A Potential Pathway to the Relapse of Fear? Conditioned Negative Stimulus Evaluation (but not Physiological Responses) Resists Instructed Extinction

Camilla C. Luck[1, 2] & Ottmar V. Lipp[1, 2]

[1]School of Psychology and Speech Pathology, Curtin University, Australia

[2]ARC-SRI: Science of Learning Research Centre

Author Notes

Address for correspondence:

Camilla C. Luck, School of Psychology and Speech Pathology, Curtin University, GPO Box U1987 Perth WA 6845, Australia. Email: camilla.luck@postgrad.curtin.edu.au

Ottmar V. Lipp, School of Psychology and Speech Pathology, Curtin University, GPO Box U1987 Perth WA 6845, Australia. Email: ottmar.lipp@curtin.edu.au

Acknowledgements

This work was supported by grant numbers DP120100750 and SR120300015 from the Australian Research Council.

Word count: 12436

Abstract

Relapse of fear after successful intervention is a major problem in clinical practice. However, little is known about how it is mediated. The current study investigated the effects of instructed extinction and removal of the shock electrode on electrodermal responding (Experiment 1), fear potentiated startle (Experiment 2), and a continuous self-report measure of conditional stimulus valence (Experiments 1 and 2) in human differential fear conditioning. Instructed extinction and removal of the shock electrode resulted in the immediate reduction of differential physiological responding, but did not affect self-reported conditional stimulus valence. A separate sample of participants (Experiment 3) who were provided with a detailed description of the experimental scenario predicted the inverse outcome, reduced differential stimulus evaluations and continued differential physiological responding, rendering it unlikely that the current results reflect on demand characteristics. These results suggest that the negative valence acquired during fear conditioning is less sensitive to cognitive interventions than are physiological indices of human fear learning and that valence reduction requires extended explicit exposure training. Persisting negative valence after cognitive intervention may contribute to fear relapse after successful treatment.

Key words: fear conditioning, instructed extinction, electrodermal responses, fear potentiated startle, evaluative learning, fear relapse.

Epidemiological data suggests that 25 percent of the population will develop an anxiety disorder at some stage in life (Kessler, Koretz, Merikangas & Wang, 2004). It is thus reassuring that efficacious treatments are available for these conditions with exposure based and cognitive therapies emerging as the most commonly used interventions in clinical practice (Ougin, 2011), both receiving consistent empirical support for a number of anxiety disorders (Bisson, 2007; Ougin, 2011; Sánchez-Meca, Rosa-Alcàzar, Marín-Martínez & Gómez-Conesa, 2010). In spite of this considerable success, approximately one to two thirds of successfully treated patients will relapse within eight years (Craske, 1999). This figure highlights the need for continued research into the mechanisms underlying fear acquisition, reduction, and relapse -- an understanding which is essential for the development of treatments with improved long term outcomes.

Fear is a basic emotion characterized by high levels of negative affect (displeasure) and physiological arousal (Lang, 1995). Classical fear conditioning models can provide a conceptual framework to study the development and treatment of human fear (Craske, Hermans & Vansteenwegen, 2006). In the laboratory setting, a differential fear conditioning paradigm is often used, involving the presentation of two neutral conditional stimuli and an aversive unconditional stimulus (Lipp, 2006). During acquisition training, one conditional stimulus (CS+) is paired with the aversive unconditional stimulus (e.g. electrotactile stimulus), whilst the other is presented alone (CS-; Lipp, 2006). During fear acquisition, the CS+ becomes a valid predictor of the aversive unconditional stimulus, leading to the development of increased physiological responding and decreased valence ratings to CS+ in comparison with the CS- (Lipp, 2006; De Houwer, Thomas, & Baeyens, 2001). Extinction training involves the repeated presentation of the CS+ without the unconditional stimulus and has been suggested as an experimental analogue to exposure based interventions (Kerkhof et al., 2012). Extinction training is very effective in eliminating differential physiological responding to CS+ and CS- and also reduces the negative valence acquired by the CS+, however, there is evidence that negative valence is more resistant to extinction than are the physiological indices of fear learning and thus requires extended extinction training (Lipp, Oughton & LeLievre, 2003).

A very common finding in human fear learning is that after successful extinction of differential responding, conditioned responding can reoccur in a post-extinction test session, in the absence of any re-training or re-exposure to the feared stimulus (for a review see Vervliet, Craske & Hermans, 2013). This phenomenon is referred to as the return of fear (Rachman, 1966). To date, three mechanisms mediating the return of fear have been uncovered; spontaneous recovery: the return of fear following the mere passage of time, renewal: the return of fear following a context change, and reinstatement: the return of fear following unpredicted presentations of the unconditioned stimulus (Bouton, 2002). It should be noted that as defined above (Lang, 1995), return of fear implies the recurrence of both physiological arousal and negative affect. However, under a less strict definition, the return of negative valence or physiological arousal alone could be interpreted as being a partial return of the fear response – an occurrence which could predispose the individual for full return of fear.

After observing that persisting negative valence towards the feared stimulus was correlated with higher reinstatement rates, Hermans et al., (2005) suggested that lingering negative valence could provide an additional pathway for the return of fear. Noting that negative stimuli preferentially associate with aversive outcomes (Hamm, Vaitl, & Lang, 1989) and that negative valence has been associated with escape and avoidance tendencies (Chen & Bargh, 1999), Kerkhof et al. (2012) developed this theory proposing, based on Lang's (1995) conceptualization of fear as a combination of high arousal and negative valence, that if negative valence persists after extinction, fear could return if the individual is put in a high arousal situation or state.

The human fear conditioning paradigm can also be used to examine the influence of cognition on the extinction of fear learning. Following, Mower's (1938) initial observation that electrodermal responding could be 'switched on and off' with signals informing the participants when an aversive unconditional stimulus was to be expected, researchers have used the instructed extinction paradigm as an experimental analogue for cognitive interventions to reduce fear. Instructed extinction involves informing one group of participants at the end of acquisition training that the aversive unconditional stimulus will no longer be presented, whilst a control group receives the same level of interaction with the experimenter, but is not informed. Frequently, the instruction that no further unconditional stimuli will be presented is accompanied by removal of the unconditional stimulus electrode (Hugdahl & Öhman, 1977; Hugdahl, 1978; see Sevenster, Beckers, & Kindt, 2012, for mere instruction effects). This manipulation has been shown to reduce the differential electrodermal responding acquired during fear conditioning unless the conditional stimuli used are pictures of snakes or spiders as fear conditioned to these stimuli seems to be encapsulated from cognition (for a recent review see Mallan, Lipp, & Cochrane, 2013). It is currently unclear, however, whether instructed extinction also affects fear learning as indexed by fear potentiated startle or the negative valence acquired during fear conditioning.

The two studies that to date have assessed the effect of instructed extinction on conditioned fear as indexed by fear potentiated startle have reached different conclusions. Whereas Mallan, Sax, and Lipp (2009) reported that, like differential electrodermal responses, instructed extinction abolished differential fear potentiated startle, Sevenster et al. (2012)

reported a dissociation between electrodermal responding and fear potentiated startle. In this study, instruction effects on differential electrodermal responses were immediate, i.e., evident on the very first trial of extinction training, whereas differential startle potentiation persisted for the first two trials of extinction. It should be noted, however, that relative to the non-instructed control group, extinction of fear as indexed by fear potentiated startle was accelerated considerably, as differential fear potentiated startle was absent after the first two extinction trials in the instructed group, but persisted across the first ten extinction trials in controls. Based on the latter finding it seems reasonable to conclude that conditioned fear as indexed by both physiological indices is subject to instructed extinction.

Whether instructed extinction affects the negative valence acquired by a CS+ during acquisition training is less clear. Lipp and Edwards (2002) and Rowles, Lipp and Mallan (2012) included post-extinction assessments of conditional stimulus valence which seemed to be unaffected by instruction. Equivalent differential evaluation of CS+ and CS- was evident in all groups regardless of the nature of the conditional stimuli used or the instructions provided. However, as conditional stimulus valence was assessed after the completion of extinction training, it is not clear whether the differential conditional stimulus evaluations reflect on insensitivity to instruction or the renewal of fear due to a context change (Bouton, 2002; Vansteenwegen, Dirikx, Hermans, Verwliet, & Eelen, 2006). Lipp et al. (2003; Experiment 2) did not find an effect of instructed extinction on conditional stimulus valence using a continuous assessment during extinction training, however, these results need to be considered with care due to fast extinction in the control group and no instruction effect on electrodermal responses.

The effect of instructed extinction on acquired conditional stimulus valence has also been examined in studies of evaluative conditioning which can inform studies of fear learning. In evaluative conditioning, pleasant and unpleasant pictures rather than aversive electrotactile stimuli are used as unconditional stimuli and conditional stimulus valence can be assessed immediately after instruction and during extinction training. Using such a paradigm, Lipp, Mallan, Libera and Tan (2010) failed to find an effect of instructed extinction on measures of conditional stimulus valence, immediate or delayed, although participants reported reduced expectancy of the unconditional stimuli immediately after instruction. Gast and De Houwer (2013) found valence measures to be sensitive to instructed extinction in their first, but not in their second experiment. However, the instructed extinction effect in Experiment 1 was not significant for participants who could correctly report the stimulus contingencies used during evaluative conditioning training. Taken together, results from evaluative conditioning seem to suggest no effect of instructed extinction on conditional stimulus valence, at least in participants who show evidence of learning during the initial training. It is unclear, however, whether these findings would transfer to fear conditioning that is acquired using biologically significant aversive unconditional stimuli, such as an electrotactile stimulus. Such an unconditional stimulus is likely to convey significantly higher levels of negative valence and emotional arousal than the presentation of an unpleasant picture.

To assess the effects of instructed extinction on electrodermal responses, fear potentiated startle, and conditional stimulus valence, two differential fear conditioning experiments were conducted using neutral faces as conditional stimuli and an aversive electric stimulus as the unconditional stimulus. In Experiment 1, electrodermal responding and conditional stimulus valence were assessed whereas fear potentiated startle and conditional stimulus valence were assessed in Experiment 2. We examined electrodermal responding and fear potentiated startle in separate experiments to avoid contamination of electrodermal responses by the noise probes used to elicit startle responses and to replicate the results for conditional stimulus valence. Following the procedure used in the majority of prior instructed extinction studies, we removed the shock electrode as part of the instructional manipulation to ensure that the participants believed the instructions. Based on the review of the prior literature we predict that instructed extinction will reduce electrodermal responses and fear potentiated startle, whereas negative valence acquired by the CS+ will persist.

Experiment 1

Method

Participants. Thirty-six (21 female) undergraduate students aged 17-52 years (*M* = 21.71) provided informed consent and volunteered participation in exchange for course credit. Participants were randomly assigned to one of two groups (instruction/removal, control). The pre-experimental ratings data of one participant was lost due to a recording error and evaluation data of three participants and the electrodermal responses of one participant were lost due problems with the recording device. These participants were included in the analyses of all remaining measures.

Apparatus/Stimuli. Color pictures of four Caucasian, male adults [NimStim database: images M_NE_C: models 20, 21, 32, 31, Tottenham et al. (2009)] displaying neutral facial expressions were used as conditioned stimuli. The pictures were 506 by 650 pixels in size, and were displayed for six seconds on a 17 inch color LCD screen. The two faces used as conditional stimuli during the experiment, the faces used as CS+/CS-, and whether the first trial of each phase was a CS+/CS- was counterbalanced across participants.

Conditional stimulus evaluations and physiological responding were recorded with a Biopac MP150 system, at a sampling frequency of 1000 Hz, using AcqKnowledge Version 3.9.1.

