse, leads to substance use again to provide relief from disappointment. In this scenario, disappointment could be considered a negative US and substance use itself could serve as a stimulus that initially provides relief from the negative US. It then becomes a backward CS that elicits positive feelings from being paired with the offset of an aversive event (relief learning). The end result is a vicious cycle where disappointment, relief, and relief learning serve to perpetuate substance use. Future research should investigate whether the concepts of disappointment and disappointment learning can be used to further our understanding of affective and substance use disorders. A limitation worth considering is that using low intensity USs in Experiment 2 reduced the chance to find backward CS contrast effects on explicit valence ratings, as no backward CS contrast effect was observed in the low intensity group in Experiment 1. However, such an effect of the low intensity USs seems unlikely as a clear assimilation effect was found in Experiment 2 and because startle blink magnitude shows that low intensity USs were sufficiently intense to elicit relief learning in both experiments. Moreover, these findings replicate that of Andreatta et al. (2013), suggesting that using low intensity USs did not preclude the observation of a backward CS contrast effect.

In summary, the current experiments show that US intensity does not moderate backward CS contrast effects to the point of attaining different patterns of startle modulation. Moreover, it shows that removing concurrent forward conditioning and adding a neutral US does not affect startle modulation during backward CSs and that both pleasant and aversive stimuli lead to contrasting emotional responses at their offset. Finally, assimilation effects were observed on explicit valence ratings in a backward conditioning paradigm (US-CS), even when relational instructions that should support backward CS contrast effects were presented. This suggests that backward conditioning is affected by simultaneous forward conditioning and that these events have a larger effect on learning than relational instructions that emphasise a-priori propositions about the relationships between stimuli.

10. 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*(53), 1-12, https://doi.org/10.3389/fnsys.2013.00053

Andreatta, M., Mühlberger, A., & Pauli, P. (2016). When does pleasure start after the end of pain? The time course of relief. *Journal of Comparative Neurology, 524,* 1653-1667. https://doi.org/10.1002/cne.23872

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*. https://doi.org/10.1098/rspb.2010.0103

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. https://doi.org/10.1016/j.nlm.2017.09.005

Annau, Z., & Kamin, L. J. (1961). The conditioned emotional response as a function of intensity of the US. *Journal of Comparative and Physiological Psychology, 54, 428-432*. https://doi.org/10.1037/h0042199

Bitar, N., Marchand, S., & Potvin, S. (2018). Pleasant pain relief and inhibitory conditioned pain modulation: A psychophysical study. *Pain Research and Management, 2018,* 1-8. https://doi.org/10.1155/2018/1935056

Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1990). Startle reflex modification: Emotion of attention? *Psychophysiology, 27,* 513-522. https://doi.org/10.1111/j.1469-8986.1990.tb01966.x

Bradley, M. M., & Lang, P. J. (2007). *The International Affective Digitized Sounds (2nd Edition; IADS- 2): Affective ratings of sounds and instruction manual. Technical report B-3.* University of Florida, Gainesville, Fl.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology, 10,* 230-241. https://doi.org/10.1017/S1138741600006491

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. https://doi.org/10.1037//0033-2909.127.6.853

Deutsch, R., Smith, K. J. M., Kordts-Freudinger, R., & Reichardt, R. (2015). How absent negativity relates to affect and motivation: An integrative relief model. *Frontiers in Psychology, 6(152),* 1-23. https://doi.org/10.3389/fpsyg.2015.00152

Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*, 297-327. https://doi.org/10.1146/annurev.psych.54.101601.145225

Felsenberg, J., Plath, J. A., Lorang, S., Morgenstern, L., & Eisenhardt. D. (2014). Short- and long-term memories formed upon backward conditioning in honeybees (Apis mellifera). *Learning and Memory, 21,* 37-45. https://doi.org/10.1101/lm.031765.113

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision makers. *Science, 321,* 1100-1102. https://doi.org/10.1126/science.1160769

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. https://doi.org/10.1086/527341

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. https://doi.org/10.1101/lm.032995.113

Green, L. J. S., Luck, C., Gawronski, B., & Lipp, O. V. (2019). Contrast effects in backward evaluative conditioning: Exploring effects of affective relief/disappointment versus instructional information. *Emotion*. Advance online publication. https://doi.org/10.1037/emo0000701

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

Hu, X., Gawronski, B., & Balas, R. (2017). 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. https://doi.org/10.1177/0146167216673351

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. https://doi.org/10.1016/j.jesp.2009.07.003

Leknes, S., Brooks, J. C. W., Wiech, K., & Tracey, I. (2008). Pain relief as an opponent process: A psychophysical investigation. *European Journal of Neuroscience, 28,* 794-810. https://doi.org/10.1111/j.1460-9568.2008.06380.x

Levey, A. B., & Martin, I. (1975). Classical conditioning of human evaluative responses. *Behaviour Research and Therapy, 13,* 221-226. https://doi.org/10.1016/0005-7967(75)90026-1

Lipp, O.V., Siddle, D.A.T., & Dall, P.J. (2003). The effects of unconditional stimulus valence and conditioning paradigm on verbal, skeletal, and autonomic indices of Pavlovian conditioning. *Learning and Motivation, 34,* 32-51. https://doi.org/10.1016/S0023-9690(02)00507-6

Lipp, O. V. (2006). Human fear learning: Contemporary procedures and measurement. In M. G. Craske, D. Hermans & D. Vansteenwegen (Eds.), (2006). Fear and learning: From basic processes to clinical implications (pp. 37-52). Washington: APA Books.

Luck. C. C., & Lipp, O. V. (2017). Startle modulation and explicit valence evaluations dissociate during backward fear conditioning. *Psychophysiology, 54,* 673-683. https://doi.org/10.1111/psyp.12834

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. https://doi.org/10.1016/j.ijpsycho.2008.02.005

Moran, T., and Bar-Anan, Y. (2013). The effect of object-valence relations on automatic evaluation. *Cognition and Emotion. 27.* 743-752. https://doi.org/10.1080/02699931.2012.732040

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. https://doi.org/101521soco2016345435

Mühlberger, A., Wieser, M. J., Gerdes, A. B. M., Frey, M. C. M., Weyers, P., & Pauli, P. (2015). Stop looking angry and smile, please: Start and stop of the very same facial expression differentially activate threat- and reward-related brain networks. *Social Cognitive and Affective Neuroscience, 6,* 321-329. https://doi.org/10.1093/scan/nsq039

Olson, M. A., & Fazio, R. H. (2001). Implicit attitude formation through classical conditioning. Psychological Science, 12, 413-417. https://doi.org/10.1111/1467-9280.00376

Pavlov, I. P. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. Oxford, England: Oxford University Press.

Unkelbach, C., & Fiedler, K. (2016). Contrastive CS-US relations reverse evaluative conditioning effects. *Social Cognition, 34,* 413-434. https://doi.org/101521soco2016345413

Vrana, S. R., Spence, E. L., & Lang, P. J. (1988). The startle probe response: A new measure of emotion? *Journal of Abnormal Psychology, 97,* 487-491. https://doi.org/10.1037/0021-843X.97.4.487

11. 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.

The authors have no conflicts of interest to declare.

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.

12. Footnotes

[1] Moran and Bar-Anan (2013) did not report the intensity to which their sound USs were set, but attempts to recreate them following their description suggest that they were below 90dBA.

[2] All materials, data, analysis files, and supplementary materials, are available at https://osf.io/q46mp/.

[3] An error in the instructions was spotted by the 27th participant. Instead of saying "The three families of creatures:" it said "The four families of creatures:", and then showed only three families. Participants after this were asked if they noticed anything about the instructions, and then if they noticed if it said "four families" at any point upon completion of the experiment. Of the 11 participants asked, four of them noticed. Three participants said they thought it was a typo, and one of them thought it was referring to the fact that there were four aliens in each family.

13. Figure Captions

13.1. *Figure 1.* Example of a positive US trial in Experiment 1. Forward and backward CSs were presented alone for 6 seconds and overlapping with the US for 2 seconds (8 seconds of total CS presentation). USs varied in duration for 10, 15, 20, 25, or 30 seconds. CS = Conditional stimulus, USpos = Positive unconditional stimulus.

13.2. *Figure 2.* Example of startle probe timing relative to F-CS, US, and B-CS onset and offset in Experiment 1. Startle probes were presented at 4.5 or 5.5 seconds after F-CS onset and 6.5 or 7.5 seconds after B-CS onset. F-CS = Forward conditional stimulus, US = Unconditional stimulus, B-CS = Backward conditional stimulus, speaker picture represents startle probe.

13.3. *Figure 3.* Startle blink magnitude (*T*-scores) by block (1, 2, and 3) for forward and backward CSs as a function of US valence (positive vs. negative) and US intensity (low vs high). Error bars represent 95% confidence intervals of the mean. .

13.4. *Figure 4.* Mean explicit valence ratings with individual participant values plotted for forward and backward CSs as a function of US valence (positive vs. negative) and US intensity (low vs high). Error bars represent 95% confidence intervals of the mean.

