The influence of contingency reversal instructions on electrodermal responding and conditional stimulus valence evaluations during differential fear conditioning

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: c.luck@curtin.edu.au

Acknowledgements

This work was supported by grants number DP120100750 and SR120300015 from the Australian Research Council.

Word count: 7069

Abstract

In differential fear conditioning, the instruction that the conditional stimulus (CS) will no longer be followed by the unconditional stimulus (US; instructed extinction) reduces differential physiological responding (expectancy learning) but leaves differential CS valence evaluations (evaluative learning) intact. This dissociation suggests that expectancy, but not evaluative learning, responds to contingency instructions. Alternatively, as instructed extinction removes the threat of receiving the US, this dissociation could be caused by a drop in participants' arousal levels which could render the physiological indices of fear learning less sensitive. To test this alternative explanation, we examined the impact of an instructed reversal manipulation on electrodermal responding and CS valence evaluations. After instructed reversal, electrodermal responses to CS+ decreased and electrodermal responses to CS- increased, in the instruction, but not in the control group. In addition, there was some evidence for an instruction dependent change in CS valence, however, this finding seems limited to changes in CS+ valence and possible explanations for this finding are discussed. Overall, the study confirms that the dissociation detected in instructed extinction studies is unlikely to be caused by a drop in the participants' arousal levels.

Key words: fear conditioning, instructed reversal, instructed extinction, evaluative learning, expectancy learning, conditional stimulus valence, electrodermal responding.

During classical fear conditioning, a neutral conditional stimulus (CS) is paired with an aversive unconditional stimulus (US). After repeated pairings, the CS generates an expectation that the US will occur (Lipp, 2006) and acquires negative valence (De Houwer, Thomas, & Baeyens, 2001). Dissociations between the predictive (expectancy) and the emotional (evaluative) components of human fear learning have been reported in response to instructed extinction (see Luck & Lipp, 2015a), generating debate about whether these components reflect different underlying mechanisms or operate under different boundary conditions.

Understanding the mechanisms underlying expectancy and evaluative learning is important from a number of viewpoints. Residual negative valence has been associated with higher relapse rates after fear extinction, and prior research suggests that CS valence may resist current fear and anxiety treatments (Hermans et al., 2005; Luck & Lipp, 2015a; Zbozinek, Hermans, Prenoveau, Liao, & Craske, 2015). From a theoretical perspective, there is some debate about whether Pavlovian conditioning can be considered the result of propositional processes alone or whether both propositional and associative processes co-occur during Pavlovian conditioning. According to single-process propositional theories, Pavlovian conditioning is the result of the formation and truth evaluation of non-automatic propositions regarding the CS-US relationship. Dual-process theories propose that automatic associations between CS and US representations also develop during CS-US pairings (see De Houwer, 2009 for a review and discussion of these theories). Some theories (see Baeyens, Eelen, Crombez, & Van den Bergh, 1992) propose that evaluative and expectancy learning are two different types of Pavlovian conditioning, both based on the formation of stimulus representations in memory. According to these theories, expectancy learning concerns the learning of predictive relationships in which the CS becomes a signal that the US will occur, whereas, evaluative learning concerns

the learning of referential relationships, in which the CS becomes a stimulus which activates the mental representation of the US without generating an expectancy that the US will occur.

Dissociations between evaluative and expectancy learning in response to the same experimental manipulation could hold the key to understanding whether or not they have the same underlying mechanism. Expectancy and evaluative learning can be examined simultaneously using a differential fear conditioning paradigm. In this paradigm, one CS, the CS+, is repeatedly paired with the US, and another, the CS-, is presented alone. Electrodermal responding, a physiological index which is very sensitive to the CS-US contingency, and CS valence evaluations are frequently collected as dependent measures, and both can be measured continuously throughout conditioning. Differential electrodermal responding and differential valence evaluations develop across training trials, such that CS+ elicits larger electrodermal responding and is rated as less pleasant than CS-. During extinction, CS+ and CS- are both presented alone and eventually the differential electrodermal responding and valence evaluations reduce and return to baseline levels. Using this paradigm, Luck and Lipp (2015a; 2015b) reported that instructed extinction, a manipulation which involves informing participants prior to the extinction phase that the US will no longer occur, results in the immediate elimination of differential electrodermal responding (and fear-potentiated startle), but leaves differential valence evaluations intact. These results can be interpreted to indicate that expectancy learning responds to the instructed CS+– noUS contingency immediately, but that evaluative learning continues to reflect the valence acquired during acquisition, requiring further Pavlovian training to reduce the negative CS+ valence. This interpretation is consistent with literature examining US expectancy and CS evaluation in picture-picture evaluative conditioning paradigms (Lipp, Mallan, Libera, & Tan, 2010). Alternatively, the elimination of differential physiological

responding after instructed extinction could occur because participants' general arousal level is reduced after being informed that they will not receive US presentations anymore. Electrodermal responding is also sensitive to stimulus valence but only under conditions of high arousal (Bradley, Codispoti, Cuthbert, & Lang, 2001). As CS evaluations are not sensitive to the overall level of arousal, the dissociation between physiological and evaluative indices of fear learning could reflect the differential sensitivity of electrodermal responding and CS evaluations to changes in arousal.