Electrodermal responding was DC amplified at a gain of 5 μSiemens per volt and monitored using two 8mm Ag/AgCl electrodes filled with Mansfield R & D TD-246 electrode paste and attached using adhesive collars. Respiration was monitored with a chest gauge to control for respiration induced artefacts in electrodermal responding. Conditional stimulus valence was measured on a trial-by-trial basis using an evaluation joystick with the anchors, very unpleasant, neutral, and very pleasant. A Grass SD9 stimulator, pulsed at 50 Hz, was used to deliver a 200 ms electrotactile stimulus to the participants' preferred forearm via a concentric electrode. DMDX3.0.2.8 software (Forster & Forster, 2003) was used to record pleasantness ratings before and after conditioning training and to control stimulus presentation and timing.

Procedure. Participants were seated in front of the monitor in an experimental room, located adjacent to a control room. A respiration belt was fitted around their waist and two electrodes were placed on the thenar and hypothenar prominences of their non-preferred hand. A shock electrode was attached with a bandage to their preferred forearm, and the participant completed a shock-work-up procedure to set the electrotactile stimulus to an intensity that was experienced as 'unpleasant, but not painful'. After the shock work-up procedure, the participants were subjected to a three minute baseline recording of their physiological responding whilst they relaxed and watched the blank computer screen.

After the baseline recording, the participants completed a pre-experimental rating task, in which participants were prompted to rate the faces on a pleasantness scale ranging from 1 to 9 (1 = unpleasant, 9 = pleasant). The CS faces were displayed on the screen until a response was made. After the pre-ratings task, participants were informed that they would view pictures of faces, and that they were required to use the joy-stick to indicate whether they felt the face was pleasant, unpleasant, or neutral. To ensure that the valence ratings were not contaminated by the presence/absence of the unconditional stimulus, the participants were informed that they should rate each face as soon as it was presented on the screen. Valence ratings were made with the participants' preferred hand to ensure the movement of the joystick would not interfere with the skin conductance recording. After the task instructions the conditioning procedure was started. Conditioning consisted of habituation, acquisition, and extinction phases; during habituation, two faces, the CS+ and the CS-, were presented four times. The habituation phase allows for the habituation of orienting responses to the conditional stimuli. The acquisition phase followed habituation immediately. During acquisition, the CS+ was presented eight times and unconditional stimulus onset coincided with CS+ offset in a 100% reinforcement schedule, whereas the CS- was presented eight times alone. Extinction involved the presentation of both the CS+ and the CS- eight times each, but no electrotactile stimulus was presented during this phase. All conditional stimuli were presented for six seconds, and a blank rest screen was presented between trials for either 15, 18, or 21 seconds, randomly. Inter-trial interval duration was varied to avoid the participants predicting and anticipating the onset of the next CS.

For both the instruction/removal group and the control group, the experimenter entered the room at the end of acquisition. Participants in the instruction/removal group were informed that in the second part of the experiment the presentations of the electrotactile stimulus would cease and the shock electrode was removed. Participants in the control group were informed that the shock electrode needed to be checked and the shock electrode was removed and reattached. All participants were informed that the experiment would continue and that they should continue to evaluate the faces. After the last trial of extinction, the participants completed a post-experimental rating procedure that was identical to the pre-rating procedure. After this, the electrodes were removed and the participants were led into the control room to complete a post-experimental questionnaire. This required an assessment of contingency awareness in which the participants were shown four neutral faces and asked to indicate which two they had seen during the experiment and which face had been followed by the electrotactile stimulus. Participants were asked to rate how pleasant or unpleasant they found the electrotactile stimulus (-3 to +3 scale) and as a manipulation check the participants were asked to indicate whether they had believed the instruction that the electrotacticle stimulus would not occur following the interruption (yes or no question; instruction/removal group only).

Scoring and Response Definition. To provide a measure of spontaneous electrodermal responding, any discernible response during the three minute baseline was counted (Dawson, Schell & Filion, 2007). Respiration traces were inspected to identify cases when electrodermal responding might have been contaminated by deep breaths or excessive movement. No cases of excessive movement were identified and therefore no electrodermal responses were discarded. Electrodermal responses during conditioning were scored in three latency windows in accordance with Prokasy and Kumpfer's (1973) recommendations for scoring electrodermal responding in fear conditioning experiments. The First Interval Responding (FIR) was defined as responses starting within 1-4 seconds of the CS onset and Second Interval Responding (SIR) was defined as responses starting within 4-7 seconds of the CS onset. Responses to the unconditional stimulus were scored during acquisition, as responses starting within 7-10 seconds of the CS+ onset (Prokasy & Kumpfer, 1973). The use of multiple response windows (as opposed to single response) is recommended by Prokasy and Kumpfer (1973) as there is evidence that first interval responding reflects orientation to the conditional stimulus, second interval responding reflects the anticipation of the unconditional stimulus, and the unconditional response window reflects the response to the unconditional stimulus (Stewart, Winokur, Stern, Guze, Pfeiffer & Hornung, 1959; Lockart, 1966.). Moreover, there is evidence that different experimental manipulations will differentially affect first and second interval responding (Prokasy & Ebel, 1967), that the occurrence of first and second interval responses is conditionally independent.

During habituation, only FIRs were scored as they reflect orienting to the novel stimuli (Öhman, 1983) and anticipation of the unconditional stimulus is not expected during this phase. The largest response starting within the latency window was scored and the magnitude of the response was calculated as the difference between response onset and peak (Prokasy & Kumpfer, 1973). Electrodermal responses were square root transformed to reduce the positive skew of the distribution (Dawson et al., 2007) and then range corrected to ensure that each participant was given an even weight in the analyses, reducing the influence of outliers (Boucsein et al., 2012; Dawson et al., 2007). The reference used for the range correction was the largest response displayed by the participant, typically the response to the first or second presentation of the unconditional stimulus. In case of multiple responses, the largest response starting within the latency window was scored, regardless of whether the peak of the response was within the same latency window (Prokasy & Kumpfer, 1973). Electrodermal responding and valence ratings were averaged into blocks of two consecutive trials to reduce the influence of trial-by-trial variability. The conditional stimulus valence ratings were scored as the largest voltage deviation form a one second pre-stimulus baseline voltage that occurred within the six second CS presentation.

First and second interval electrodermal responding and conditional stimulus valence evaluations were subjected to separate 2 × 2 × n [Group (Instruction/removal, Control) × CS (CS+, CS-) × Block (Habituation = 2, Acquisition = 4, Extinction = 4)] factorial ANOVAs. As the influence of the instructional manipulation is expected in early extinction, additional 2 × 2 × 2 [Group (Instruction/removal, Control) × CS (CS+, CS-) × Block (2)] factorial ANOVAs were performed on the electrodermal responding and conditional stimulus valence during early extinction. To examine whether differential responding was still present during the last block of extinction 2 × 2 factorial ANOVAs [Group (Instruction/removal, Control) × CS (CS+, CS-)] were performed on the electrodermal responding and conditional stimulus valence during the last block of extinction. Unconditional electrodermal responding during acquisition was subjected to a 2 × 4 [Group (Instruction/removal, Control) × Block (4)] factorial ANOVA. Pre- and post-experimental ratings were subjected to a 2 × 2 × 2 [Group (Instruction/removal, Control) × CS (CS+, CS-) × Phase (Pre- and Post-experimental)] factorial ANOVA.

Multivariate F values (Phillai's Trace) and partial eta-squares are reported for all main effects and interactions. All main and simple effect comparisons were conducted using Bonferroni adjustments to protect against the accumulation of α –error and adjusted *p* values are reported for these follow-up analyses. IBM SPSS Statistics 21 was used to conduct all analyses, and the significance level was set at .05 for all analyses.

Results

Preliminary Checks. Preliminary analyses revealed no difference between groups in age (Instruction/removal: *M* = 21.17 years, *SD* = 4.30 years; Control: *M* = 21.71 years, *SD* = 8.91 years), $t(33) = 0.23$, $p = .820$, the number of spontaneous electrodermal responses during the three minute baseline period (Instruction/removal: *M* = 20.58 responses, *SD* = 13.26 responses; Control: *M* = 21.47 responses, *SD* = 9.81 responses), t(34) = 0.23, $p = .822$, the US intensity set by the participant (Instruction/removal: *M* = 36.32 V, *SD* = 11.28 V; control: *M* = 35.00 V, *SD* = 9.35 V), t(34) = 0.38, $p = .708$, and the rated US unpleasantness (Instruction/removal: *M* = -1.61, *SD* = 1.40; Control: *M* = -2.06, *SD* = 0.43), t(34) = 1.28, $p = .209$. The female to male ratio was larger in the control group (13:4) in comparison with the instruction/removal group (7:12), $\chi^2(1) = 5.71$, $p = .017$. Analysis of the unconditioned electrodermal responses (responses to the electrotactile stimulus) during acquisition revealed a main effect of block, $F(3, 31) = 25.77$, $p < .001$, $\eta p^2 = .714$. Electrodermal responding in block one was significantly higher than in blocks two, $p < .001$, three, $p < .001$, and four, $p < .001$. One participant in the control group failed to indicate which face had preceded the electrotactile stimulus. The analyses were re-run excluding this participant, but as this exclusion did not alter the pattern of results, the analyses of the entire sample have been reported. All participants in the instruction/removal group indicated that they believed the instructions.

Electrodermal responses. The first interval electrodermal responses for habituation are presented in the left panel of Figure 1. During habituation, first interval electrodermal responses declined from block one (*M* = 0.28, *SD* = 0.20), to block two (*M* = 0.20, *SD* = 0.20), as indicated by a main effect of block, $F(1, 33) = 10.19$, $p = .003$, $\eta p^2 = .236$. All remaining main effects and interactions did not attain significance, closest (main effect of group) $F(1, 33) = 1.02$, $p = .321$, $\eta p^2 = .030$.

The first and second interval electrodermal responses during acquisition are summarized in Figure 1 (middle panel), and Figure 2 (left panel), respectively. During acquisition differential responding between the CS+ and the CS- emerged in both the first and the second interval responses for both groups. Analysis of the first interval responses, revealed a main effect of CS, $F(1, 33) = 46.07$, $p < .001$, $\eta p^2 = .583$, a main effect of block, $F(3, 31) = 4.86$, $p = .007$, $\eta p^2 = .320$, and a CS × block interaction, $F(3, 31) = 3.55$, $p = .026$, $\eta p^2 = .256$. This interaction confirmed that the CS+ and the CS- elicited similar levels of responding at block one, $F(1, 33) =$

1.41, p = .244, ηp^2 = .041, but that responding to the CS+ was larger than responding to the CS- in blocks two, $F(1, 33) = 24.79$, $p < .001$, ηp^2 = .429, three $F(1, 33) = 31.89$, $p < .001$, ηp^2 = .491, and four $F(1, 33) = 13.08$, p = .001, ηp^2 = .284. All other main effects and interactions did not attain significance, closest (main effect of group), $F(1, 33) = 1.40$, p = .245, ηp^2 = .041.