13.5. *Figure 5.* Mean priming scores (RT based) with individual participant values plotted from the affective priming task for forward and backward CSs as a function of US intensity. Positive scores suggest assimilation effects. Error bars show 95% confidence intervals of the mean.

13.6. *Figure 6.* Example of a positive US trial in Experiment 2. Backward CSs were presented alone for 6 seconds and overlapping with the US for 2 seconds (8 seconds of total CS presentation). USs varied in duration for 10, 15, 20, 25, or 30 seconds. CS = Conditional stimulus, USpos = Positive unconditional stimulus.

13.7. *Figure 7.* Example of startle probe timing relative to US and B-CS onset and offset in Experiment 2. Startle probes were presented at 6.5 or 7.5 seconds after B-CS onset. US = Unconditional stimulus, B-CS = Backward conditional stimulus, speaker picture represents startle probe.

13.8. *Figure 8.* Startle blink magnitude (*T*-scores) by block (1, 2, and 3) for backward CS as a function of US valence (positive vs. neutral vs. negative). Error bars represent 95% confidence intervals of the mean.

13.9. *Figure 9.* Mean explicit valence ratings with individual participant values plotted for backward CSs as a function of US valence (positive vs. neutral vs. negative). Error bars represent 95% confidence intervals of the mean.

13.10. *Figure 10.* Difference scores (positive target words – negative target words) for reaction times with individual participant values plotted from the affective priming task for backward CSs as a function of US valence (positive vs. neutral vs. negative). Assimilation effects are represented by negative scores for positive CS primes and by positive scores for negative CS primes. Error bars show 95% confidence intervals of the mean.

14. Table 1

Table 1: US intensities in the two groups (measured by a handheld digital sound level meter C-DSM1)

Group	Positive US – Melody	Negative US – Scream
Low Intensity	47dBA	72dBA
High Intensity	74dBA	88dBA

Note. dBA = Decibel A-Scale. US = Unconditional Stimulus

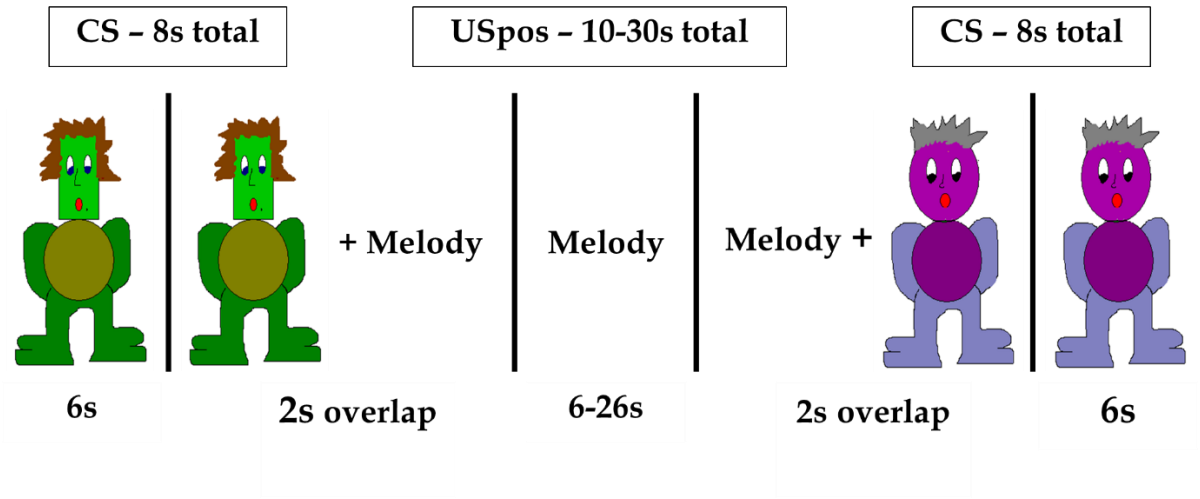

Figure 1. Example of a positive US trial in Experiment 1. Forward and backward CSs were presented alone for 6 seconds and overlapping with the US for 2 seconds (8 seconds of total CS presentation). USs varied in duration for 10, 15, 20, 25, or 30 seconds. CS = Conditional stimulus, USpos = Positive unconditional stimulus.

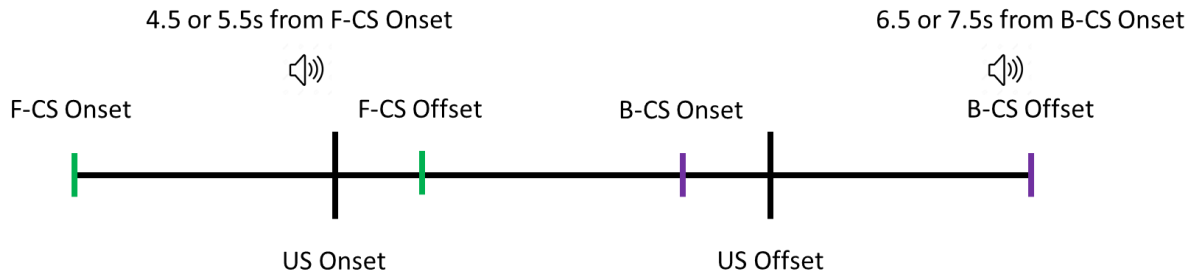

Figure 2. Example of startle probe timing relative to F-CS, US, and B-CS onset and offset in Experiment 1. Startle probes were presented at 4.5 or 5.5 seconds after F-CS onset and 6.5 or 7.5 seconds after B-CS onset. F-CS = Forward conditional stimulus, US = Unconditional stimulus, B-CS = Backward conditional stimulus, speaker picture represents startle probe.

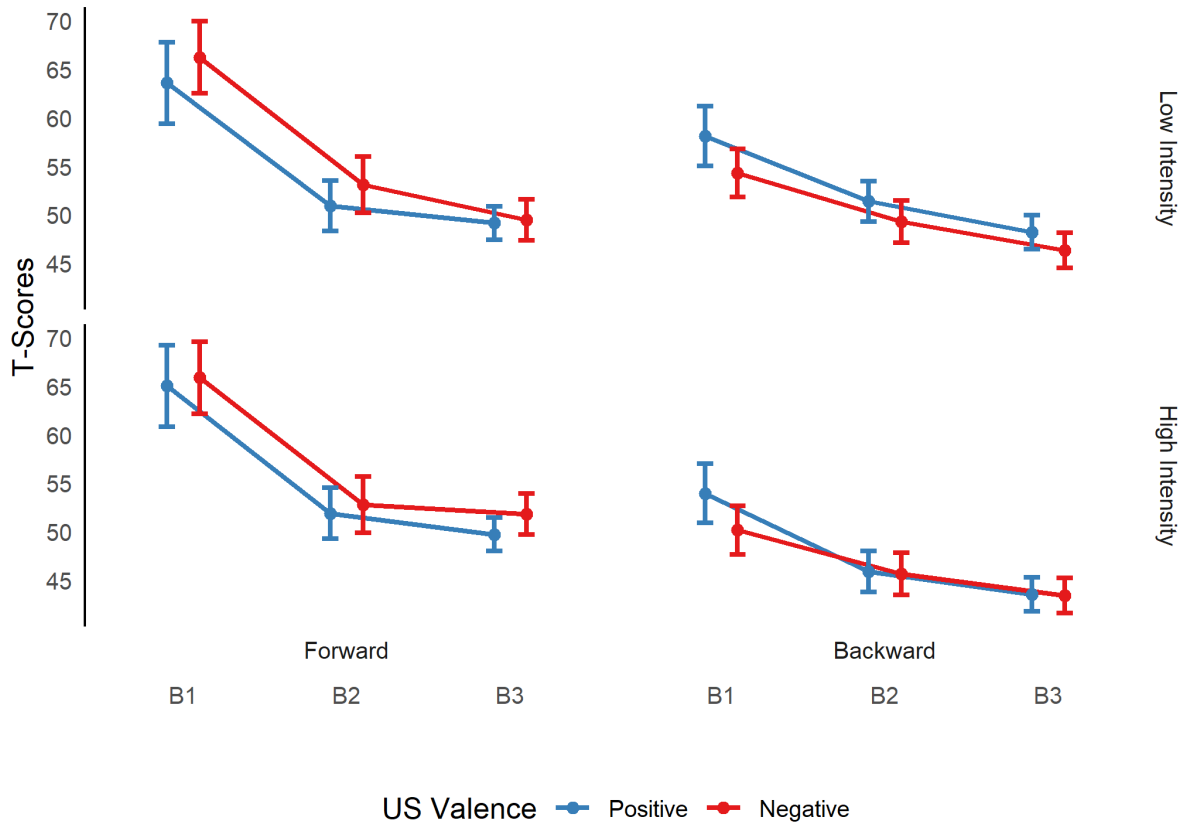

Figure 3. Startle blink magnitude (*T*-scores) by block (1, 2, and 3) for forward and backward CSs as a function of US valence (positive vs. negative) and US intensity (low vs high). Error bars represent 95% confidence intervals of the mean. .