An instructed reversal manipulation (Grings, Schell, & Carey, 1973) involves informing participants after acquisition training, that the contingencies will switch, such that CS+ will no longer be followed by the US, but that the US will now be presented after the CS-. This manipulation is unlikely to cause a drop in participants' overall arousal because of the ongoing threat of receiving the US and therefore provides a test of the arousal account described above. While instructed extinction involves examining safety instructions to the CS+, instructed reversal allows for the examination of both safety instructions to the CS+ and danger instructions to the CS-, providing a more comprehensive examination of the effects of instructions.

Effects of the instructional manipulation can be examined across the entire reversal phase or on the very first trial after the instruction was provided. Although differences between the instruction and control groups may be observed in both cases, the two assessments can indicate different processes. Instruction effects detected across the entire reversal phase could indicate that instructions facilitate learning of the new contingency (Instruction × Training interaction) and not necessarily a reversal change caused by the instructions alone. Differences on the first reversal trial, however, can be considered the effects of the instructional manipulation alone and provide for the strongest test of the instructed reversal manipulation. The nature of the first trial

(CS+/CS-) presented after instruction should also be controlled because experiencing a contingency change on the first reversal trial (i.e. unreinforced CS+ or reinforced CS-) could lead participants to infer that the experimental contingencies have changed.

Using a differential fear conditioning paradigm, we examined whether electrodermal responding and trial-by-trial CS valence would respond to an instructed reversal manipulation. To be able to examine the effects of instructed reversal without any influence of additional learning (or inference), half of the participants received a CS+ as the first reversal trial and the others received a CS- as the first reversal trial. We hypothesized, based on the results of Luck and Lipp (2015a; 2015b), that electrodermal responding to CS+ would decrease and that electrodermal responding to CS- would increase on the first reversal trial in the instruction group but not in the control group. It was further hypothesized that CS valence would not be affected in either group.

Method

Participants

One hundred and forty-nine undergraduate students (95 female), aged between 17 – 43 years (M = 23.16) provided informed consent and volunteered participation in exchange for course credit or monetary compensation of AU$15. Participants were assigned to different CS order conditions[1] and then were randomly assigned to the control or instruction group. Twenty participants failed to correctly verbalize the experimental contingencies and were removed from the analyses. An additional 7 participants reported that they did not believe the reversal instructions and were removed from the reversal and instruction analyses. Five participants' electrodermal responses and two participants' conditional stimulus (CS) valence evaluations

[1] Two experiments were conducted which were identical except for which CS was presented first during the reversal phase. To streamline the report, we have combined the experiments and added the factor CS order to the analyses.

were lost due to problems with the recording device, and five participants did not provide complete before and after rating datasets. These participants have been included in the analyses of the remaining measures.

Apparatus/Stimuli

The CSs were 4 pictures of Caucasian, male adults [NimStim database: images M_NE_C: models 20, 21, 32, 31, Tottenham et al. (2009)] displaying neutral facial expressions. The pictures were presented on a 17-inch color LCD screen for 6 s. A pseudorandom trial sequence was used, such that a CS+/CS- was not presented more than twice consecutively. Counterbalancing was performed between participants, varying the nature of the first trial during acquisition (CS+/CS-), the face used as CS+/CS-, and the two faces used in the experiment. The unconditional stimulus (US) was a 200 ms electrotactile stimulus pulsed at 50 Hz and delivered by a Grass SD9 stimulator to the participants' preferred forearm. Physiological responding and CS evaluations 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 CS evaluations were measured on a trial-by-trial basis using an evaluation joystick with the anchors 'very unpleasant', 'neutral', and 'very pleasant'. DMDX 3.0.2.8 software (Forster & Forster, 2003) was used to control the stimulus presentation and timing and to record the pleasantness ratings (Ratings A and B).

Procedure

Participants washed their hands, provided informed consent, and were seated in front of a monitor in a separate room adjacent to the control room. The respiratory effort transducer was fitted around their waist, and the electrodermal electrodes were attached to the thenar and hypothenar prominences of their non-dominant hand. The shock electrode was attached to their

dominant forearm, and a shock-work up procedure was performed to set the US intensity to a level that was experienced as subjectively ‘unpleasant, but not painful’. Participants were then asked to relax and watch the blank computer screen while a 3-min baseline of their electrodermal activity (EDA) was recorded. After the baseline recording, participants rated the CS faces on a 1 to 9 (1= unpleasant, 9=pleasant) Likert scale (ratings A) and were informed that they would see the faces displayed on the screen throughout the experiment. They were asked to use the evaluation joystick throughout the experiment to indicate how pleasant/unpleasant they found each face, and to make this evaluation as soon as the face was presented on the screen with their preferred hand – ensuring that the movement did not interfere with the electrodermal recording and that the presence/absence of the US, on a given trial, did not influence the evaluations.