Insert Figure 1 about here

Analysis of the second interval responses revealed a main effect of CS, $F(1, 33) = 16.37$, $p < .001$, ηp^2 = .332, a main effect of group, $F(1, 33) = 6.27$, p = .017, ηp^2 = .160, a CS × block interaction, $F(3, 31) = 9.67$, $p < .001$, ηp^2 = .483, and a CS × block × group interaction, $F(3, 31)$ = 3.83, p = .019, ηp^2 = .270. Follow up analyses revealed that in the control group, responding to the CS- was larger than to CS+ at block one, $F(1, 33) = 8.29$, p = .007, ηp^2 = .201. At block two, responses to the CS+ and the CS-, did not differ, $F(1, 33) = 0.19$, p = .669, ηp^2 = .006, and at blocks three, $F(1, 33) = 34.11$, $p < .001$, ηp^2 = .508, and four $F(1, 33) = 4.94$, p = .033, ηp^2 = .130, CS+ elicited larger responses than CS-. In the instruction/removal group, no difference in responding between CS+ and CS- was detected at blocks one, $F(1, 33) = 0.21$, p = .649, ηp^2 = .006, or two, $F(1, 33) = 1.62$, p = .212, ηp^2 = .047, whilst larger responding was elicited by the CS+ at blocks three, $F(1, 33) = 6.59$, p = .015, ηp^2 = .166, and four, $F(1, 33) = 4.35$, p = .045, ηp^2 = .116. All other main effects and interactions did not attain significance, closest (block × group), $F(3, 31) = 1.78$, p = .172, ηp^2 = .147.

Insert Figure 2 about here

The first and second interval electrodermal responses recorded during extinction are summarized in the right panels of Figures 1 and 2. The differential responding between the CS+ and the CS-, acquired during acquisition, was not present in the first interval responses of extinction. This was the result of increased responding to the CS-, in the control group, and decreased responding to the CS+, in the instruction/removal group. A main effect of block, $F(3, 31) = 10.88$, $p < .001$, $\eta p^2 = .513$, a main effect of group, $F(1, 33) = 10.26$, $p = .003$, $\eta p^2 = .237$, and a block × group interaction, $F(3, 31) = 6.79$, $p = .001$, $\eta p^2 = .396$, confirmed these impressions. Follow up analyses revealed that at block one, $F(1, 33) = 18.01$, $p < .001$, $\eta p^2 = .353$, and two, $F(1, 33) = 13.06$, $p = .001$, $\eta p^2 = .284$, responding was larger in the control group in comparison with the instruction/removal group, whilst at block three, $F(1, 33) = 2.97$, $p = .094$, $\eta p^2 = .082$, and four, $F(1, 33) = 1.40$, $p = .245$, $\eta p^2 = .041$, the groups did not differ in responsiveness. All other main effects and interactions did not attain significance, closest (CS × group), $F(1, 33) = 1.80$, $p = .188$, $\eta p^2 = .052$.

This group effect was confirmed when the analyses were run only examining block one and two of extinction. Responding was larger in the control group ($M = 0.25$, $SD = 0.20$) in comparison with the instruction/removal group ($M = 0.07$, $SD = 0.10$) as confirmed by a main effect of group, $F(1, 33) = 17.66$, $p < .001$, $\eta p^2 = .349$. A main effect of block confirmed that responding was larger in block one ($M = 0.18$, $SD = 0.18$) in comparison with block two ($M = 0.13$, $SD = 0.18$), $F(1, 33) = 10.19$, $p = .003$, $\eta p^2 = .236$. The remaining main effects and interactions did not reach significance, closest (CS × block), $F(1, 33) = 1.87$, $p = .181$, $\eta p^2 = .054$. To further confirm that differential responding had been eliminated in both groups an 2 × 2 × 2 [Group (Instruction/removal, Control) × CS (CS+, CS-) × Phase (Last block of acquisition, First block of extinction) factorial ANOVA was performed, yielding a main effect of CS, $F(1, 33) = 15.92$, $p < .001$, $\eta p^2 = .325$, a main effect of group, $F(1, 33) = 8.68$, $p = .006$, $\eta p^2 = .325$ and a phase × group interaction, $F(1, 33) = 9.13$, $p = .005$, $\eta p^2 = .217$. Follow-up analyses revealed that during the last block of acquisition there was no difference in responding between the instruction/removal and control groups, $F(1, 33) = 1.59$, $p = .217$, $\eta p^2 = .046$, however during the first block of acquisition responding in the instruction/removal group was significantly reduced in comparison to the control group. The remaining main effects and interactions did not reach significance, closest (phase × CS), $F(1, 33) = 3.39$, $p = .075$, $\eta p^2 = .093$. When responses in the last block of extinction were analyzed no main effects or interactions attained significance, closest (main effect of group), $F(1, 33) = 1.40$, $p = .245$, $\eta p^2 = .041$, confirming that differential first interval electrodermal responding between the CS+ and the CS- had been eliminated in both groups at the end of extinction.

Inspection of the right panel of Figure 2 suggests that differential second interval electrodermal responding was present during early extinction in the control group, but not in the instruction/removal group. Analyses of responses from the entire extinction phase revealed a main effect of block, $F(3, 31) = 3.60$, $p = .024$, $\eta p^2 = .258$, a main effect of group, $F(1, 33) = 6.05$, $p = .019$, $\eta p^2 = .155$, and a block × group interaction, $F(3, 31) = 4.26$, $p = .012$, $\eta p^2 = .292$. Follow-up analyses confirmed that at block one responding was larger in the control group, in comparison with the instruction/removal group, $F(1, 33) = 10.66$, $p = .003$, $\eta p^2 = .244$. No differences in responding were detected between the groups at blocks two, $F(1, 33) = 1.40$, $p = .246$, $\eta p^2 = .041$, three, $F(1, 33) = 0.11$, $p = .747$, $\eta p^2 = .003$, and four, $F(1, 33) = 3.85$, $p = .058$, $\eta p^2 = .105$. All other main effects and interactions failed to attain significance, closest (main effect of CS), $F(1, 33) = 3.43$, $p = .073$, $\eta p^2 = .094$.

As the influence of the instructional manipulation on differential responding was expected in early extinction, the analyses were run including only blocks one and two. This revealed a main effect of group, $F(1, 33) = 7.62$, $p = .009$, $\eta p^2 = .187$, a block × group interaction, $F(1, 33) = 4.76$, $p = .036$, $\eta p^2 = .126$, a CS × block interaction, $F(1, 33) = 4.31$, $p = .046$, $\eta p^2 = .115$, and a CS × block × group interaction, $F(1, 33) = 6.63$, $p = .015$, $\eta p^2 = .167$.

Follow up analyses revealed that, in the control group, the CS+ elicited larger responses than the CS- at block one, $F(1, 33) = 10.74$, $p = .002$, $\eta p^2 = .246$, but not at block two, $F(1, 33) = 0.32$, $p = .579$, $\eta p^2 = .009$. Conversely, in the instruction/removal group, there was no differential responding at block one, $F(1, 33) = 0.03$, $p = .865$, $\eta p^2 = .001$, and block two, $F(1, 33) = 0.36$, $p = .552$, $\eta p^2 = .011$. All other main effects and interactions did not reach significance, largest (main effect of block), $F(1, 33) = 3.32$, $p = .078$, $\eta p^2 = .091$.

When only the last block of extinction was included in the analyses no main effects of interactions attained significance, closest (main effect of group), $F(1, 33) = 3.85$, $p = .058$, $\eta p^2 = .105$. This marginal group effect reflected increased responding in the control group in comparison with the instruction removal group. In both groups differential second interval responding between the CS+ and the CS- was no longer present at the end of extinction.

Conditional Stimulus Valence Evaluations. The conditional stimulus valence evaluations obtained during habituation (left), acquisition (middle), and extinction (right), for both groups are summarized in Figure 3. Analysis of the valence evaluations recorded during habituation revealed a CS × block × group interaction, $F(1, 31) = 4.55$, $p = .041$, $\eta p^2 = .128$. The CS- was rated less pleasant in block one than in block two in the instruction/removal group, $F(1, 31) = 4.81$, $p = .036$, $\eta p^2 = .134$. All other comparisons failed to reach significance, largest (control CS+, block one in comparison with block two), $F(1, 31) = 1.45$, $p = .237$, $\eta p^2 = .045$. The remaining main effects and interactions did not reach significant, largest (main effect of block), $F(1, 31) = 1.11$, $p = .300$, $\eta p^2 = .035$.

Insert Figure 3 about here

At the beginning of acquisition, the pleasantness ratings of CS+ and CS- did not differ, but as the experiment progressed, the CS+ was rated less pleasant than the CS- in both groups.

A main effect of CS, $F(1, 31) = 12.44$, $p = .001$, $\eta p^2 = .286$, and a CS × block interaction, $F(3, 29) = 5.19$, $p = .005$, $\eta p^2 = .349$, confirmed these impressions. Follow up analyses revealed that the CS+ and the CS- were rated similarly at block one, $F(1, 31) = 1.18$, $p = .286$, $\eta p^2 = .037$, but at blocks two, $F(1, 31) = 12.94$, $p = .001$, $\eta p^2 = .294$, three, $F(1, 31) = 12.36$, $p = .001$, $\eta p^2 = .285$, and four, $F(1, 31) = 13.32$, $p = .001$, $\eta p^2 = .300$, the CS+ was rated as less pleasant than the CS-. The remaining main effects and interactions did not reach significant, largest (CS × block × group interaction), $F(3, 29) = 1.60$, $p = .211$, $\eta p^2 = .142$.

During extinction, both groups gave lower pleasantness ratings to the CS+ ($M = -0.82$, $SD = 0.81$) in comparison with the CS- ($M = -0.12$, $SD = 0.87$), and both conditional stimuli were rated as more pleasant as extinction progressed. The analyses confirmed these impressions revealing main effects of CS, $F(1, 31) = 15.79$, $p < .001$, $\eta p^2 = .337$, and block, $F(3, 29) = 5.49$, $p = .004$, $\eta p^2 = .362$. Follow up analyses revealed that when compared with block one, the evaluations in block two, $p = .014$, and three, $p = .025$, were more pleasant. All other comparisons were not significant, largest (block one in comparison with block four), $p = .094$. All other main effects and interactions did not attain significance, largest (CS × block) $F(3, 29) = 2.15$, $p = .116$, $\eta p^2 = .182$.

As the influence of the instructional manipulation was expected in early extinction, the analyses were run excluding blocks three and four. This did not change the pattern of results, with analyses revealing a main effect of CS, $F(1, 31) = 18.36$, $p < .001$, $\eta p^2 = .372$, and a main effect of block, $F(1, 31) = 11.05$, $p = .002$, $\eta p^2 = .263$. All remaining main effects and interactions did not attain significance, largest (block × group, $F(1, 31) = 1.66$, $p = .207$, $\eta p^2 = .051$. To examine whether differential valence ratings were still present at the end of extinction the analyses were re-run including only block 4. This revealed a main effect of CS, $F(1, 31) =$

7.78, p = .009, ηp^2 = .201, which confirmed that across groups CS+ (M = -0.64, SD = 0.89) was rated as less pleasant than CS- (M = -0.10, SD = 0.99) during the last block of extinction. The remaining main effects and interactions did not reach significance, largest (main effect of group), $F(1, 31) = 1.03$, p = .318, ηp^2 = .032.

Pre- and Post-Experimental Pleasantness Ratings. The pleasantness ratings recorded before habituation and after extinction are displayed on the left side of Figure 4. Before habituation the CS+ and the CS- received similar pleasantness ratings, however after extinction the CS+ was rated less pleasant than the CS-. This pattern emerged consistently for both groups. The analyses confirmed these impressions, revealing a period × CS interaction, $F(1, 32) = 10.04$, p = .003, ηp^2 = .239. Follow-up analyses revealed that before the experiment, pleasantness ratings of the CS+ and the CS- did not differ, $F(1, 32) = 0.87$, p = .358, ηp^2 = .026, but after the experiment the CS+ was rated less pleasant than the CS-, $F(1, 32) = 9.19$, p = .005, ηp^2 = .223. All remaining effects did not reach significance, largest (main effect of CS), $F(1, 32) = 3.43$, p = .073, ηp^2 = .097.