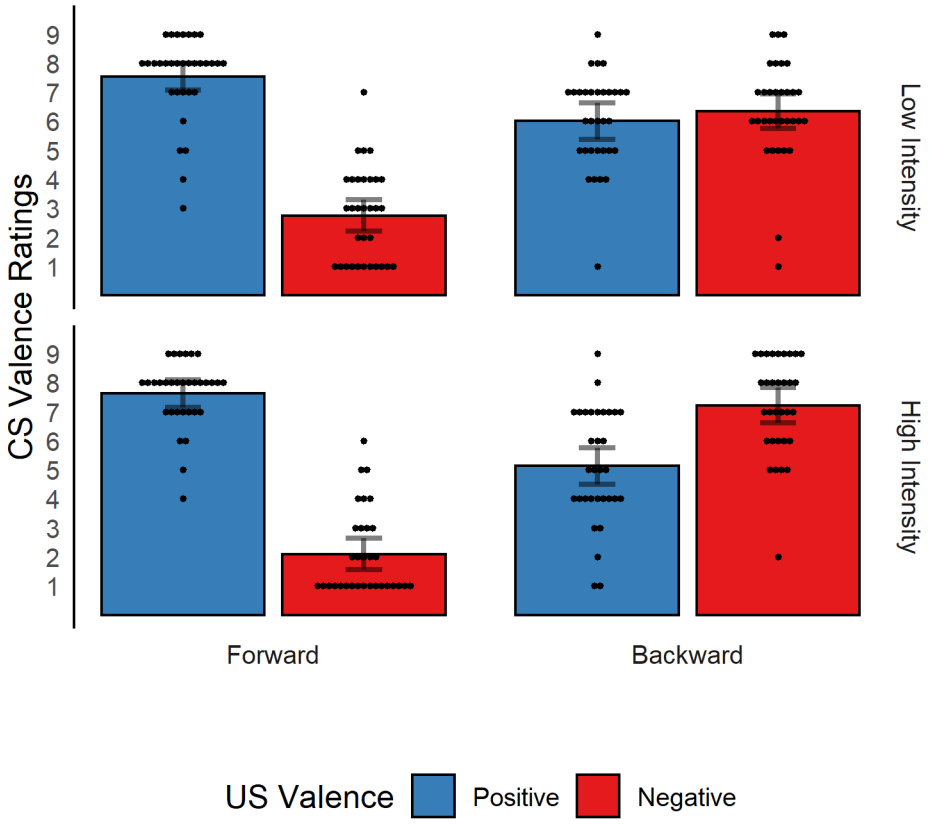

Figure 4. Mean explicit valence ratings with individual participant values plotted for forward and backward CSs as a function of US valence (positive vs. negative) and US intensity (low vs high). Error bars represent 95% confidence intervals of the mean.

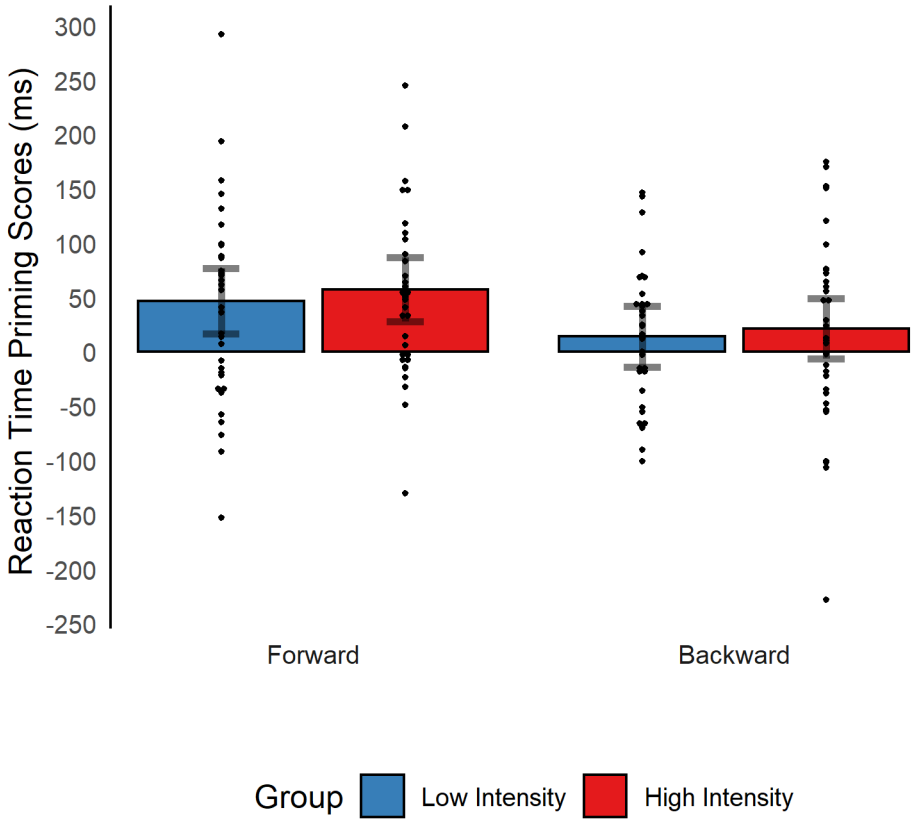

Figure 5. Mean priming scores (RT based) with individual participant values plotted from the affective priming task for forward and backward CSs as a function of US intensity. Positive scores suggest assimilation effects. Error bars show 95% confidence intervals of the mean.

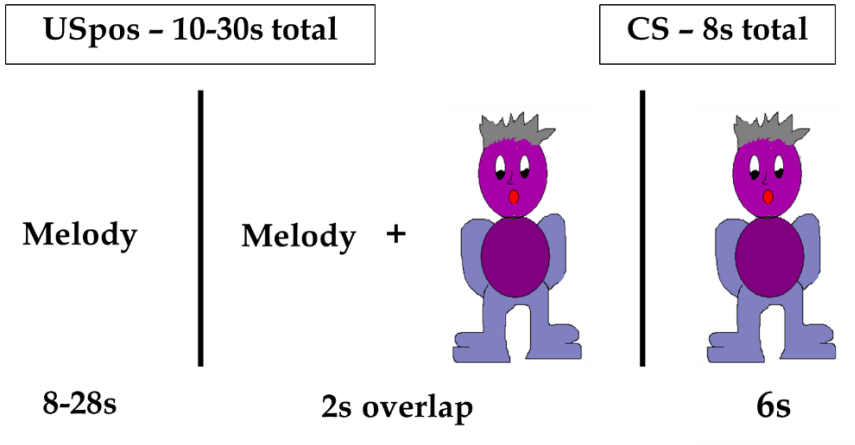

Figure 6. Example of a positive US trial in Experiment 2. Backward CSs were presented alone for 6 seconds and overlapping with the US for 2 seconds (8 seconds of total CS presentation). USs varied in duration for 10, 15, 20, 25, or 30 seconds. CS = Conditional stimulus, USpos = Positive unconditional stimulus.

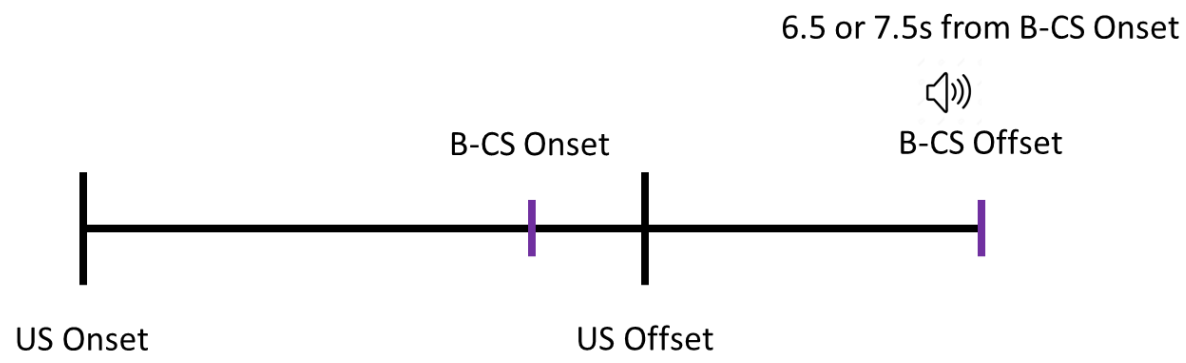

Figure 7. Example of startle probe timing relative to US and B-CS onset and offset in Experiment 2. Startle probes were presented at 6.5 or 7.5 seconds after B-CS onset. US = Unconditional stimulus, B-CS = Backward conditional stimulus, speaker picture represents startle probe.