After the participant confirmed that they understood what was required, the conditioning task, consisting of habituation, acquisition, and reversal phases, was started. During habituation, both CS+ and CS- were presented 4 times alone. During acquisition, the CS+ was presented 8 times, with the offset of the CS+ coinciding with the onset of the US in a 100% reinforcement schedule, while the CS- was presented 8 times alone. During habituation and acquisition, CS+ and CS- were presented in a pseudorandom sequence with the restrictions that the first 2 stimuli in a phase were a CS+ and a CS- and that no more than 2 consecutive stimuli were the same. After acquisition, the experimenter entered the participants’ room and informed them that the mid-point of the experiment had been reached and that the electrodes needed to be checked, before appearing to visually inspect the electrodermal electrodes. Participants in the control group did not receive information about the CS-US contingency. Participants in the instruction group were informed that in the second part of the experiment the electrotactile stimulus would no longer be presented after the stimulus it had previously followed, but would switch to follow

the other stimulus. Participants were asked to confirm they understood the instructions and told the experiment would continue. During the reversal phase, the CS+ (CS terminology from acquisition will be used consistently throughout both phases) was presented 8 times alone, and the CS- was presented 8 times with the offset of the CS- coinciding with the onset of the US in a 100% reinforcement schedule. The first 3 trials of the reversal phase differed depending on CS order group. Participants in the CS+ first group viewed 2 consecutive presentations of the CS+, followed by a CS- and then the counterbalanced pseudorandom trial sequence. Participants in the CS- first group viewed 2 consecutive presentations of the CS-, followed by a CS+ and then the counterbalanced pseudorandom trial sequence. Inter-trial intervals lasted 11s, 13s, or 15s from CS offset to CS onset and were randomly varied throughout the experiment. After the last reversal trial, participants completed another rating task (ratings B), which was identical to the one performed before conditioning, the electrodes were removed and the participant was led into the control room for the post-experimental questionnaire. The questionnaire required participants to identify which faces were presented in the experiment and which face was followed by the electrotactile stimulus in the first and second part of the experiment. As a manipulation check, participants were asked to indicate whether they believed the instructions (instruction group only; yes or no question). Participants then rated the pleasantness of the electrotactile stimulus and the CS faces on a (-3 [*very unpleasant*] to +3 [*very pleasant*]) pleasantness scale (ratings C), before being debriefed and thanked.

Scoring and Response Definition

Electrodermal responding was scored in multiple latency windows as recommended by Prokasy and Kumpfer (1973) and Luck and Lipp (2016). First interval responding was defined as responses starting within 1-4 s of CS onset and second interval responding was defined as

responses starting within 4-7 s of CS onset. The largest response starting within the latency window was scored and the response magnitude was calculated as the difference between response onset and peak (Prokasy & Kumpfer, 1973). The electrodermal responses were square root transformed to reduce the positive skew of the distribution (Dawson, Schell, & Filion, 2007) and then range corrected (using the largest response as a reference) to reduce the effect of individual differences in response size (Boucsein et al., 2012; Dawson et al., 2007). During habituation only first interval responses were scored as they reflect orienting to novel stimuli (Öhman, 1973). As a measure of spontaneous EDA, any discernible response displayed during the baseline period was counted (Dawson et al., 2007). The CS valence ratings provided with the response joystick were recorded by the Biopac MP150 system as voltage deviations. The joystick was spring loaded, such that after a response was made the joystick would return to the 'neutral' position. The valence ratings made during the 6 s CS presentation were scored as the largest voltage deviation from mean baseline voltage recorded 1 s prior to CS onset. To reduce the influence of trial by trial variability, electrodermal responding and CS valence evaluations were averaged into blocks of 2 consecutive trials[2]. All analyses were conducted with IBM SPSS Statistics 22 with a significance level of .05, and Pillai's trace statistics have been reported.

Results

Preliminary Analyses

Two Pearson's chi-square tests were performed to ensure that the gender ratio did not differ in the instruction or CS order groups. To check for baseline differences between the groups a series of 2 (Group: instruction, control) × 2 (CS order: CS+ first, CS- first) univariate ANOVAs were performed on age, spontaneous EDA, US intensity, and US valence. The means

[2] As the influence of the instructional manipulation is expected during the first reversal trial the analyses concerned with the instruction effect are based on single trials.

and standard deviations for these variables are displayed in Table 1. The instruction groups, $\chi^2(1) = .240$, $p = .624$, and CS order groups, $\chi^2(1) = .362$, $p = .547$, did not differ in gender ratio. The CS- first group was older than the CS+ first group, $F(1, 125) = 5.75$, $p = .018$, $\eta p^2 = .044$, and the CS+ first group set the US intensity marginally higher than the CS- first group, $F(1, 125) = 3.28$, $p = .073$, $\eta p^2 = .026$. No other comparisons reached significance, all F's < 2.71, p's $< .102$, ηp^2's $< .021$.