Insert Figure 4 about here

Experiment Two

Method

Participants. Forty-four (26 female) undergraduate students volunteered participation in exchange for course credit. The participants' ages ranged from 16-59 (M = 22.77). All participants consented to the experiment and were fully informed. Participants were randomly assigned to one of two groups (instruction/removal, control). Recording error resulted in the loss of two participants' pre-experimental ratings data, four participants' post-experimental ratings data, and one participants' fear potentiated startle data. These participants were included in the analyses of all remaining measures.

Apparatus/Stimuli. Orbicularis oculi electromyography (EMG) was measured using two 4mm Ag/AgCl electrodes, one placed directly underneath the participants' left eye, and another below the corner of the left eye, approximately 1 cm to the left of the first electrode. A reference electrode was placed in the middle of the participants' forehead. All electrodes were fitted with adhesive collars and filled with a standard electrode gel, and impedances were checked to ensure they were lower than 10 kΩ. Orbicularis oculi EMG was recorded using AcqKnowledge Version 3.9.1 with a Biopac MP150 system, at a sampling frequency of 1000 Hz, and an amplification factor of 5000. Raw EMG was bandpass filtered with a low cut-off of 10 Hz and a high cut-off of 500 Hz. Startle blinks were elicited with a 105 dB bursts of white noise lasting 43 ms with an instantaneous rise time, generated by a custom built noise generator and presented through Sennheiser headphones. Startle probes were presented 3.5 s or 4.5 s after the onset of the conditional stimulus and during the inter-trial intervals, seven seconds after the conditional stimulus offset and eight seconds before the onset of the next conditional stimulus. Before any stimulus presentations, three startle probes were presented to habituate startle responding, and to allow for a comparison of baseline startle magnitude between the groups. Two, four, and six startle probes were presented during CS+ and CS- in habituation, acquisition and extinction, respectively. Four probes were presented in the inter-trial interval of habituation, eight in acquisition, and twelve in extinction. During habituation, startle probes were placed in the 2^{nd} and 4^{th} presentation of both the CS+ and the CS-. During acquisition startle probes were placed in the 3^{rd}, 4^{th}, 6^{th} and 8^{th} presentation of the CS-; and the 2^{nd}, 4^{th}, 6^{th}, and 8^{th} presentation of the CS+. During extinction startle probes were placed in the 2^{nd}, 4^{th}, 6^{th}, 7^{th}, 10^{th} and 12^{th} presentation of the CS-; and the 2^{nd}, 4^{th}, 5^{th}, 8^{th}, 10^{th} and 12^{th} presentation of the CS+.

Procedure. Eight additional trials (four CS+ and four CS-) were added during extinction, in order to allow sufficient time to examine changes in fear potentiated startle. Counterbalancing and the remainder of the procedure were conducted in the same manner as Experiment 1.

Scoring and Response Definition. Raw EMG , was filtered offline, (Band stop at 50 Hz followed by a bandpass filter, low cut-off of 30 Hz and a high cut-off of 500 Hz) rectified and smoothed (five point moving average). Blink startle magnitude was defined as the maximum of the rectified and smoothed response curve occurring within 120 ms of the stimulus onset (Blumenthal et al., 2005). A trial was defined as missing if the baseline EMG recorded 50 ms prior to probe onset was judged by visual inspection to be unstable, or if a spontaneous or voluntary blink immediately preceded the startle probe onset. A trial was defined as a non-response trial if no response onset could not be identified within 20-60 ms of probe onset. Blink startle magnitudes elicited during the conditional stimuli were averaged into blocks of two consecutive trials, yielding one block for habituation, two blocks for acquisition and three blocks in extinction. Using all startles measured during conditioning as the reference distribution, t-scores were calculated to reduce the impact of individual differences.

Startle magnitudes were subjected to separate 2 × 2 × n [Group (Instruction/removal, control) × CS (CS+, CS-) × Block (Habitation = 1, Acquisition = 2, Extinction = 3)] factorial ANOVAs. The remaining analyses were conducted in the same manner as for Experiment 1.

Results

Preliminary Checks. No differences between the groups were detected for age (Instruction/removal: *M* = 22.77 years, *SD* = 9.82; Control: *M* = 22.36 years, *SD* = 6.07), $t(42) = 0.17$, $p = .869$, gender (Instruction/removal: 8 male: 14 female; Control: 10:12), $\chi^2(1) = 0.38$, $p =$

.540, or the magnitude of the blink startle responses elicited during the baseline period (Instruction/removal: *M* = 190, *SD* = 118; Control: *M* = 190 , *SD* = 132), *t*(40) = 0.06, *p* = .954. The US intensity level set (Instruction/removal: *M* = 32.91 V, *SD* = 7.28 V; Control *M* = 31.68 V, *SD* = 10.97 V), t(42) = 0.44, *p* = .664, and the rated US unpleasantness level (Instruction/removal: *M* = -1.82, *SD* = 0.66; Control *M* = -1.86, *SD* = 0.71), t(42) = 0.22, *p* = .828, were similar in both groups.

Two participants in the control group, and one participant in the instruction/removal group failed to correctly identify which face had been paired with the US. One participant in the instruction/removal group reported not believing the instructional manipulation. The analyses were run excluding these participants, and this strengthened the pattern of fear potentiated startle magnitude results reported for the CS × group interaction during extinction. The results from the exclusion sample have been reported in addition to the results of the entire sample for this interaction, but as the remaining results were not altered, they are reported for the entire sample.

Fear Potentiated Startle. The magnitude of the blink startle responses recorded during habituation (left), acquisition (middle), and extinction (right) are summarized in Figure 5. In habituation, there were no differences in startle magnitude during CS+ and CS-, or between the groups, largest (CS × group), *F*(1, 41) = 0.22, *p* = .641, ηp^2 = .005. During acquisition, fear potentiated startle magnitude was larger during the CS+ (*M* = 55.12, *SD* = 7.56), than during the CS- (*M* = 51.05, *SD* = 7.21), and fear potentiated startle magnitude decreased from block one (*M* = 55.50, *SD* = 7.81), to block two (*M* = 50.67, *SD* = 6.96). The analyses confirmed these impressions yielding main effects of CS, *F*(1, 40) = 11.18, *p* = .002, ηp^2 = .218, and block, *F*(1, 40) = 16.91, *p* < .001, ηp^2 = .297. The remaining main effects and interactions did not attain significance, largest (group), *F*(1, 40) = 3.33, *p* = .075, ηp^2 = .077. This marginal group effect indicated that startle magnitude was marginally higher in the instruction/removal group (*M* = 54.14, *SD* = 7.25) in comparison to the control group (*M* = 52.02, *SD* = 7.37).

Insert Figure 5 about here

During extinction, startle magnitude decreased with time, as confirmed by a main effect of block, $F(2, 37) = 10.52$, $p < .001$, $\eta p^2 = .363$. Follow up analyses revealed that startle magnitude was larger during block one, in comparison to block two, $p < .001$, and three, $p < .001$, but that bocks two and three did not differ, $p > .999$. A marginal block × group interaction, $F(2, 37) = 3.12$, $p = .054$, $\eta p^2 = .146$, revealed that responding in the control group differed significantly between blocks 1 and 2 ($p < .001$) and blocks 1 and 3 ($p = .007$), $F(2, 37) = 9.86$, $p < .001$, $\eta p^2 = .348$, whereas responding in the instruction/removal group only differed between blocks 1 and 3 ($p = .026$), $F(2, 37) = 3.82$, $p = .031$, $\eta p^2 = .171$. The remaining main effects and interactions did not attain significance, largest (main effect of CS), $F(1, 38) = 3.39$, $p = .073$, $\eta p^2 = .082$.

As the influence of the instructional manipulation was expected in early extinction, the analyses were run excluding blocks two and three. This revealed a marginal CS × group interaction, $F(1, 41) = 3.90$, $p = .055$, $\eta p^2 = .087$, for the entire sample. The remaining main effects and interactions did not attain significance, largest (main effect of group), $F(1, 41) = 3.38$, $p = .073$, $\eta p^2 = .076$. When the analyses were re-run excluding the three participants who failed to identify the experimental contingencies correctly and the participant who did not believe the instructional manipulation, the CS × group interaction attained significance, $F(1, 37) = 4.84$, $p = .034$, $\eta p^2 = .116$. Follow up analyses revealed that, in the control group, startle magnitude was larger during the CS+ than during the CS-, $F(1, 37) = 4.54$, $p = .040$, $\eta p^2 = .109$ (Full Sample: $F(1, 41) = 4.34$, $p = .043$, $\eta p^2 = .096$), but that no difference was present in the instruction/removal group, $F(1, 37) = 0.94$, $p = .339$, $\eta p^2 = .025$ (Full Sample: $F(1, 41) = 0.48$, $p = .492$, $\eta p^2 = .012$). Examining the last block of extinction revealed that across groups responding to the CS+ ($M = 46.54$, $SD = 5.27$) was still marginally larger than responding to the CS- ($M = 44.29$, $SD = 5.02$), $F(1, 40) = 3.79$, $p = .059$, $\eta p^2 = .087$. The remaining effects did not reach significance, largest (main effect of group), $F(1, 40) = 2.54$, $p = .119$, $\eta p^2 = .060$.

Conditional Stimulus Valence Evaluations. The conditional stimulus valence evaluations recorded during habituation (left), acquisition (middle), and extinction (right) are summarized in Figure 6. During habituation, no significant differences were detected between the groups or between the conditional stimuli, largest $F(1, 42) = 3.31$, $p = .076$, $\eta p^2 = .073$, (block × group). In acquisition, a main effect of CS was detected, $F(1, 42) = 10.33$, $p = .003$, $\eta p^2 = .197$, revealing that the CS+ ($M = -0.94$, $SD = 0.95$) was rated less pleasant than the CS- ($M = -0.42$, $SD = 1.05$). The remaining main effects and interactions did not attain significance, closest (CS × block), $F(3, 40) = 2.65$, $p = .062$, $\eta p^2 = .166$

Analyses of the extinction phase, revealed a main effect of CS, $F(1, 41) = 6.58$, $p = .014$, $\eta p^2 = .138$, confirming that the CS+ ($M = -0.84$, $SD = 0.94$), continued to be rated less pleasant than the CS- ($M = -0.47$, $SD = 0.94$). A main effect of block was detected, $F(5, 37) = 2.62$, $p = .040$, $\eta p^2 = .262$, revealing that ratings in block four were more pleasant, than ratings in block two, $p = .030$. The other comparisons failed to reach significance, largest, $p = .062$ (responses in block five in comparison with block two). The remaining main effects and interactions failed to reach significance, largest (CS × block), $F(5, 37) = 2.36$, $p = .059$, $\eta p^2 = .242$ (CS+ was rated less pleasant than CS- in blocks one and two but did not differ in the remaining blocks).