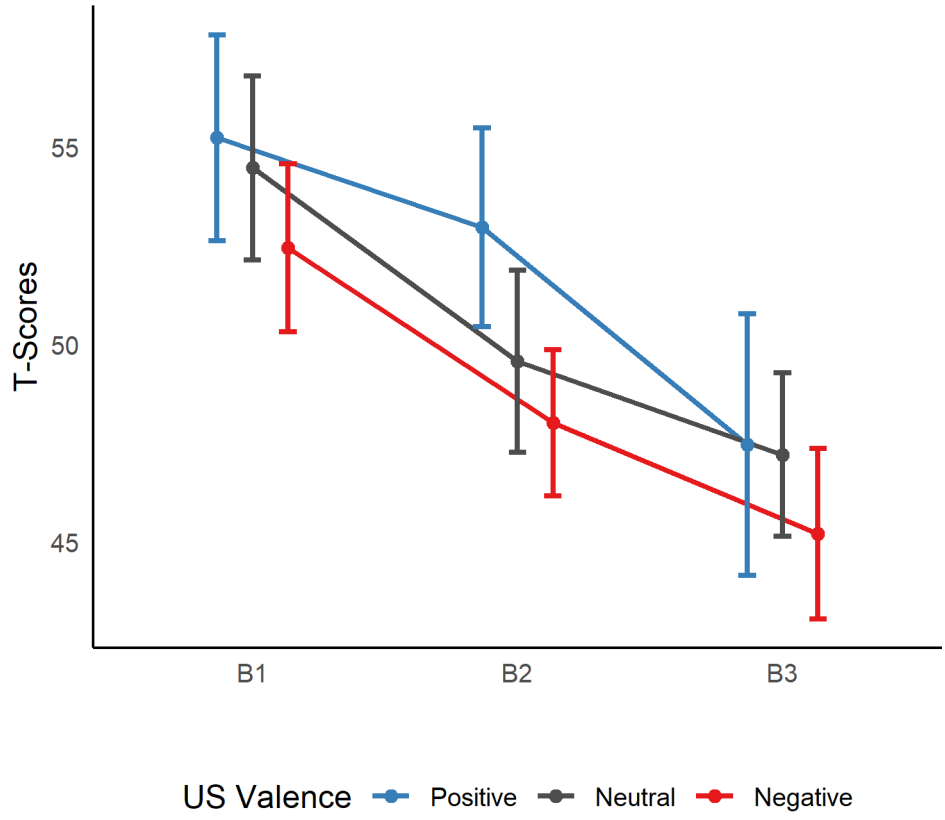

Figure 8. Startle blink magnitude (*T*-scores) by block (1, 2, and 3) for backward CS as a function of US valence (positive vs. neutral vs. negative). Error bars represent 95% confidence intervals of the mean.

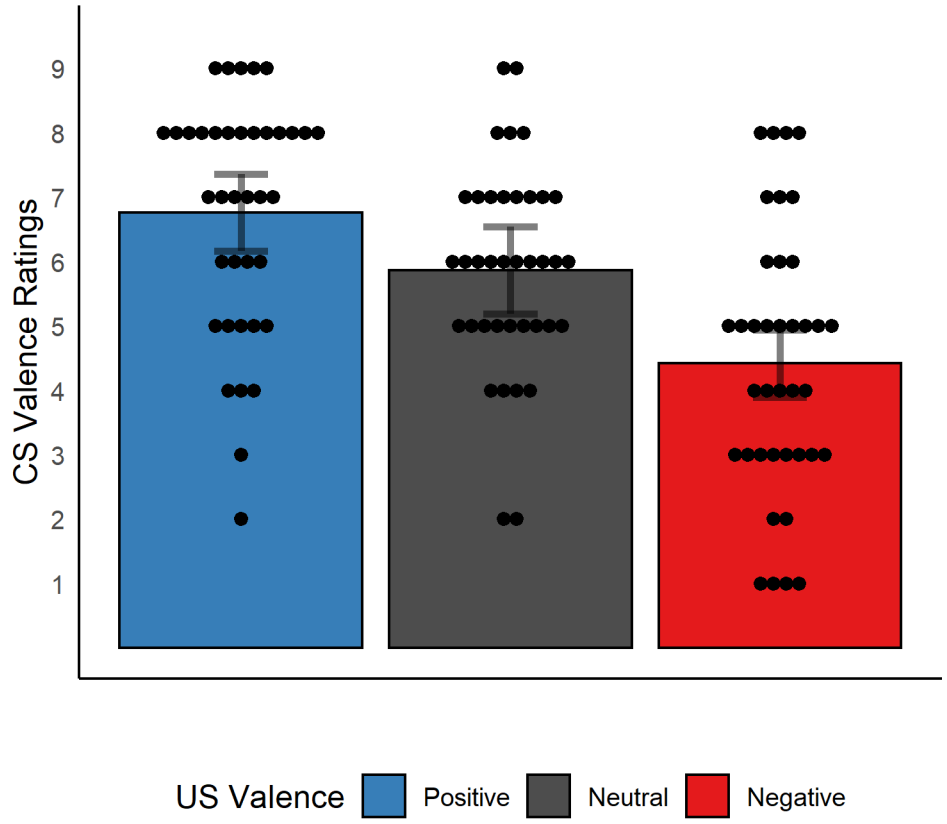

Figure 9. Mean explicit valence ratings with individual participant values plotted for backward CSs as a function of US valence (positive vs. neutral vs. negative). Error bars represent 95% confidence intervals of the mean.

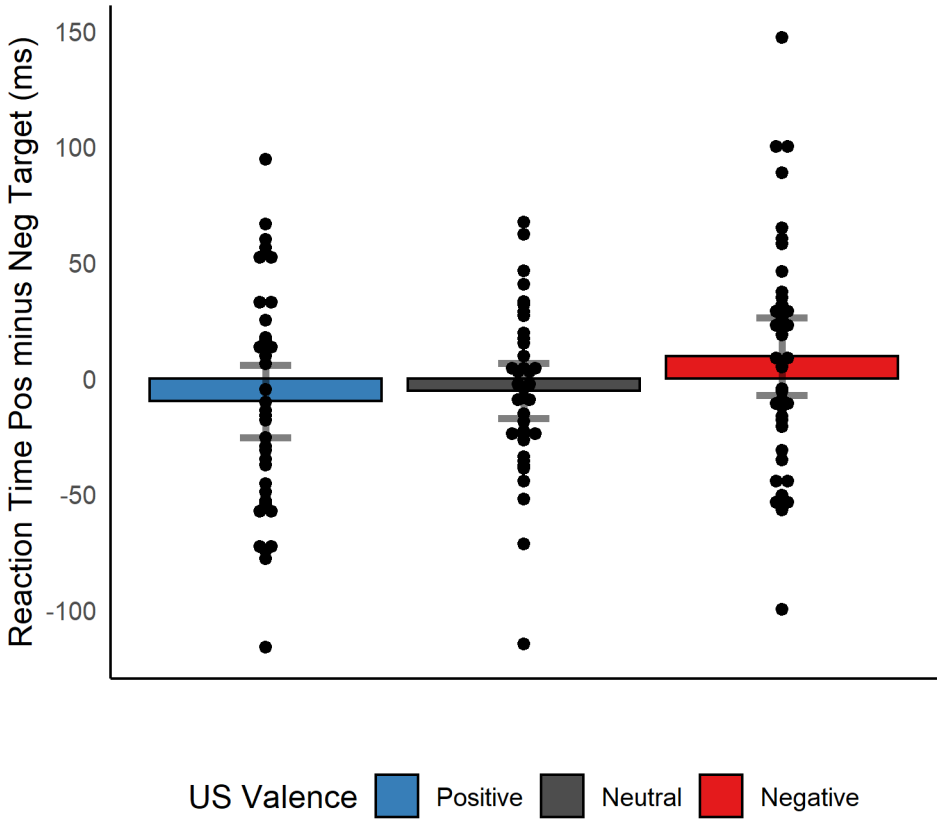

Figure 10. Difference scores (positive target words – negative target words) for reaction times with individual participant values plotted from the affective priming task for backward CSs as a function of US valence (positive vs. neutral vs. negative). Assimilation effects are represented by negative scores for positive CS primes and by positive scores for negative CS primes. Error bars show 95% confidence intervals of the mean.

Supplementary Material – Experiment 1

Startle blink magnitude – Extinction

Figure S1 shows habituation occurring across blocks regardless of group or conditioning type. A main effect of block, F (2, 58) = 58.674, p < .001, ηp^2 = .669, was qualified by a marginal US valence x block interaction, F (2, 58) = 3.083, p = .053, ηp^2 = .096. Follow-up analyses revealed larger responses to CSs paired with positive USs at block 1, F (1, 59) = 4.176, p = .045, ηp^2 = .066, and no differences between CSs at blocks 2, F (1, 59) = 0.057, p = .812, ηp^2 = .001, and 3, F (1, 59) = 1.403, p = .241, ηp^2 = .023. As the interaction was marginal, we also followed up the main of effect of block, and found larger responses at block 1 when compared with block 2, t (59) = 10.65, p < .001, d = 1.36, and block 3, t (59) = 9.06, p < .001 d = 1.16.