Habituation

The CS valence evaluations and first interval responding recorded during habituation (see left panels of Figures 1 and 2, respectively) were subjected to separate 2 (Group: instruction, control) × 2 (CS order: CS+ first, CS- first) × 2 (CS: CS+, CS-) × 2 (Block: 1, 2) mixed-model factorial ANOVAs.

Conditional Stimulus Valence. A CS × CS order interaction, $F(1, 123) = 4.12$, $p = .045$, $\eta p^2 = .032$, revealed that participants in the CS- first group evaluated CS+ as less pleasant than CS-, $F(1, 123) = 5.16$, $p = .025$, $\eta p^2 = .040$, whereas evaluations did not differ in the CS+ first group, $F(1, 123) = 0.40$, $p = .530$, $\eta p^2 = .003$.

First Interval Responding. Responding decreased from block 1 to block 2, $F(1, 121) = 61.50$, $p < .001$, $\eta p^2 = .337$, and responding was larger in the CS+ first group than in the CS- first group, $F(1, 121) = 5.65$, $p = .019$, $\eta p^2 = .045$.

Acquisition

The CS valence evaluations, first interval responding, and second interval responding recorded during acquisition were subjected to separate 2 (Group: instruction, control) × 2 (CS order: CS+ first, CS- first) × 2 (CS: CS+, CS-) × 4 (Block: 1, 2, 3, 4) mixed model factorial

ANOVAs and are presented in Figures 1 (middle panels), 2 (middle panels), and 3 (left panels), respectively.

Conditional Stimulus Valence. A main effect of CS, $F(1, 123) = 23.31$, $p < .001$, $\eta p^2 = .159$, and a CS × Block interaction, $F(3, 121) = 14.53$, $p < .001$, $\eta p^2 = .265$, were moderated by a CS × Block × Group interaction, $F(3, 121) = 3.48$, $p = .018$, $\eta p^2 = .079$. Differential valence was not present in either group during block 1 (*F's* $(1, 123) < 2.72$, *p's* $> .101$, ηp^2's $< .023$), however, during subsequent blocks CS+ was evaluated as less pleasant than CS- in both groups (all *F's* $(1, 123) < 4.90$, *p's* $> .028$, ηp^2's $< .037$). Although differential valence was present in both groups, valence evaluations to CS+ and CS- changed across blocks in the control groups, *F's* $(3, 121) > 5.58$, *p's* $< .002$, ηp^2's $> .121$, but not in the instruction groups, *F's* $(3, 121) < 2.21$, *p's* $> .090$, ηp^2's $> .053$.

First Interval Responding. Responses were larger in the CS+ first group than in the CS- first group, $F(1, 121) = 4.94$, $p = .028$, $\eta p^2 = .039$. A main effect of CS, $F(1, 121) = 60.38$, $p < .001$, $\eta p^2 = .333$, and a main effect of block, $F(3, 119) = 11.28$, $p < .001$, $\eta p^2 = .221$, were moderated by a CS × Block interaction, $F(3, 119) = 13.66$, $p < .001$, $\eta p^2 = .256$. Follow-up analyses revealed that responding to CS+ and CS- did not differ during block 1, $F(1, 121) = 0.52$, $p = .470$, $\eta p^2 = .004$, but during subsequent blocks responding to CS+ was larger than to CS-, all *F's* $(1, 121) > 24.27$, *p*'s $< .001$, ηp^2's $> .166$.

Second Interval Responding. A main effect of CS, $F(1, 121) = 42.33$, $p < .001$, $\eta p^2 = .259$, was moderated by a CS × Block interaction, $F(3, 119) = 9.07$, $p < .001$, $\eta p^2 = .186$. Follow-up analyses revealed that responding to CS+ and CS- did not differ during block 1, $F(1, 121) = 0.46$, $p = .497$, $\eta p^2 = .004$, but responding to CS+ was larger than to CS- during subsequent blocks, all *F's* $(1, 121) > 4.67$, *p's* $< .034$, ηp^2's $> .036$.

Reversal

The CS valence evaluations, first interval responding, and second interval responding recorded during reversal were subjected to separate 2 (Group: instruction, control) × 2 (CS order: CS+ first, CS- first) × 2 (CS: CS+, CS-) × 4 (Block: 1, 2, 3, 4) mixed-model factorial ANOVAs and can be seen in the right panels of Figures 1, 2, and 3, respectively.