Insert Figure 6 about here

To assess the influence of the instructional manipulation in early extinction, the analyses were re-run including only blocks one and two. This revealed a main effect of CS $F(1, 42) = 15.94$, $p < .001$, $\eta p^2 = .275$, and a CS × block interaction $F(1, 42) = 8.32$, $p = .006$, $\eta p^2 = .165$. Follow up analyses revealed that during block one the CS+ was rated less pleasant than during block two, $F(1, 42) = 4.78$, $p = .034$, $\eta p^2 = .102$, whilst the pleasantness evaluations did not differ between blocks for the CS-, $F(1, 42) = 3.31$, $p = .076$, $\eta p^2 = .073$. The block × group interaction attained marginal significance, $F(1, 42) = 3.89$, $p = .055$, $\eta p^2 = .085$, follow-up analyses revealed marginally more positive evaluations in block one, in comparison with block two, in the control group, $F(1, 42) = 2.95$, $p = .093$, $\eta p^2 = .066$, but not in the instruction/removal group, $F(1, 42) = 1.14$, $p = .291$, $\eta p^2 = .026$. However, the CS × block × group interaction was not significant, $F(1, 42) = 0.07$, $p = .797$, $\eta p^2 = .002$, confirming that the instructional manipulation did not differentially affect the conditional stimulus valence evaluations. The main effects and interactions did not reach significance, closest (main effect of group), $F(1, 42) = 0.41$, $p = .527$, $\eta p^2 = .010$. When only the last block of extinction was included in the analyses no main effects or interactions attained significance, closest (main effect of CS), $F(1, 41) = 2.70$, $p = .108$, $\eta p^2 = .062$, confirming that differential ratings of the CS+ and the CS- had extinguished in both groups.

Pre- and Post-Experimental Pleasantness Ratings. The right panel of Figure 4 summarizes the pleasantness ratings recorded before habituation and after extinction. Before habituation, the CS+ and the CS- received similar pleasantness ratings, however after extinction the CS+ was rated less pleasant than the CS-, a pattern that emerged consistently for both groups. These impressions were confirmed by a main effect for CS, $F(1, 37) = 7.86$, $p = .008$, $\eta p^2 = .175$ and a period × CS interaction, $F(1, 37) = 4.69$, $p = .037$, $\eta p^2 = .112$. Follow-up analyses revealed that before the experiment, pleasantness ratings of the CS+ and the CS- did not differ, $F(1, 37) =$

0.47, p = .499, ηp^2 = .012, but after the experiment the CS+ was rated less pleasant than the CS-, $F(1, 37)$ = 7.95, p = .008, ηp^2 = .177. The remaining main effects and interactions did not attain significance, largest, (CS × group), $F(1, 37)$ = 1.08, p = .306, ηp^2 = .028.

Discussion

Experiments 1 and 2 aimed to assess the influence of instructed extinction and removal of the shock electrode on electrodermal responses, fear potentiated startle, and conditional stimulus valence during fear conditioning. In Experiment 1 instructed extinction and removal of the shock electrode resulted in the elimination of differential electrodermal responding, but did not affect conditional stimulus valence ratings. This pattern of results was replicated and extended in Experiment 2. Instructed extinction and removal of the shock electrode eliminated differential startle modulation at the beginning of extinction, whilst, the differential valence evaluations were not affected.

The current findings suggest that instructed extinction and removal of the shock electrode results in the immediate decline of differential physiological responding, but does not affect indices of conditional stimulus valence. As modulation of the startle reflex is not under conscious control (Lang, Bradley, & Cuthbert, 1990), the results of the physiological measures used in Experiment 2 are unlikely to reflect demand characteristics. Conversely, subjective valence ratings are susceptible to the effects of demand characteristics (Mitchell, Anderson, & Lovibond, 2003), as they are under the conscious control of the participants. To ensure that the findings of Experiments 1 and 2 reflect a true dissociation between physiological measures and conditional stimulus valence, an explanation of the current results based on demand characteristics should be excluded.

Experiment Three

Demand characteristics can influence the outcome of an experiment when the participants can correctly infer the experimental hypotheses and desire to respond according to them (Mitchell et al., 2003). The participants in Experiments 1 and 2 might have hypothesized that consistent differential responding throughout the experiment was expected and therefore continued to differentially rate the conditional stimuli throughout extinction. If so, the results obtained could reflect demand characteristics rather than a failure of instructed extinction to affect conditional stimulus valence.

Cacioppo, Marshall-Goodell, Tassinary, and Petty (1992) developed a method to assess demand characteristics explicitly in a separate sample of participants. To determine whether the participants might have been able to infer the experimental hypothesis, and respond accordingly, they asked participants to read a detailed description of a particular experiment and predict its outcome. They argued that a demand characteristic explanation would be implausible if the participants were not able to predict the results of the prior experiment. In Experiment 3, we utilized this methodology to examine whether the results of the Experiments 1 and 2 could reflect demand characteristics.

Method

Participants. Sixty-three (56 female; age range: 17-42; *M* = 20.54) undergraduate students who had not participated in Experiments 1 or 2 volunteered participation in exchange for course credit and provided informed consent.

Demand Questionnaire Measure. The demand characteristic questionnaire is shown in Appendix 1. The questionnaire consisted of a description of the acquisition and extinction phase of the instructed extinction experiment, as well as a series of questions requiring the participants to predict the results of the experiment. Heart rate was chosen as an example of a physiological response as it seemed more familiar than electrodermal responding or fear potentiated startle to a first year undergraduate sample.

Procedure. Participants were instructed to read the descriptions and questions carefully and to answer as if they were trying to predict the outcome of the study.

Questionnaire Scoring. The responses to each question were examined and coded as describing either an increase, a decrease, or no change. For example, a response like ‘the pleasantness rating will drop’ would be recorded as a ‘decrease’; a response like ‘the physiological responses will increase in response to the face paired with the shock’ would be recorded as an ‘increase’; and a responses like ‘I don’t think the pleasantness rating of the CS- will change’ was recorded as a ‘no change’. If the participant’s response could not be categorized into one of the three response categories it was recorded as missing. For each question, the results were calculated as a percentage of people who predicted each outcome.

Results

Demand Questionnaire Responses. The predictions obtained from the demand characteristic questionnaire are displayed in Table 1. In the acquisition scenario, the most common pattern of results reported was that the CS+ would become more unpleasant, and elicit larger physiological responses throughout acquisition; whereas the CS- would become more pleasant, and result in reduced physiological responses throughout acquisition. In the instructed extinction scenario, the most common pattern of results predicted was that on the first trial of extinction, the physiological responses to both the CS+ and the CS- would not change, whereas the evaluations of the CS+ would increase in pleasantness, and the evaluations of the CS- would stay the same.

Insert Table 1 about here

Discussion

Experiment 3 aimed to assess whether the findings of Experiments 1 and 2 reflect on demand characteristics. The method used by Cacioppo et al. (1992) was implemented asking a separate sample of participants to predict the outcome of the experiment after reading a detailed description of the instructed extinction procedure. The majority of participants predicted that the physiological responding would not change, but that the ratings of the CS+ would become more pleasant on the first trial after the instructional manipulation. That is, they predicted a dissociation between the physiological indices of fear learning and conditional stimulus valence in the opposite direction to that observed in Experiments 1 and 2. These results suggest that the results of Experiments 1 and 2 are unlikely to reflect on demand characteristics.

General Discussion

The current study examined the effect of instructed extinction and removal of the shock electrode on physiological indices of human fear learning and conditional stimulus valence. In Experiment 1, instructed extinction resulted in the immediate elimination of differential second interval electrodermal responding (Experiment 1) and differential startle magnitude (Experiment 2) in the instruction/removal group, while differential responding remained intact at the beginning of extinction in the control group. In both experiments conditional stimulus valence ratings did not respond to instructed extinction as shown by continued differential ratings between CS+ and CS- in both groups at the beginning of extinction. This is to our knowledge the first study showing that instructed extinction has no effect on conditional stimulus valence in a differential fear conditioning paradigm, whilst simultaneously showing an effect on the physiological indices of human fear learning. This pattern of results replicates previous instructed extinction studies (Mallan et al. 2009; Rowles et al., 2012) and extends work suggesting conditional stimulus valence is not responsive to instructed extinction (Lipp et al., 2010) into a fear conditioning setting.

It should be noted that in Experiment 1 the experimental manipulation affected differential first interval responding in both participant groups, although in a different manner. In the instruction/removal group responding to the CS+ decreased from the last block of acquisition to the first block of extinction, whereas in the control group, responding to CS- increased from the last block of acquisition to the first block of extinction. These changes led to a between group difference in responding at the beginning of extinction but not the expected Group x CS interaction. Rowles et al. (2012) reported a similar increase in first interval responding to the CS- in the control group during early extinction. It is likely that increased responding to the CS- in the control group reflects sensitization of the orienting reflex to CS- due to the interaction with the experimenter, an effect not seen in the instruction/removal group as they were provided with safety information. The expected Group x CS interaction was evident in second interval responding which is less affected by orienting and more selectively sensitive to unconditional stimulus anticipation. This differential pattern of results across response windows supports the notion of using separate latency windows when scoring electrodermal responding (Prokasy & Kumper, 1973).

In Experiment 2 the Group x CS interaction attained marginal significance ($p = 0.55$) when the entire sample was considered and was significant after removal of two participants who were unable to verbalize the experimental contingencies and one participant who reported not believing the instructions ($p = .034$). Follow-up analyses of both interactions (full or excluded sample) revealed that startle magnitude was larger during the CS+ than during the CS- in the control group, but not in the instruction/removal group. Exclusion of participants who fail to provide evidence of learning in a differential fear conditioning paradigm or fail a manipulation check is not uncommon in human fear research and the fact that similar patterns emerge even if these participants are retained speaks to the robustness of the results.

As conditional stimulus evaluations can be susceptible to demand characteristics, we explicitly assessed participants' predictions of the experimental results in Experiment 3. After reading a detailed description of the study, the majority of participants predicted that instructed extinction would affect the conditional stimulus evaluations, but not physiological responding. As this prediction is not consistent with the pattern of results observed in Experiments 1 and 2, it seems unlikely that these results reflect on demand characteristics. It is possible that the demand characteristics of the participants predicting the outcome of a study they read about may differ from those of participants who are actually in the experimental situation. However, it seems unlikely that the demand characteristics developed in the latter group would be opposite to those developed in the former.

The current findings suggest that the negative valence acquired during fear conditioning is not responsive to cognitive interventions, a finding with significant clinical importance as cognitive interventions are commonly used in treatments for anxiety disorders. If persisting negative valence does drive return of fear as proposed by Kerkhof et al. (2012) and suggested in the data of Hermans et al. (2005) then the current findings highlight the importance of using extended extinction training to reduce negative valence of the feared stimulus. Conditional stimulus valence has been shown to resist extinction in comparison with physiological indices of human fear learning, however extended extinction training can be effective at reducing differential valence ratings (Lipp, Oughton & LeLievre, 2003). This is supported by the finding that a significant differential valence evaluation of CS+ and CS- was still present at the end of extinction in Experiment 1, but not in Experiment 2 which utilized a larger number of extinction trials.

The current study highlights the importance of future research to identify ways in which conditional stimulus valence can be effectively reduced. Although the current study provides evidence that conditional stimulus valence is not sensitive to verbal instructions that target the stimulus contingencies, instructions that target the valence of the conditional or unconditional stimuli may effectively reduce the negative valence acquired by the conditional stimulus. Future research should examine whether instructions aimed at increasing the valence of the CS+ without any reference to the unconditional stimulus can affect the valence of the CS+. Consistent with this idea, past research on evaluative conditioning has shown that changing the affective value of an unconditional stimulus will change the affective valence of a CS+ that was associated with it (US re-valuation; Baeyens, Eelen, van den Bergh, & Crombez, 1992).

Like a number of previous studies, the current research combined verbal instruction with removal of the shock electrode to implement the instructed extinction manipulation. This was done to reduce the number of participants who did not believe the instructions but renders it impossible to attribute any change in conditional responding to the provision of verbal information alone. It speaks, however, to the robustness of the differential valence evaluations which were maintained even though presentation of further unconditional stimuli was impossible. Future research should examine whether the presence of the electrode influences the effect of the verbal manipulation as one could argue that it increases participants' arousal. We would predict that retaining the stimulus electrode will not alter the effect of instructed extinction on conditional stimulus valence, but may influence the physiological indices of fear learning.