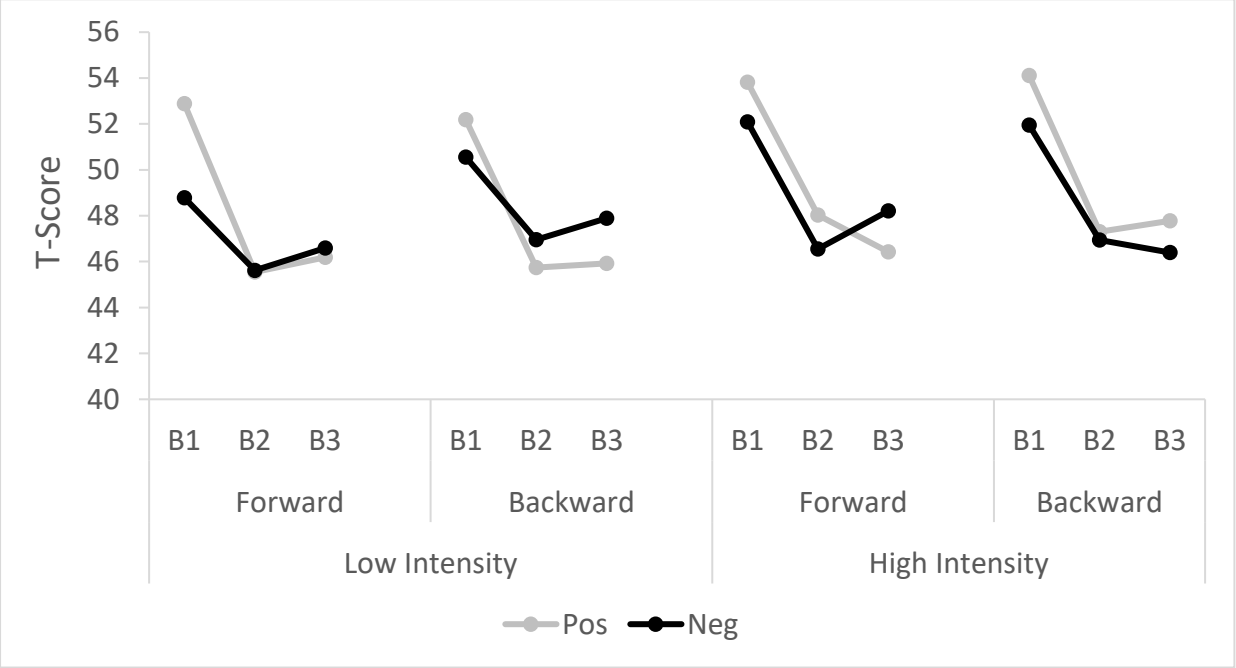

Figure S1. Startle blink magnitude (*T*-scores) by block (1, 2, and 3) during extinction, for CSs that were paired with positive and negative USs for forward and backward conditioning in the low intensity and high intensity groups.

Startle blink latency – Acquisition

Figure S2 below suggests faster blink onset following probes presented during CSs paired with negative USs than CSs paired with positive USs for forward conditioning, and faster following probes presented during CSs paired with positive USs than CSs paired with negative USs for backward conditioning, in both groups. Main effects of conditioning type, F (1, 56) = 190.209, p < .001, ηp^2 = .773, and block, F (2, 55) = 21.857, p < .001, ηp^2 = .443, a group x conditioning type interaction, F (1, 56) = 29.188, p < .001, ηp^2 = .343, and a conditioning type x US valence interaction, F (1, 56) = 35.373, p < .001, ηp^2 = .387, were qualified by a conditioning type x US valence x block interaction, F (2, 55) = 4.446, p = .016, ηp^2 = .139. Follow-up analyses revealed an assimilation effect for forward conditioning, as blink onset was faster during CSs paired with negative USs than CSs paired with positive USs, on blocks 1, F (1, 56) = 13.14, p = .001, ηp^2 = .190, and 2, F (1, 56) = 6.559, p = .013, ηp^2 = .105, and a contrast effect for backward conditioning, as blink onset was faster during CSs paired with positive USs than CSs paired with negative USs, on blocks 1, F (1, 56) = 18.417, p < .001, ηp^2 = .247, and 3, F (1, 56) = 4.465, p = .039, ηp^2 = .074.

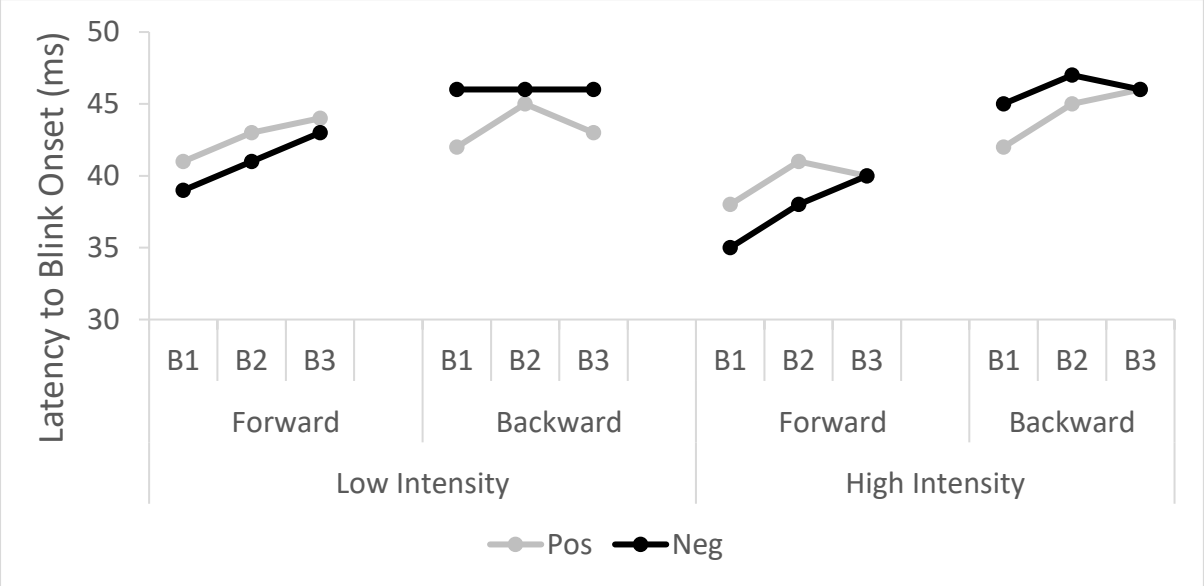

Figure S2. Time until blink onset following startle probe in milliseconds, during forward and backward CSs paired with positive and negative USs, in the low and high intensity groups across blocks 1, 2, and 3 during acquisition.

Startle blink latency – Extinction

As shown in figure S3, blink latency slows gradually across blocks in the low intensity group, and slows between blocks 1 and 2, then increases between blocks 2 and 3 in the high intensity group. A main effect of block, F (2, 48) = 14.327, $p < .001$, $\eta p^2 = .374$, was qualified by a group x block interaction, F (2, 48) = 5.514, $p = .007$, $\eta p^2 = .187$. Follow-up analyses revealed no differences between blocks in the low intensity group, F (2, 48) = 2.597, $p = .085$, $\eta p^2 = .098$, and faster responses for block 1 and 3 than block 2, t (49) = 5.62, $p < .001$, $d = 0.79$, and t (49) = 4.15, $p < .001$, $d = 0.58$, respectively, F (2, 48) = 18.836, $p < .001$, $\eta p^2 = .440$, in the high intensity group.

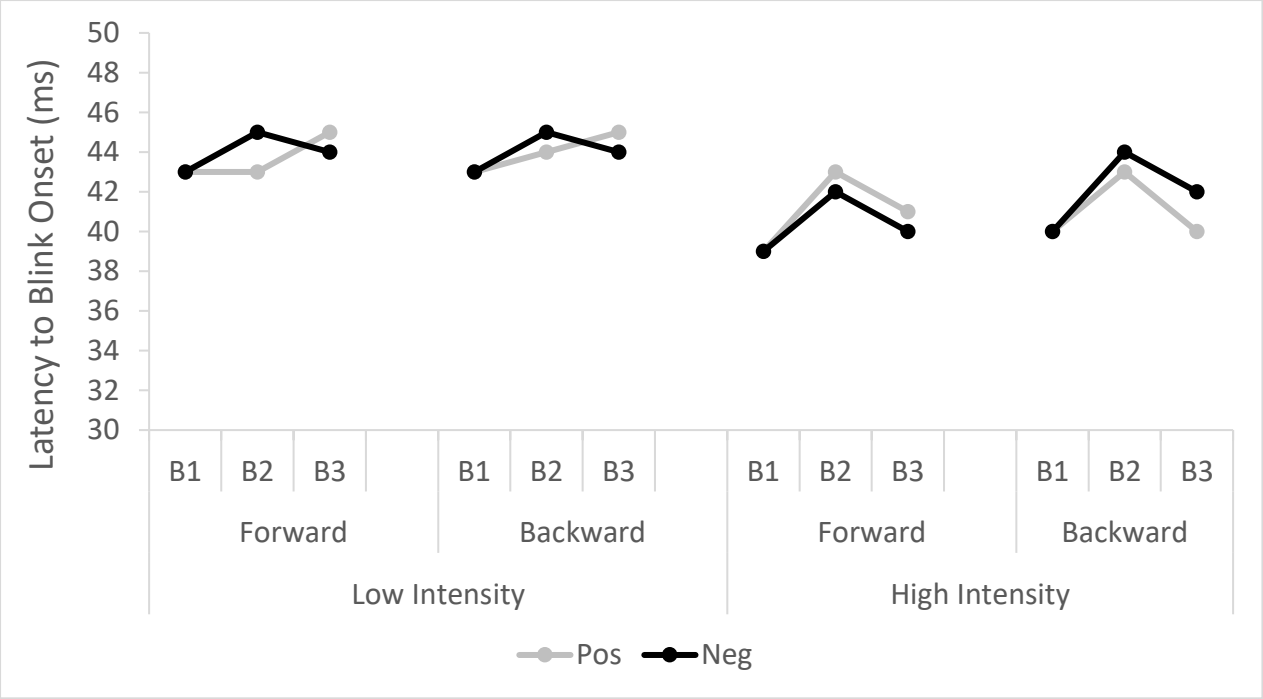

Figure S3. Time until blink onset following startle probe in milliseconds, during forward and backward CSs paired with positive and negative USs, in the low and high intensity groups across blocks 1, 2, and 3 during extinction.