Conditional Stimulus Valence. A main effect of CS, $F(1, 117) = 20.42$, $p < .001$, $\eta p^2 = .149$, was moderated by a CS × Group × CS order interaction, $F(1, 117) = 3.99$, $p = .048$, $\eta p^2 = .033$. If a CS+ was presented first, the instruction group evaluated CS- as less pleasant than CS+, $F(1, 117) = 9.18$, $p = .003$, $\eta p^2 = .073$, whereas evaluations did not differ in controls, $F(1, 117) = 2.38$, $p = .126$, $\eta p^2 = .020$. If a CS- was presented first, the instruction group did not evaluate CS+ and CS- differently, $F(1, 117) = 0.99$, $p = .321$, $\eta p^2 = .008$, but the control group evaluated CS- as less pleasant than CS+, $F(1, 117) = 12.08$, $p = .001$, $\eta p^2 = .094$. A CS order × Block interaction, $F(3, 115) = 3.46$, $p = .019$, $\eta p^2 = .083$, revealed when CS+ was presented first, overall evaluations did not differ across blocks, $F(3, 115) = 0.87$, $p = .461$, $\eta p^2 = .022$, but when CS- was presented first, evaluations in block 1 were more pleasant than evaluations in subsequent blocks, all p's $< .037$, $F(3, 115) = 4.31$, $p = .006$, $\eta p^2 = .101$. A CS × Block interaction, $F(3, 115) = 17.60$, $p < .001$, $\eta p^2 = .315$, revealed that differential evaluations were not present during the first reversal block, $F(1, 117) = 0.25$, $p = .616$, $\eta p^2 = .002$, but CS- was evaluated as less pleasant than CS+ during subsequent blocks, all F's$(1, 117) > 17.87$, p's $< .001$, ηp^2's $> .132$. The CS × Block × Group interaction approached significance, $F(3, 115) = 2.64$, $p = .053$, $\eta p^2 = .064$, but follow-up analyses revealed the same pattern of differential valence in both groups.

First Interval Responding. Main effects of CS, $F(1, 114) = 89.86$, $p < .001$, $\eta p^2 = .441$, and block, $F(3, 112) = 10.94$, $p < .001$, $\eta p^2 = .227$, and a CS × Block interaction, $F(3, 112) = 3.88$, $p = .011$, $\eta p^2 = .094$, were moderated by a CS × Block × Group interaction, $F(3, 112) = 3.67$, $p = .014$, $\eta p^2 = .089$. In the control group, responding between CS+ and CS- did not differ during block 1, $F(1, 114) = 0.13$, $p = .724$, $\eta p^2 = .001$, but during subsequent blocks responding to CS- was larger than responding to CS+, all *F's* $(1, 114) > 13.76$, *p's* $< .001$, ηp^2's $> .107$. In the instruction group, however, CS- elicited larger responding than CS+ during all blocks, block 1: $F(1, 114) = 32.05$, $p < .001$, $\eta p^2 = .219$, subsequent blocks: all *F's* $(1, 114) > 14.06$, *p's* $< .001$, ηp^2's $> .109$. A CS × Group × CS order interaction, $F(1, 114) = 6.39$, $p = .013$, $\eta p^2 = .053$, revealed that across reversal, responding to CS- was larger in the CS+ first instruction group in comparison with the CS+ first control group, $F(1, 114) = 4.62$, $p = .034$, $\eta p^2 = .039$; no other differences between the groups reached significance, all *F's* $(1, 114) < 0.12$, *p's* $> .745$, ηp^2's $< .002$.

Second Interval Responding. A main effect of CS, $F(1, 114) = 90.03$, $p < .001$, $\eta p^2 = .441$, was moderated by a CS × Block × Group interaction, $F(3, 112) = 5.79$, $p = .001$, $\eta p^2 = .134$. In both groups, CS- elicited larger responding than CS+ during all 4 blocks, all *F's* $(1, 114) > 3.97$, *p's* $< .049$, ηp^2's $> .033$; however, during block 1, responding to the CS+ was larger in the control group than in the instruction group, $F(1, 114) = 5.46$, $p = .021$, $\eta p^2 = .046$, and responding to the CS- was larger in the instruction group than in the control group, $F(1, 114) = 4.69$, $p = .033$, $\eta p^2 = .039$. During block 2, responding to the CS+ was marginally larger in the instruction group than in the control group, $F(1, 114) = 3.77$, $p = .055$, $\eta p^2 = .032$. The instruction and control group did not differ in responding to CS+ or CS- during any other stage of the reversal phase, all *F's* $(1, 114) < 0.70$, *p's* $> .403$, ηp^2's $< .007$.