Past research has shown that physiological indices of emotion are critically dependent on emotional arousal (Lang, Greenwald, Bradley, & Hamm, 1993; Bradley et al. 2001). Electrodermal responses are enhanced to arousing emotional stimuli, regardless of valence and the affect startle effect, startle facilitation during unpleasant stimuli and inhibition during pleasant stimuli, is observed if the stimuli are arousing, but not if they are low in arousal (Cuthbert, Bradley, & Lang, 1996). Thus, it may be that verbal instruction and removal of the shock electrode reduced arousal sufficiently to eliminate differential physiological responses while leaving self-reported valence unaffected. It should be noted that no evidence in support of this explanation was found when analyzing the tonic level of electrodermal activity one second prior to conditional stimulus onset. Instructed extinction and removal of the shock electrode did not differentially affect this index of general arousal, however, it may be that the manipulation did affect stimulus specific arousal rather than general arousal levels.

The arousal explanation offered above can be assessed utilizing an instructed counter-conditioning procedure. Rather than advising participants that no more unconditional stimuli will be presented, counter-conditioning involves the instruction that from now on the unconditional stimulus will be presented after the CS-. This manipulation should maintain the general level of arousal as well as the arousal level associated with one of the conditional stimuli. Extrapolating from the current results, we would predict that after instructed counter-conditioning electrodermal responses and fear potentiated startle will be enhanced during the CS-, whereas the CS+ would retain its negative valence and counter-conditioning trials would be required to alter this.

Regardless of the outcome of the future studies described above, the current results have significant practical implications. They suggest that even in the analogue procedure implemented in the laboratory, physiological indices of fear learning respond well to cognitive interventions but that negative valence towards a feared stimulus is durable and may resist cognitive intervention. As suggested by Kerkhof et al. (2012) this residual negative valence may play a critical role in the return of fear after treatment. To elaborate – it may be that after successful treatment for an anxiety disorder, the negative conditional stimulus valence comes to the fore again once a client is placed in a high arousal situation or faced with isolated presentations of aversive stimuli. It may well be that persistent negative valence provides a pathway for the return of fear.

References

Baeyens, F., Eelen, P., van den Bergh, O., & Crombez, G. (1992). The content of learning in human evaluative conditioning: Acquired valence is sensitive to US-revaluation. *Learning and Motivation*, *23*, 200-224

Blumenthal, T. D., Cuthbert, B.N., Filion, D.L., Hackley, S., Lipp O.V., & Boxtel, A.V. (2005). Committee report: Guidelines for human startle electromyographic studies. *Psychophysiology, 42*, 1-15. Doi: 10.1111/j.1469-8986.2005.00271.x

Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry, 52*, 976-986. doi: 10.1016/S0006-3223(02)01546-9

Bisson, J., & Andrew, M. (2007). Psychological treatment of post-traumatic stress disorder (PTSD). *Cochrane Database of Systematic Reviews*, 2007 (3), doi: 10.1002/14651858.CD003388.pub3.

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W.T., Dawson, M.E., & Filion, D. L. (2012). Publication recommendations for electrodermal measures. *Psychophysiology, 49*, 1017-1034. doi:10.1111/j.1469-8986.2012.01384.x

Bouton, M. E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry, 52*, 976-986. doi: 10.1016/S0006-3223(02)01546-9

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion, 1*(3), 276–298. doi:10.1037//1528-3542.1.3.276

Cacioppo, J.T., Marshall-Goodell, B.S., Tassinary, L.G., & Petty, R.E. (1992). Rudimentary determinants of attitudes: Classical conditioning is more effective when prior knowledge about the attitude stimulus is low than high. *Journal of Experimental Social Psychology,* *28*, 207-233.

Chen, M., & Bargh, J. A. (1999). Consequences of Automatic Evaluation: Immediate Behavioral Predispositions to Approach or Avoid the Stimulus. *Personality and Social Psychology Bulletin*, *25*(2), 215–224. doi:10.1177/0146167299025002007

Craske, M. G. (1999). Anxiety disorder: Psychological approaches to theory and treatment, Boulder, CO: Westview Press.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L.G. Tassinary & G.G. Bernston (Eds.), (2007). Handbook of Psychophysiology (pp. 159-181). Cambridge: Cambridge University Press.

M. G. Craske, D. Hermans & D. Vansteenwegen (Eds.), (2006). Fear and learning: From basic processes to clinical implications. Washington: APA Books.

Cuthbert, B.N., Bradley, M.M., & Lang, P.J. (1996). Probing picture perception: Activation and emotion. *Psychophysiology, 33*, 103-111.

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*(6), 853–869. doi:10.1037//0033-2909.127.6.853

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers, 35*, 116-124.

Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, *44*(4), 312–325. doi:10.1016/j.lmot.2013.03.003

Hamm, a O., Vaitl, D., & Lang, P. J. (1989). Fear conditioning, meaning, and belongingness: a selective association analysis. *Journal of Abnormal Psychology*, *98*(4), 395–406.

Hermans, D., Dirikx, T., Vansteenwegen, D., Vansteenwegenin, D., Baeyens, F., Van den Bergh, O., & Eelen, P. (2005). Reinstatement of fear responses in human aversive conditioning. *Behaviour Research and Therapy*, *43*(4), 533–51. doi:10.1016/j.brat.2004.03.013

Hugdahl, K. (1978). Electrodermal conditioning to potentially phobic stimuli: effects of instructed extinction. *Behaviour Research and Therapy*, *16*, 315–321. doi:10.1016/0005-7967(78)90001-3

Hugdahl, K., & Öhman, A. (1977). Effects of instruction on acquisition and extinction of electrodermal responses to fear-relevant stimuli. *Journal of Experimental Psychology. Human Learning and Memory*, *3*, 608–18. doi:10.1037/0278-7393.3.5.608

Kerkhof, I., Vansteenwegen, D., Beckers, T., Dirikx, T., Baeyens, F., D'Hooge, R., & Hermans, D. (2012). The role of negative affective valence in return of fear. In A. D. Gervaise (Ed), (2012). Psychology of fear: New research (pp. 153-170). New York: Nova Science Publishers.

Kessler, R. C., Koretz, D., Merikangas, K. R., & Wang, P.S. (2004). The epidemiology of adult mental disorders. In B.L. Levin, J. Petrilia, & K.D. Hennessy (Eds.), (2004). Mental health services: A public health perspective. New York: Oxford University Press.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention and the startle reflex. *Psychological Review*, *97*, 377-395.

Lang, P. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, 50, 372-385.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology, 30,* 261-273. doi: 10.1111/j.1469-8986.1993.tb03352.x

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.

Lipp, O.V., & Edwards, M.S. (2002). Effect of instructed extinction on verbal and autonomic indices of Pavlovian learning with fear-relevant and fear-irrelevant conditional stimuli. *Journal of Psychophysiology, 16,* 176-186. doi: 10.1027//0269-8803.16.3.176

Lipp, O. V., Oughton, N., & LeLievre, J. (2003). Evaluative learning in human Pavlovian conditioning: Extinct, but still there? *Learning and Motivation, 34,* 219-239. doi: doi:10.1016/S0023-9690(03)00011-0

Lipp, O.V., Mallan, K.M., Libera, M., & Tan, M. (2010). The effects of verbal instruction of affective and expectancy learning. *Behaviour Research and Therapy, 48,* 203-209. doi: 10.1016/j.brat.2009.11.002

Lockhart, R. A. (1966). Comments regarding multiple response phenomena in long interstimulus interval conditioning. *Psychophysiology, 3,* 108-114. doi: 10.1111/j.1469-8986.1966.tb02687.x

Mallan, K.M., Lipp, O.V., & Cochrane, B. (2013). Slithering snakes, angry men and out-group members: What and whom are we evolved to fear? *Cognition and emotion, 27,* 1168-1180. doi:10.1080/02699931.2013.778195

Mallan, K.M., Sax, J., & Lipp, O.V. (2009). Verbal instruction abolishes fear conditioned to racial out-group faces. *Journal of Experimental Social Psychology*, *45*, 1303-1307. doi:10.1016/j.jesp.2009.08.001

Mitchell, C.J., Anderson, N.E., & Lovibond, P.F. (2003). Measuring evaluative conditioning using the implicit association test. *Learning and Motivation, 34*, 203-217. doi:10.1016/S0023-9690(03)00003-1

Mowrer, O. H. (1938). Preparatory set (expectancy)--a determinant in motivation and learning. *Psychological Review*, *45*(1), 62–91. doi:10.1037/h0060829

Öhman, A. (1983). The orienting response during Pavlovian conditioning. In D. A. T. Siddle (Ed.), *Orienting and habituation: Perspectives in human research* (pp. 315-370). New York: Wiley.

Ougrin, D. (2011). Efficacy of exposure versus cognitive therapy in anxiety disorders: systematic review and meta-analysis. *BMC Psychiatry*, *11*(1), 1-12. doi:10.1186/1471-244X-11-200

Prokasy, W. F., & Ebel, H.C. (1967). Three components of the classically conditioned GSR in human subjects. Journal of Experimental Psychology, *73*, 247-256. doi:10.1037/h0024108

Prokasy, W.F., & Kumpfer, K.L. (1973). Classical conditioning. In W. F. Prokasy & D. C. Raskin (Eds.), *Electrodermal activity in psychological research* (pp. 157-202). San Diego: Academic Press.

Rachman, S. (1966). Studies in desensitization III: speed of generalization. Behaviour research and therapy, *4*, 7-15. doi:10.1016/0005-7967(66)90038-6

Rowles, M. E., Lipp, O.V., & Mallan, K.M. (2012). On the resistance to extinction of fear conditioned to angry faces. *Psychophysiology, 49*, 375-380. doi:10.1111/j.1469-8986.2011.01308.x

Sánchez-Meca, J., Rosa-Alcàzar, A., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: A meta-analysis. *Clinical Psychology Review*, *30*, 37-50. doi:10.1016/j.cpr.2009.08.011

Stewart, M.A., Winokur, G., Stern, J.A., Guze, S.B., Pfeiffer, E., & Hornung, F. (1959). Adaptation and conditioning of the galvanic skin response in psychiatric patients. *The British Journal of Psychiatry, 105*, 1102-1111. doi:10.1192/bjp.105.441.1102

Sevenster, D., Beckers, T., & Kindt, M. (2012). Instructed extinction differentially affects the emotional and cognitive expression of associative fear memory. *Psychophysiology*, *49*(10), 1426–35. doi:10.1111/j.1469-8986.2012.01450.x

Tottenham, N., Tanaka, J.W., Leon, A.C., McCarry, T., Nurse, M., Hare, T. A., . . . Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*, *168*, 242–249. doi:10.1016/j.psychres.2008.05.006

Vansteenwegen, D., Dirikx, T., Hermans, D., Verwliet, B., & Eelen, P. (2006). Renewal and reinstatement of fear: Evidence from human conditioning research. In M. G. Craske, D. Hermans & D. Vansteenwegen (Eds.), (2006). Fear and learning: From basic processes to clinical implications (pp. 197-215). Washington: APA Books.