Affective priming – Errors

Figures S4 below suggests assimilation effects for forward conditioning in both groups, and for backward conditioning in the low intensity group. In the high intensity group for backward conditioning, a contrast effect is suggested. The group x conditioning type

interaction was significant, F (1, 59) = 9.257, p = .003, ηp^2 = .136. Follow-up analyses showed that priming scores in the low intensity group were significantly greater than 0 for backward conditioning, t (29) = 3.181, p = .003, d = 0.58, but not forward conditioning, t (29) = 1.511, p = .142, d = 0.28, and that forward and backward conditioning did not differ significantly from each other, t (28) = 0.876, p = .384, d = 0.16. In the high intensity group, follow-up analyses showed that forward conditioning was greater than 0, t (30) = 3.437, p = .002, d = 0.62, while backward conditioning was not, t (30) = 1.070, p = .293, d = 0.19, and that forward and backward conditioning differed significantly from each other, t (29) = 3.448, p = .001 d = 0.62.

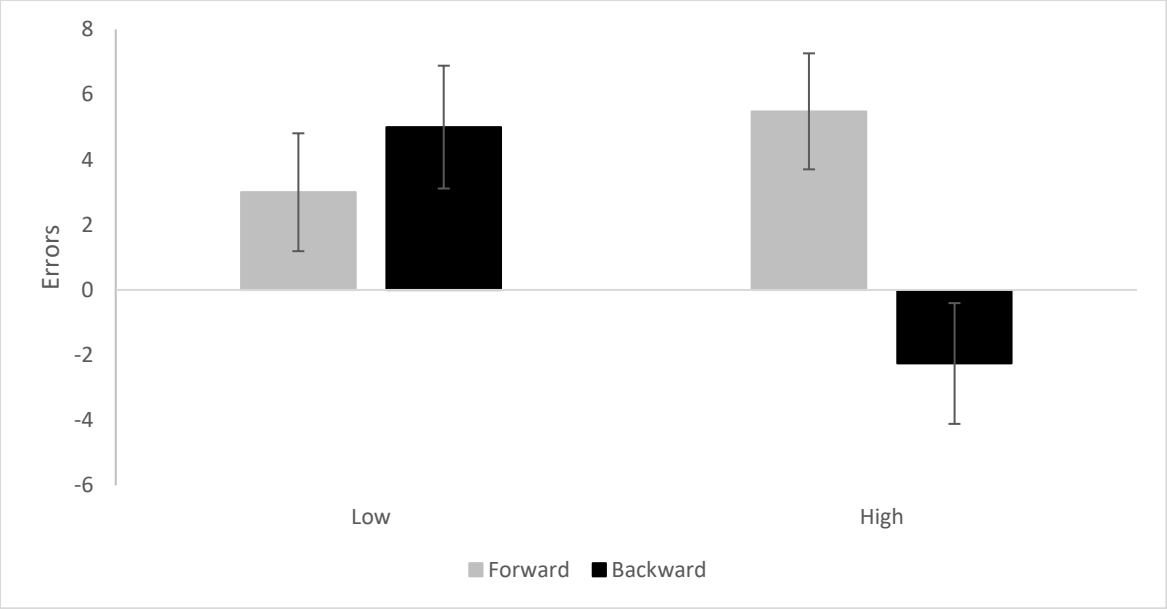

Figure S4. Priming scores for errors from the affective priming task for forward and backward conditioning in the low and high intensity groups. Error bars show 95% confidence intervals for the mean.

Skin conductance responding

Self-adhesive isotonic electrodes were attached to the thenar and hypothenar eminences of the non-preferred hand to record SCRs throughout the experiment. Responses were amplified at a gain of 5 μSiemens per volt by a Biopac MP150 system and recorded using AcqKnowledge 4.1.0 at a sampling rate of 1000 Hz. Respiration was measured by

fitting a respiration belt around the participant's waist to control for SCR artefacts. SCR's were scored offline using AcqKnowledge 4.1.0. Responses were square root transformed and range correct by dividing each participant's response by their largest response, to reduce the skew of the data prior to analysis. Only first interval responses (1-4s) during forward CSs and US onset were analysed, as startle probes and CS/US overlap precluded any meaningful analysis of second interval responding or responding during backward CSs. Responses were then aggregated into five blocks, with each block containing the average of two consecutive trials. These data during forward CSs and US onset were then subjected to separate 2 (group: low intensity vs high intensity) x 2 (US valence: positive vs negative) x 5 (block: 1, 2, 3, 4, 5) mixed model ANOVA's. Four participants from the low intensity group and five participants from the higher intensity group were removed for being non-responders.

Skin conductance responding – First interval – Forward CS

Figure S5 shows habituation across blocks, and larger responses to CSs paired with negative USs regardless of group. This was confirmed by a main effect of US valence, F (1, 53) = 5.658 p = .021, ηp^2 = .096, showing larger responses to CSs paired with negative USs over CSs paired with positive USs, and a main effect of block, F (4, 50) = 23.861, p < .001, ηp^2 = .656, which shows larger responses at block 1 than blocks 2, t (54) = 9.84, p < .001, d = 1.33, 3, t (54) = 8.11, p < .001, d = 1.09, 4, t (54) = 6.94, p < .001, d = 0.94, and 5, t (54) = 6.93, p < .001 d = 0.93. While it may look like group interacts with valence, this was not the case, Group × Valence, F (1, 53) = 2.197 p = .144, ηp^2 = .040, and Group × Valence × Block, F (4, 50) = 0.381 p = .821, ηp^2 = .030.

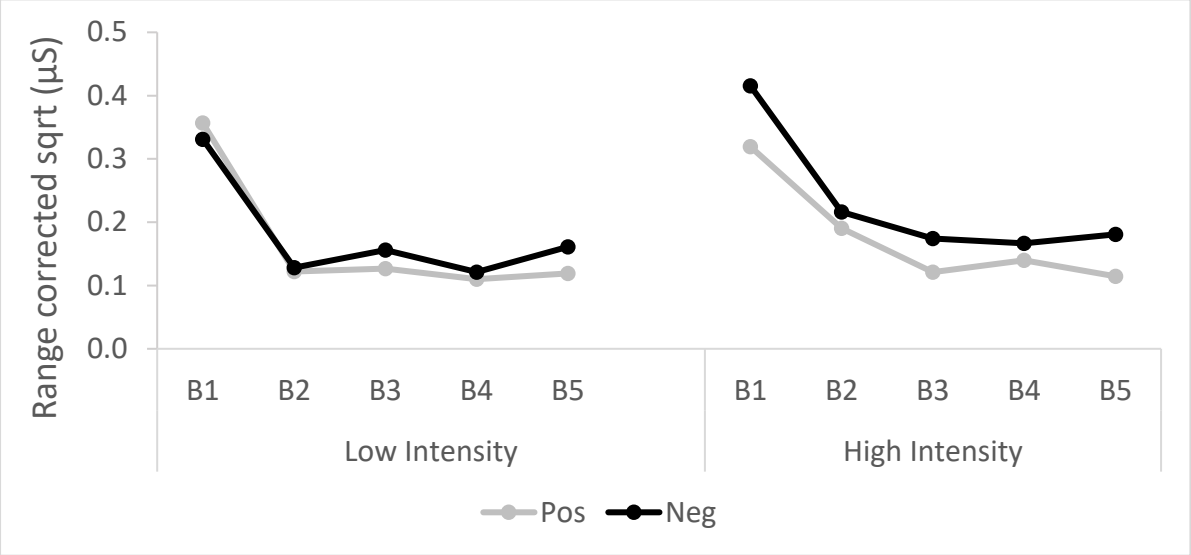

Figure S5. First interval skin conductance responses to CSs paired with positive and negative USs during acquisition, presented in blocks of 2 averaged responses per block (blocks 1, 2, 3, 4, and 5), for the low intensity and high intensity groups.

Skin conductance responding – First interval – US onset

Figure S6 shows habituation to the US, with larger responses to the negative US in the high intensity group. This was confirmed by a main effect of block, F (4, 50) = 14.267, p < .001, ηp^2 = .533, and a group x US valence interaction, F (1, 53) = 5.111, p = .028, ηp^2 = .088. Follow-up analyses revealed no differences between positive and negative USs in the low intensity group, F (1, 53) = 0.151 p = .669, ηp^2 = .003, and larger responses to the negative US than the positive US in the high intensity group, F (1, 53) = 7.764 p = .007, ηp^2 = .128.