First Trial Instruction Effects

In order to examine the effects of the instructions on responding to CS+ and CS- independent of any additional learning that may have occurred as a result of the initial reversal trial, a change score [first reversal trial – last acquisition trial] was calculated for evaluations of and electrodermal responses to CS+ in the CS+ first groups and CS- in the CS- first groups. To compare the magnitude of the instruction effects for CS+ (instructions should increase pleasantness and reduce electrodermal responses) and CS- (instructions should decrease pleasantness and increase electrodermal responses), the change scores in the CS- first group were inverted[3] and 2 (Group: instruction, control) × 2 (CS order: CS+ first, CS- first) between groups ANOVAs were performed and the 95% confidence intervals for the change scores were inspected. The (non-inverted) change scores for CS valence, first interval, and second interval responding are displayed in the left, middle, and right, panels of Figure 4, respectively.

Conditional Stimulus Valence. The 2 x 2 factorial ANOVA yielded no significant differences, largest $F(1, 117) = 2.66$, $p = .105$, $\eta p^2 = .022$ (Group × CS order interaction) indicating that the change in stimulus evaluations did not differ across the 4 groups. The change score for CS+ valence in the instruction group, however, was significantly different from 0 as suggested by the 95% confidence interval [0.178, 0.837]. This was not the case in the other groups 95% CI [Instruction CS-: -0.501, 0.103; Control CS+: -0.278, 0.336; Control CS-: -0.514, 0.062].

First Interval Responding. As can be seen in the middle panel of Figure 4, the change in first interval responding was larger in the instruction than in the control groups, $F(1, 114) = 4.39$,

[3] The signs for the CS- first group were inverted in order to remove the direction of the instruction effect (while still keeping individual variability). As some participant's instructions scores are positive others are negative taking the absolute values of the scores is not accurate as it does not take into account this variability. Inversing the score removes the direction while keeping the magnitude.

p = .038, ηp^2 = .037, and larger in the CS- first group than in the CS+ first group, $F(1, 114)$ = 9.50, p = .003, ηp^2 = .077. Inspection of the 95% confidence intervals suggests that the increase in first interval responding to CS- in the instruction group was significant [0.154, 0.357], whereas there was no difference in the three other groups 95% [Instruction CS+: -0.140, 0.082; Control CS+: -0.089, 0.124; Control CS-: -0.017, 0.179].

Second Interval Responding. The change in electrodermal second interval responding was larger in the instruction than in the control groups, $F(1, 114) = 8.33$, p = .005, ηp^2 = .068. Second interval responses to CS+ decreased in the instruction, 95% CI [-0.230, -0.050], but not the control group, 95% CI [-0.092, 0.081], whereas second interval responses to CS- increased in the instruction group, 95% CI [0.037, 0.201], but not in the control group, 95% CI [-0.072, 0.087].

Pre/Post Pleasantness Ratings

Before analysis, the post-experimental pleasantness ratings (ratings C) were transformed from a 7 to a 9 point Likert scale. Pleasantness evaluations taken before habituation (ratings A), after reversal (ratings B), and post-experimentally were subjected to a 2 (Group: instruction, control) × 2 (CS order: CS+ first, CS- first) × 2 (CS: CS+, CS-) × 3 (Phase: ratings A, ratings B, ratings C) factorial ANOVA, see Figure 5. A main effect of phase, $F(2, 120) = 7.38$, p = .001, ηp^2 = .109, was moderated by a CS × Phase interaction, $F(2, 120) = 11.27$, $p < .001$, ηp^2 = .158. Ratings of CS+ and CS- did not differ before habituation, $F(1, 121) = 0.11$, p = .746, ηp^2 = .001, however after reversal, CS- was given lower pleasantness ratings than CS+, $F(1, 121) = 15.07$, $p < .001$, ηp^2 = .111. After the experiment, ratings of CS+ and CS- did not differ, $F(1, 121) = 0.30$, p = .585, ηp^2 = .002.

Discussion

In the current study, we examined the effect of reversal instructions on electrodermal responding and online conditional stimulus (CS) valence evaluations after differential fear conditioning. Prior studies of instructed extinction have reported that instructions eliminate differential physiological responding, while leaving differential CS valence evaluations intact (Luck and Lipp, 2015a; 2015b). This dissociation could indicate that different mechanisms underlie expectancy learning and evaluative learning. Alternatively, it could occur because instructed extinction reduces arousal levels, rendering the physiological indices less sensitive to residual stimulus valence. An instructed reversal design permits the assessment of this proposition as the threat of receiving the unconditional stimulus (US), and therefore arousal, is maintained. Based on studies of instructed extinction we hypothesized that instructed reversal would reduce electrodermal responding to CS+, and increase electrodermal responding to CS-, in the instruction groups, but not the control groups. CS valence, however, was predicted to remain unchanged in both groups.

Throughout acquisition, differential first and second interval electrodermal responding was acquired, such that presentations of CS+ elicited larger responses than presentations of CS-. Differential valence evaluations were also acquired such that CS+ acquired negative valence relative to CS-. Reversal instructions affected electrodermal responses to CS+ and CS- as predicted. Analysis of the change in electrodermal responses from the last trial of acquisition to the first trial of reversal revealed that the instruction decreased electrodermal second interval responding to CS+ and increased electrodermal first and second interval responding to CS-. This change was evident on the very first trial of reversal, i.e., in the absence of any additional Pavlovian training. The finding that the instructed CS+ first group showed a decrease in electrodermal second interval responding to CS+, even though US presentations were expected

on subsequent trials, indicates that the elimination of differential electrodermal responding after instructed extinction is not caused by a decrease in arousal levels.