Vervliet, B., Craske, M., & Hermans, D. (2013). Fear extinction and relapse: State of the Art. *Annual Review of Clinical Psychology*, *9*, 215-48. doi:10.1146/annurev-clinpsy-050212-185542

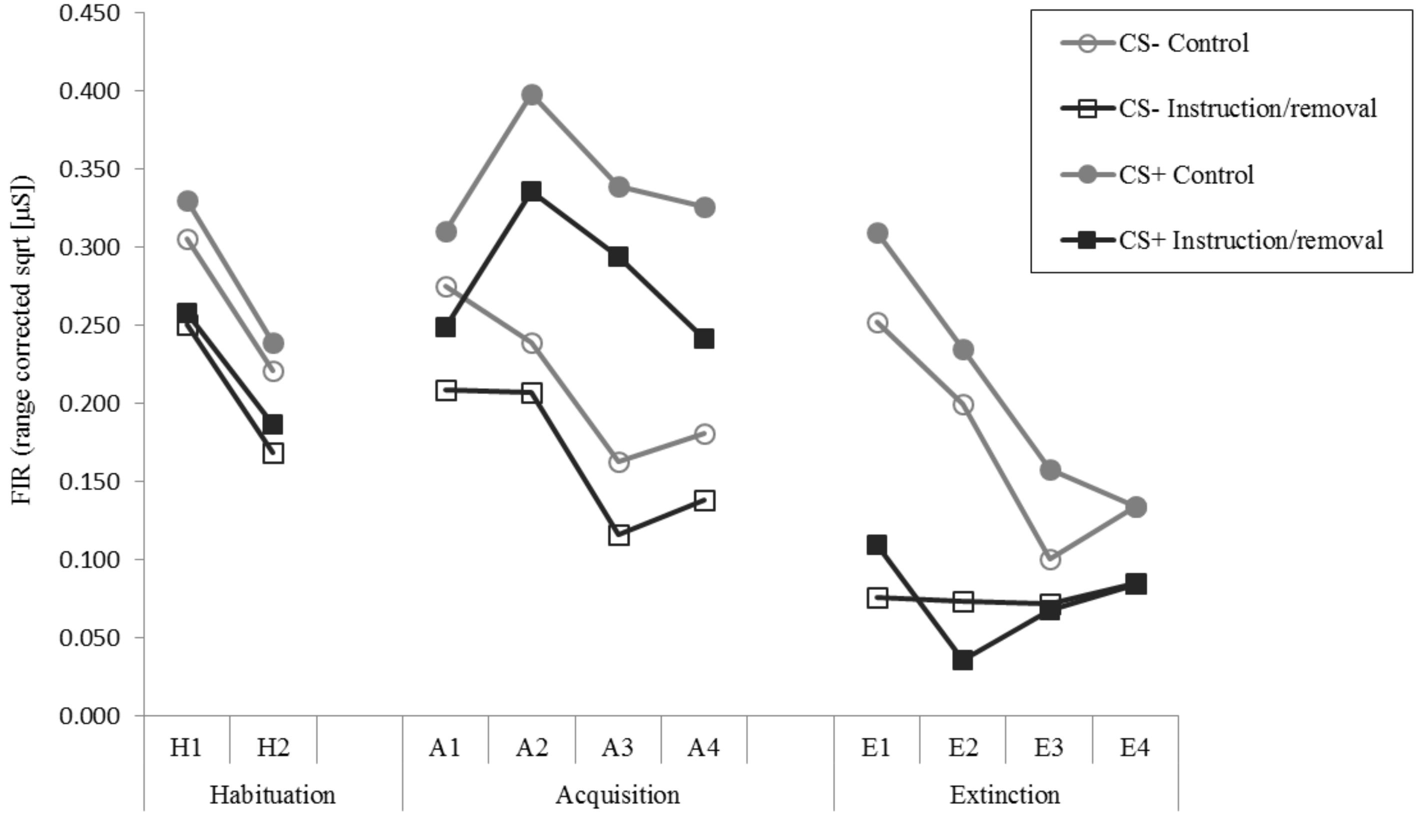

Figure 1. Mean electrodermal FIRs, for instruction/removal and control groups during habituation, acquisition, and extinction.

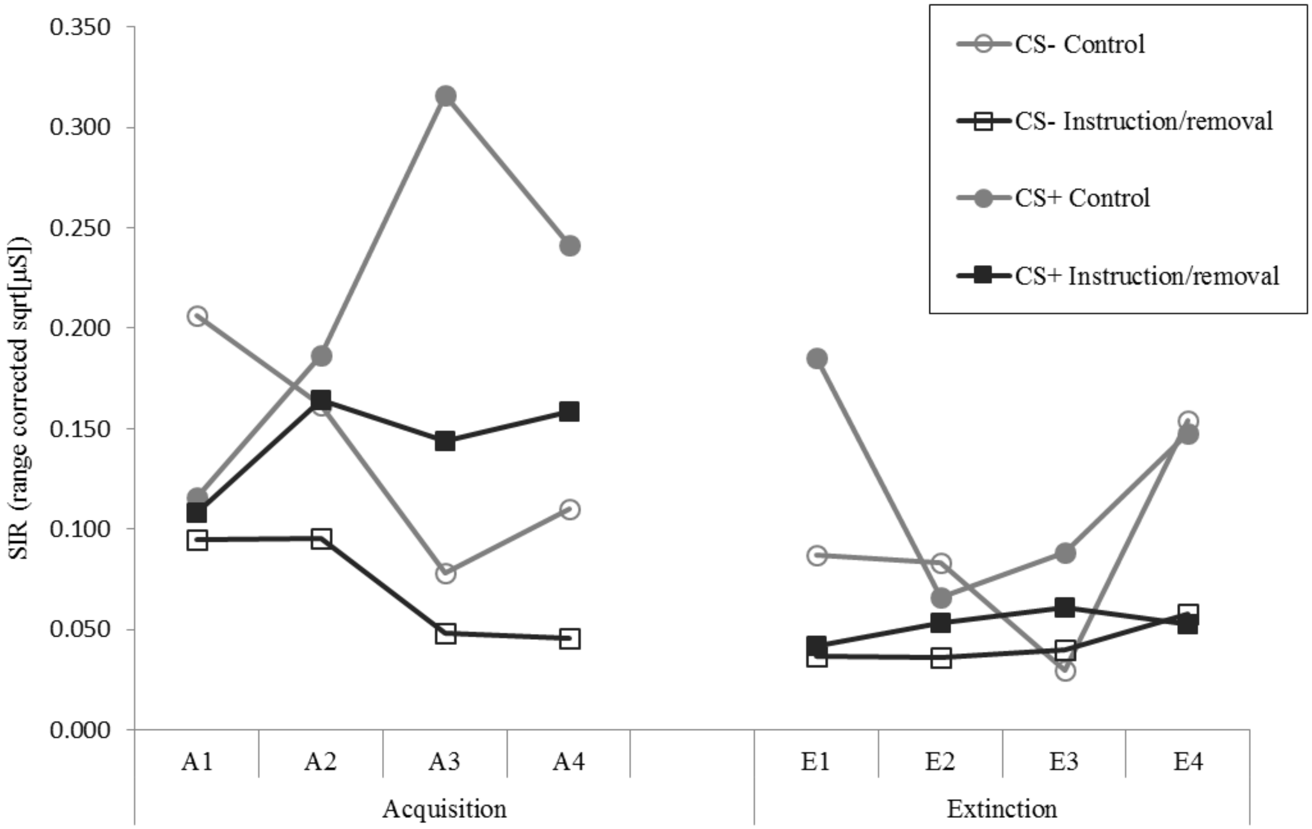

Figure 2. Mean electrodermal SIRs, for instruction/removal and control groups during acquisition, and extinction.

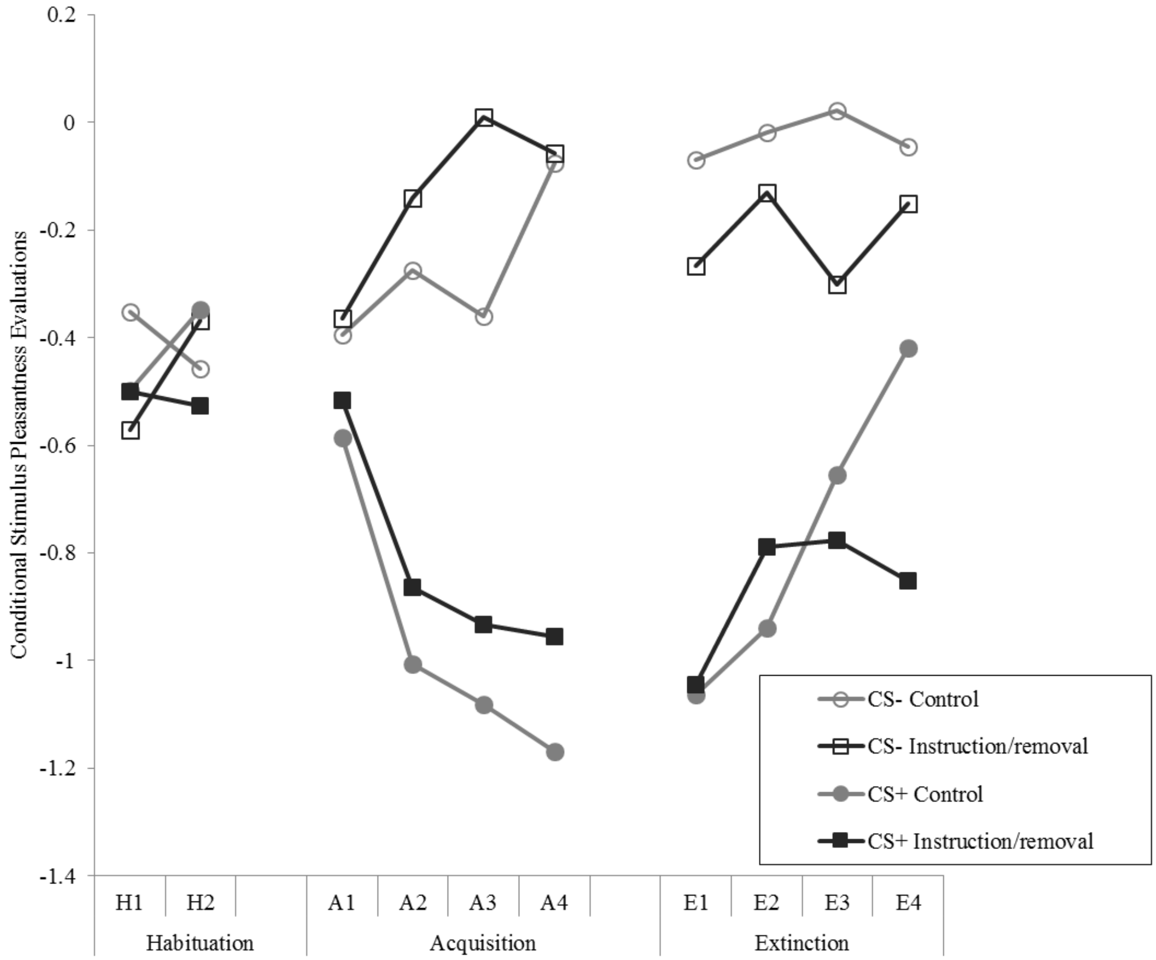

Figure 3. Conditional stimulus evaluations, for instruction/removal and control groups during habituation, acquisition, and extinction.