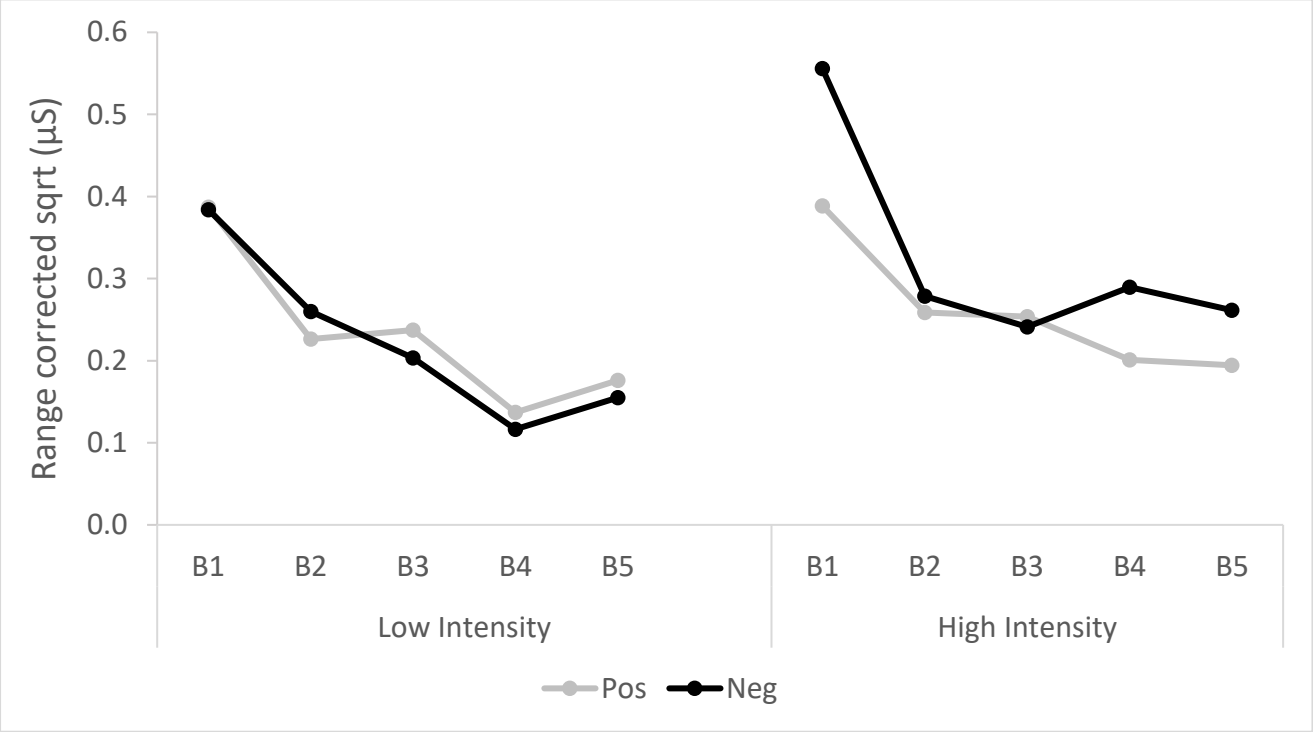

Figure S6. First interval skin conductance responses positive and negative USs during acquisition, presented in blocks of 2 averaged responses per block (blocks 1, 2, 3, 4, and 5), for the low intensity and high intensity groups.

Supplementary Material – Experiment 2

Startle blink magnitude – Extinction

Figure S7 shows a decrease in response from blocks 1 to 2. This is confirmed by the tests of within-subjects contrasts which showed a linear trend for block, F (1, 36) = 13.91, p = .001, ηp^2= .279.

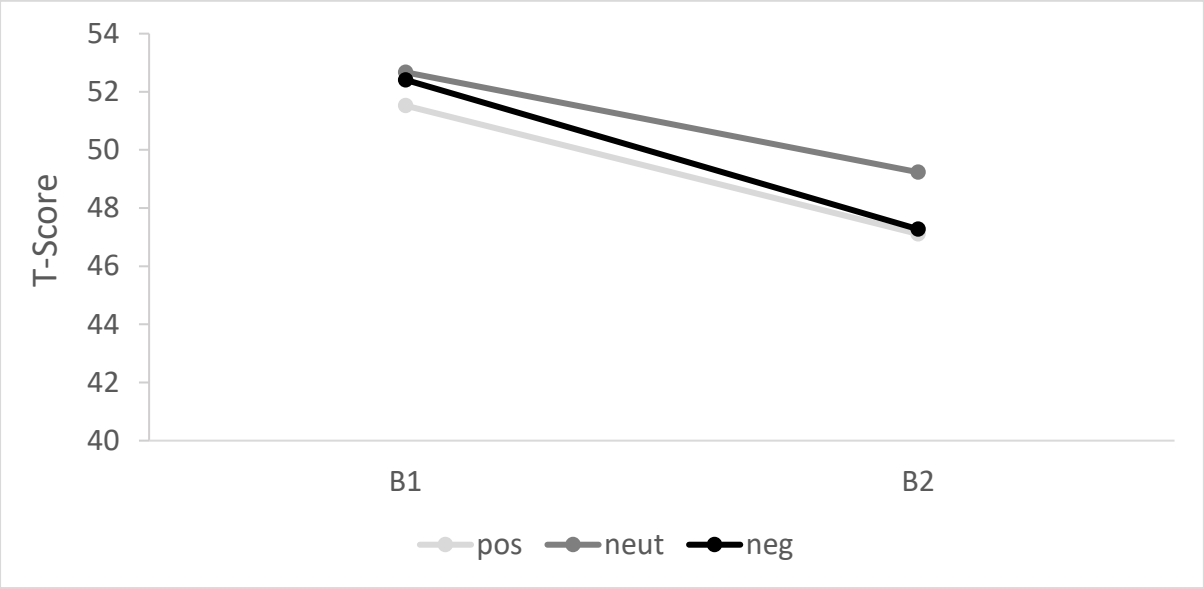

Figure S7. Startle blink magnitude (*T*-scores) by block (1 and 2) during extinction, for CSs that were paired with positive, neutral, and negative USs following backward conditioning.

Startle blink latency – Acquisition

Figure S8 shows faster blink onset during CSs paired with the positive US than CSs paired with the neutral US, and CSs paired with the negative US. This was confirmed by the tests of within-subject contrasts which showed a linear trend for US valence, F (1, 33) = 19.801, $p < .001$, ηp^2= .375

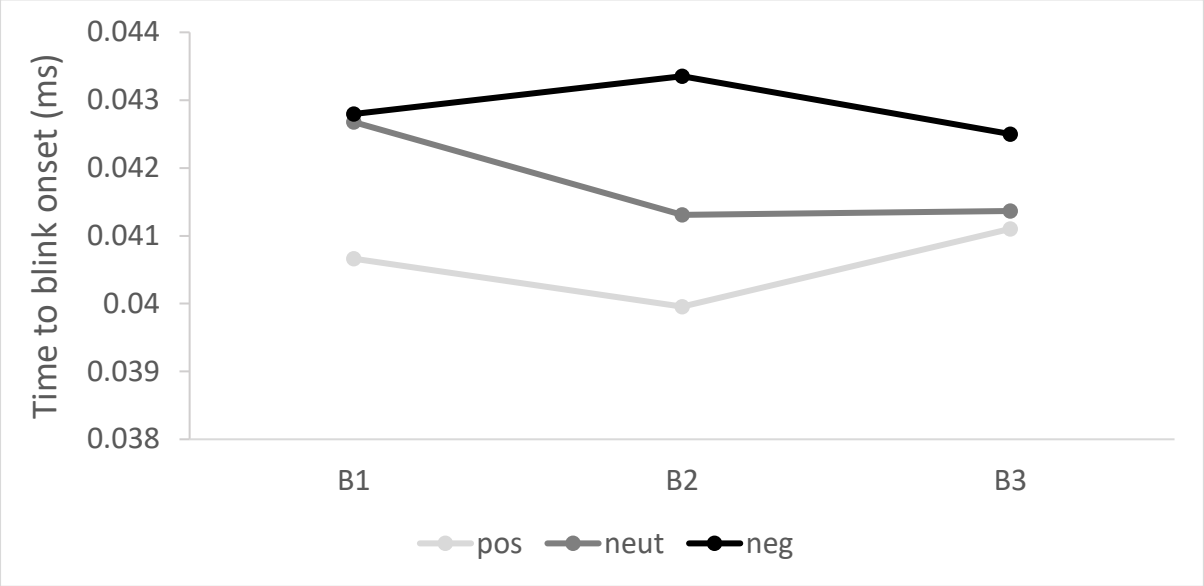

Figure S8. Time until blink onset following startle probe in milliseconds, during backward CSs paired with positive, neutral, and negative USs, across blocks 1, 2, and 3 during acquisition.