While significant changes in second interval responding in response to instructed reversal were observed in both CS order groups, a change in first interval responding was significant only in the CS- first group. The absence of significant instruction effects in electrodermal first interval responding is not uncommon and has been reported in past studies of instructed extinction (see Luck & Lipp, 2015a; 2015b; Rowles, Lipp, & Mallan, 2012). It is likely that this is a side effect of the experimental manipulation as the interaction with the experimenter may increase orienting. The finding of differences between first and second interval responding in an instructed reversal design supports the argument that multiple response scoring is important, especially in instructional designs (see Luck & Lipp 2016 for more details and a FIR/SIR vs. EIR scoring comparison).

The overall analysis of the change from the last trial of acquisition to the first trial of reversal did not provide evidence for a significant change in CS valence evaluations; however, inspection of Figure 4 and the 95% CI suggests that CS+ valence in the instructed CS+ first group became more pleasant after the instruction. Although inspection of Figure 4 suggests that a similar change may have been evident for the instructed CS- first group, this change was not significant and occurred in both instructed and control participants. The pattern of results observed in the instructed CS+ first group may suggest that there are differences between the effects of instructed extinction and instructed reversal, with the latter able to affect both CS valence evaluations and electrodermal responses.

The differences between instructional designs could occur because, while instructed extinction only affects the valence of the CS+, reversal instructions target the valence of both

CS+ and CS-. In the reversal design, not only does the absolute valence change (the CS+ is no longer paired with an aversive event), but also the relative valence (the CS+ is no longer the more negative of the two CSs). Differences between instructed extinction and instructed reversal could be explained by this CS- valence change if the participants make their evaluations in a relative fashion. It should be noted, however, that no such effect of instructed reversal was evident in the instructed CS- first condition or in Lipp et al's (2010) study of instruction effects on evaluative conditioning. Alternatively, a change in CS+ evaluation, but not CS- evaluation, may have been observed because the presentation of the CS+ alone during habituation allowed participants to form a CS+ –noUS representation which they could retrieve in response to the reversal instructions. No CS- –US pairings were presented before the reversal phase, and therefore participants would not have had the opportunity to form this representation. As electrodermal responding was immediately altered by the reversal instructions, it seems clear that relational propositions can be formed in response to instructions, but it is possible evaluative representations may not be able to form in a similar way based on instructions, but can be retrieved after instructions if a prior representation is available. This interpretation would be consistent with the failure of Lipp et al. (2010) to find an effect of instructed reversal on evaluative learning in a picture-picture paradigm as, unlike the current study, the picture-picture paradigm did not involve a habituation phase. It would not account for findings that instructed extinction failed to influence CS+ evaluations (Luck & Lipp, 2015a,b) as these experiments did include a habituation phase. As this interpretation is post-hoc it should be treated with caution until it has been empirically validated.

It is also possible that pre-existing valence differences in the CS- first group may have dampened the influence of the reversal instructions, leading to the observation that CS- valence

did not respond to instruction. The CS- first group evaluated CS+ as less pleasant than CS- during habituation, and this intrinsic negativity may have reduced the impact of instructed reversal on CS- valence if participants evaluated the stimuli in a relative fashion. A counterbalanced trial sequence was used and any valence differences occurring before the experiment are likely to be chance effects. Despite this, if the CS+ was intrinsically a negative stimulus for the some participants they may have been more reluctant to evaluate CS- more negatively than CS+ after the reversal instructions. Inspection of the reversal phase data in Figure 1 supports these suggestions, as participants in the control CS- first group evaluated the CS- as more negative than the instruction CS- first group, even at the end of the reversal phase. It is not possible to exclude the possibility that these pre-existing valence differences could have dampened the effects of instructed reversal on CS- valence, and therefore more work seems to be required to clarify this inconsistency

In addition to online ratings of stimulus valence, participants also provided ratings of CS valence in Likert scales before and after Pavlovian training (Ratings A and B), and after completion of the experiment (Ratings C). The pleasantness evaluations taken immediately after reversal training (Ratings B) revealed the same pattern of results as present in the online ratings throughout reversal training, i.e., the CS- was rated as more negative than the CS+. Interestingly however, when participants were asked to rate the faces in a different context (Ratings C), participants did not evaluate CS+ and CS- differently. This finding is in line with reports that participants integrate stimulus valence across an entire experiment when providing post-experimental ratings in a context (defined in this instance by place and mode of measurement) that is different from that in which the most recent experimental contingency was experienced (Lipp & Purkis, 2006). More broadly, it highlights the importance of assessing the emotional

response to an event in different contexts when assessing the effects of an intervention in experimental or applied settings.