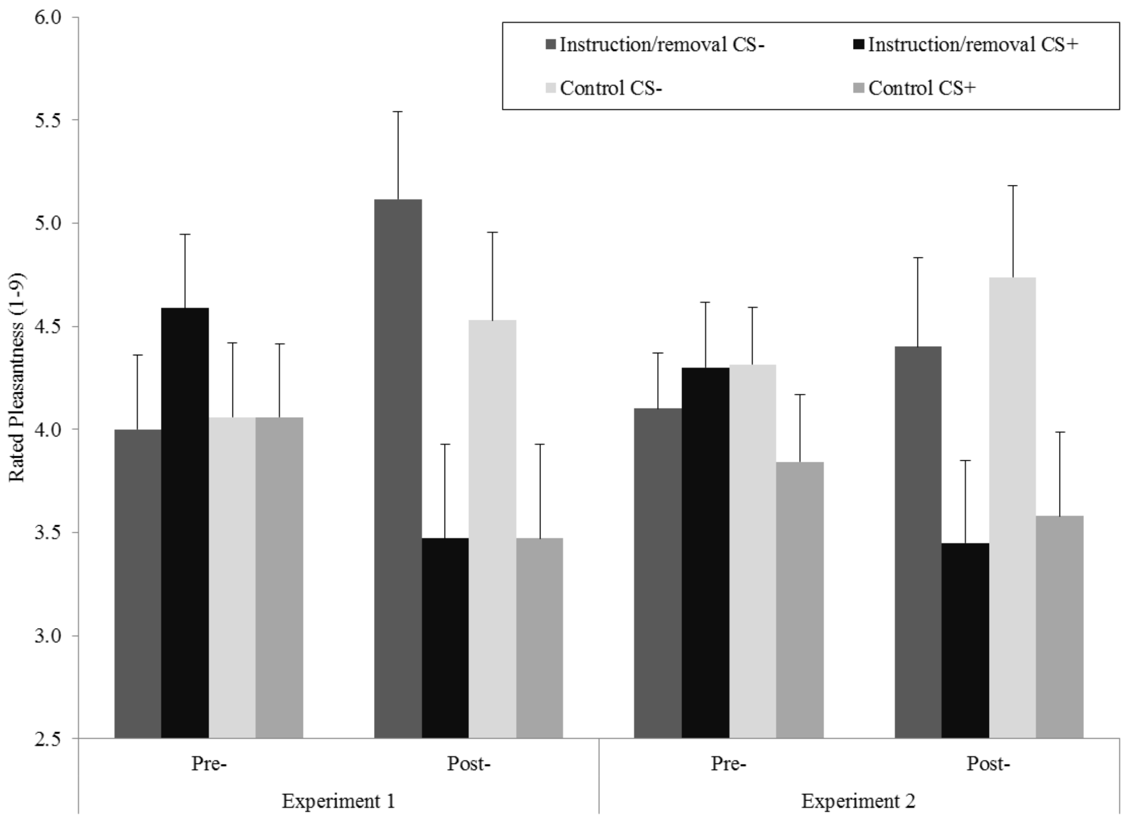

Figure 4. Pleasantness ratings collected pre- and post-experimentally for instruction/removal and control groups in Experiment 1 and 2 (error bars indicate standard errors of the mean).

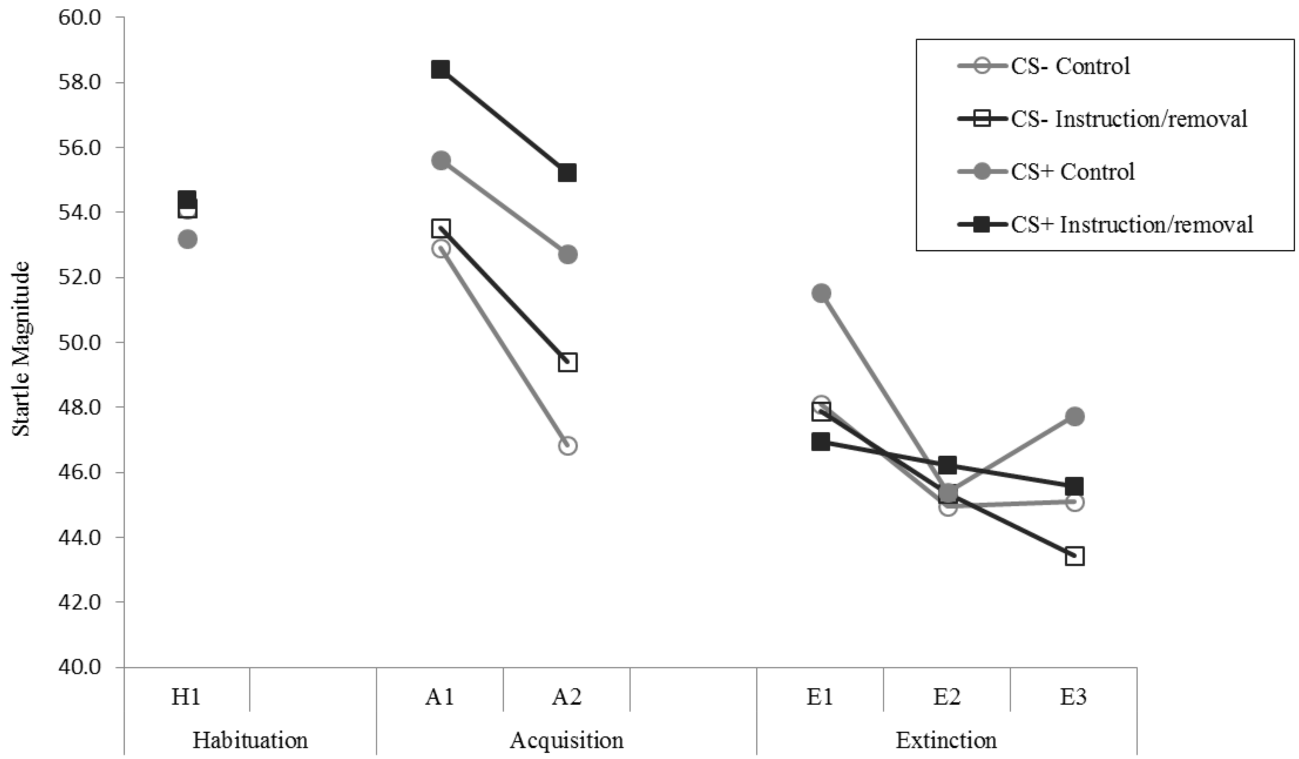

Figure 5. Startle magnitude elicited during habituation, acquisition, and extinction for instruction/removal and control groups.

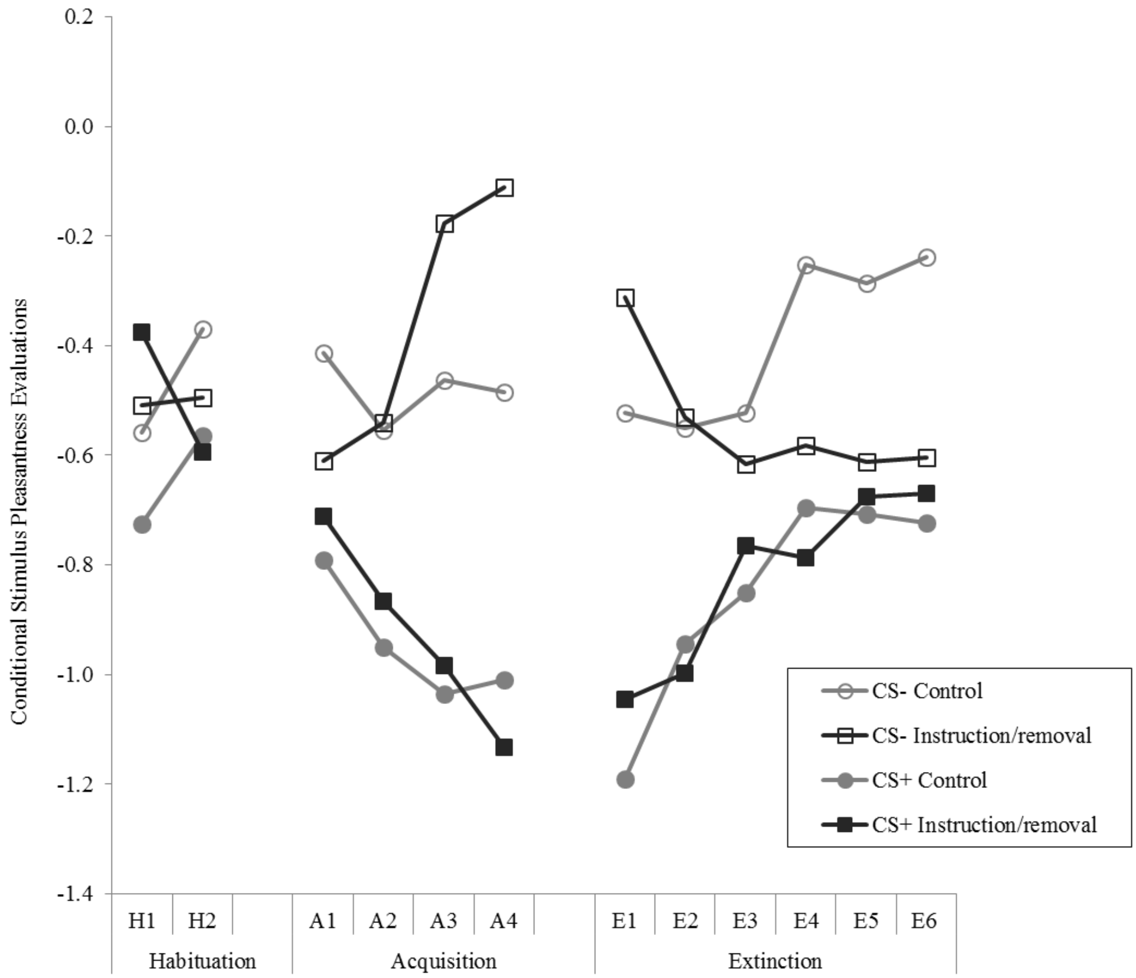

Figure 6. Conditional stimulus evaluations for instruction/removal and control groups during habituation, acquisition, and extinction.

Table 1

Response predictions expressed as a percentage of the total participants.

	Before instructional manipulation (acquisition)				After instructional manipulation (first trial of extinction)			
	Physiological		Evaluations		Physiological		Evaluations	
Response	CS+	CS-	CS+	CS-	CS+	CS-	CS+	CS-
Increase	98%	14%	2%	64%	14%	0%	61%	7%
Decrease	0%	61%	98%	31%	13%	2%	6%	5%
No Change	2%	25%	0%	5%	73%	98%	33%	88%

Appendix 1. Demand Characteristics Questionnaire

Please read the following description carefully and answer the questions:

An experimenter is conducting a fear learning experiment looking at how associations are formed between different stimuli. The participant views repeated presentations of two different faces throughout the experiment. On each presentation one of them is followed by an unpleasant (but not painful) electric stimulus, and the other is presented alone. As a measure of fear, physiological responses (e.g. heart rate) to the faces are recorded throughout the experiment. The participant is also required to rate how they feel about the faces every time they are shown on the screen (i.e. whether they perceive the face as pleasant, unpleasant or neutral).

1. How do you think the **physiological responses** to the face *paired with the electric stimulus* will develop across the experiment*?*
2. How do you think the **physiological responses** to the *face presented alone* will develop across the experiment?
3. What do you think will happen to the participants' **pleasantness ratings** to the face *paired with the electric stimulus* throughout the experiment?
4. What do you think will happen to the participants' **pleasantness ratings** to the face *presented alone* throughout the experiment?

Halfway through the experiment, the experimenter informs the participant that the electric stimulus will no longer be presented, but that they will continue to view and rate the same two faces for the remainder of the experiment.

5. The first time the participants view the face that was previously *paired with the electric stimulus* after receiving the instructions, do you think their **physiological responses** will change?

6. The first time the participants view the face that was previously *presented alone* after receiving the instructions, do you think their **physiological responses** will change?
7. The first time the participants view the face that was previously *paired with the electric stimulus* after receiving the instructions, what do you think will happen to the **pleasantness rating** of the face?
8. The first time the participants view the face that was previously *presented alone* after receiving the instructions, what do you think will happen to the **pleasantness rating** of the face?