Startle blink latency – Extinction

Figure S9 suggests slower blinks to the CS paired with the negative US compared with CSs paired with neutral and positive USs at block 1, with faster responses to CSs paired with the negative US at block 2. However, no tests of within-subjects contrasts were significant, Fs < 0.738, ps > .396, $\eta p^2 s$ < .021.

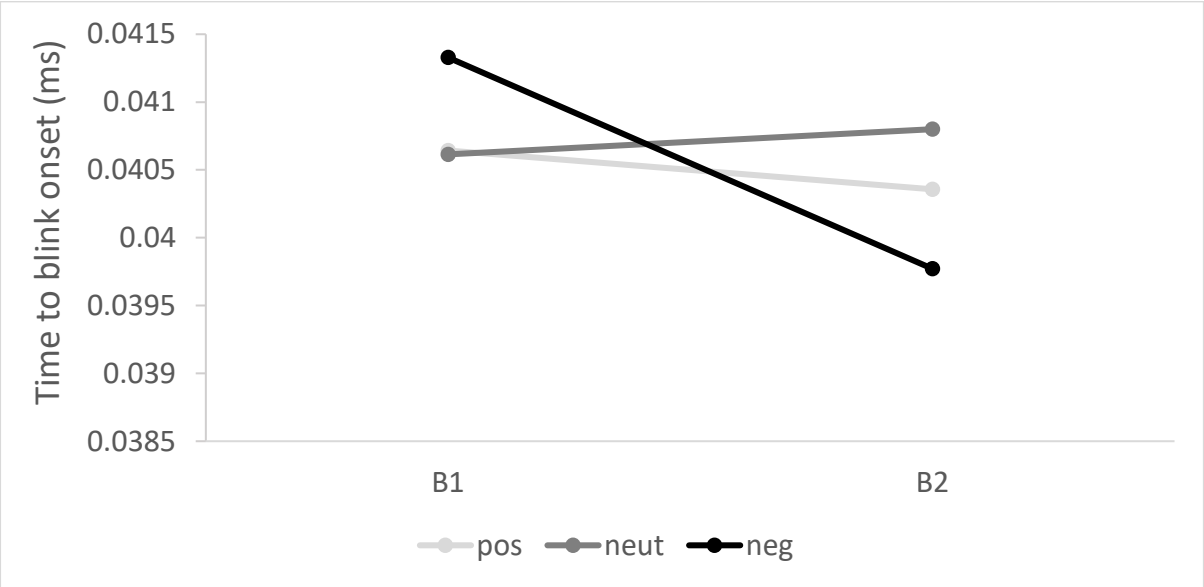

Figure S9. Time until blink onset following startle probe in milliseconds, during backward CSs paired with positive, neutral, and negative USs, across blocks 1 and 2 during extinction.

Affective priming – Errors

Figure S10 suggests a quadratic trend, however neither the quadratic, F (1, 37) = 1.488, p = .230, ηp^2= .039, or linear, F (1, 37) = 1.716, p = .198, ηp^2= .044, trend analyses were significant.

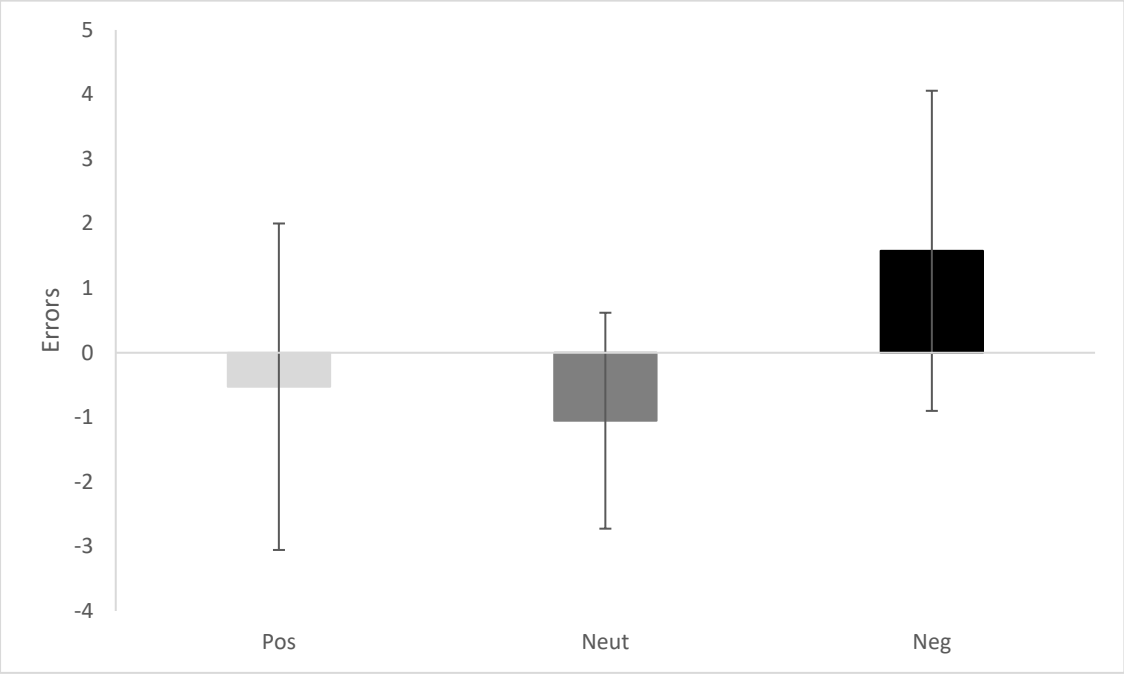

Figure S10. Difference scores (positive target words – negative target words) for errors from the affective priming task for CSs paired with positive, neutral, and negative USs. Error bars show 95% confidence intervals of the mean.

Skin Conductance Responding – US Onset – First Interval Response

Figure S11 shows skin conductance responses decreasing across blocks, with larger skin conductance responses to the negative US than the neutral and positive USs. This was confirmed by main effects of US valence, F (2, 35) = 15.827, p < .001, ηp^2 = .475, and block, F (3, 34) = 13.188, p < .001, ηp^2 = .538, and a significant US valence x block interaction, F (6, 31) = 4.86, p < .001, ηp^2 = .485. At block's 1 and 2, responses to USneg were greater than

responses to USpos and USneut, F (2, 35) = 23.568, p < .001, ηp^2= .574, and F (2, 35) = 6.774, p = .003, ηp^2= .279. No differences in responses size were found between US valence at block 3, F (2, 35) = 1.564, p = .224, ηp^2= .082, and block 4, F (2, 35) = 0.947, p = .398, ηp^2= .051.

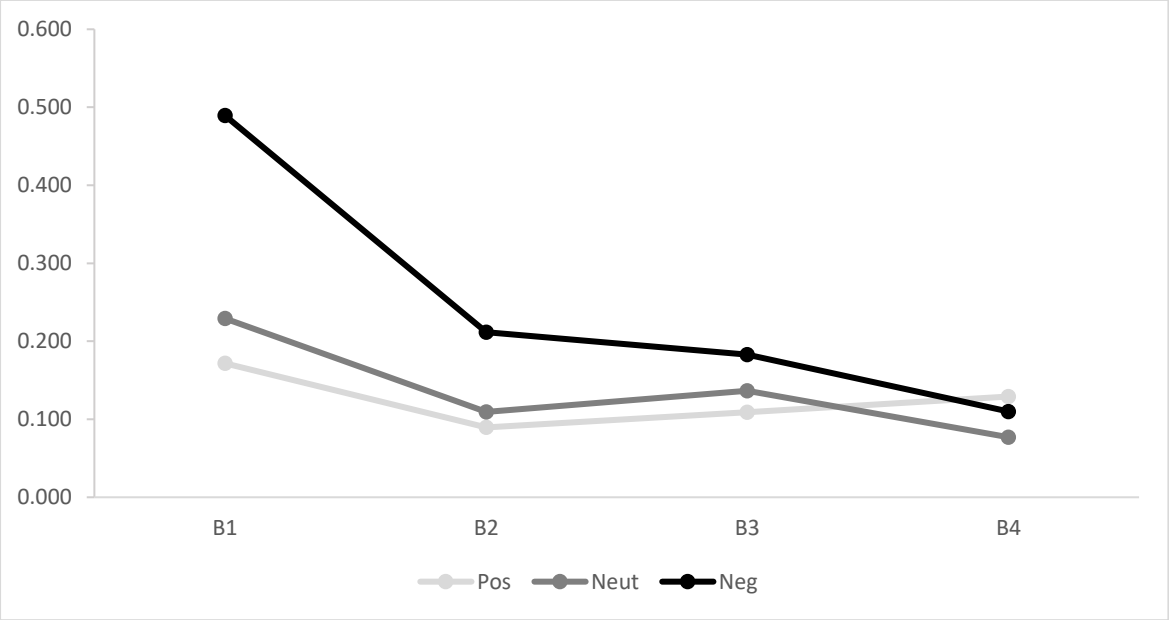

Figure S11. First interval skin conductance responses to positive, neutral, and negative USs during acquisition, presented in blocks of 2 averaged responses per block (blocks 1, 2, 3, and 4).