The current investigation confirms that the reduction of the physiological indices in response to instructed extinction does not occur because of a drop in arousal levels. Furthermore, the current study suggests that an instructional manipulation may also influence evaluative learning. Demonstrating that both expectancy and evaluative learning respond to the same manipulation provides some support for the propositional learning account, but strong theoretical conclusions cannot be drawn on the basis of the current data as the difference in valence changes between CS+ and CS- first groups needs further investigation. If CS+, but not CS-, evaluations respond to instructed reversal, the pattern of results would be more in line with dual process models. More research will be required to investigate whether changes in the evaluations of CS+ and CS- differ on the process level and to disentangle the mechanisms underlying evaluative learning.

Table 1. Means and Standard Deviations for the Variables Assessed in the Preliminary Analyses

	CS+ First		CS- First	
	Instruction	Control	Instruction	Control
Gender Ratio (male:female)	10:21	10:21	11:22	14:20
Age	21.19 (4.15)	22.65 (4.36)	24.18 (5.63)	23.47 (3.68)
Spontaneous EDA	21.50 (15.00)	17.03 (16.82)	16.13 (12.65)	17.74 (14.13)
US Level	3.25 (1.07)	3.36 (0.96)	3.08 (0.74)	2.95 (0.82)
US Valence	-1.94 (0.59)	-1.82 (0.78)	-1.61 (1.06)	-1.94 (0.55)

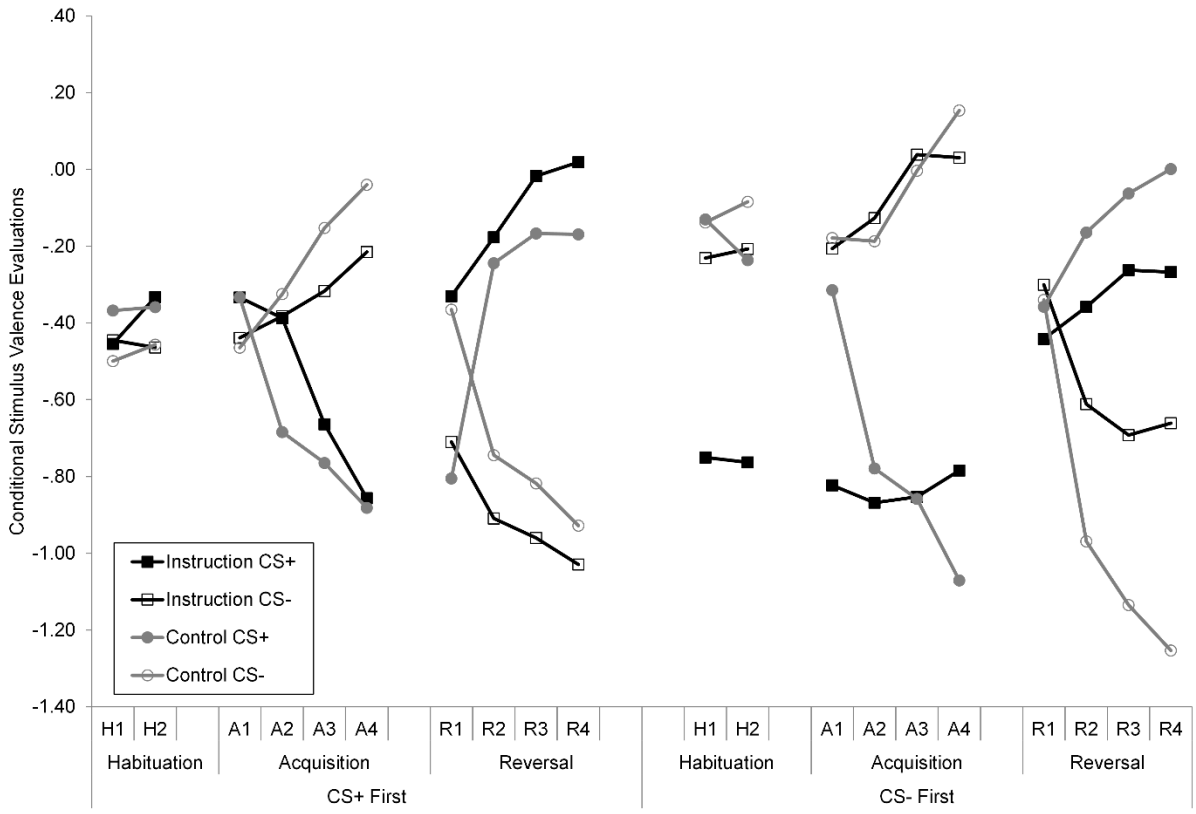

Figure 1. Conditional stimulus valence evaluations recorded throughout habituation, acquisition, and reversal

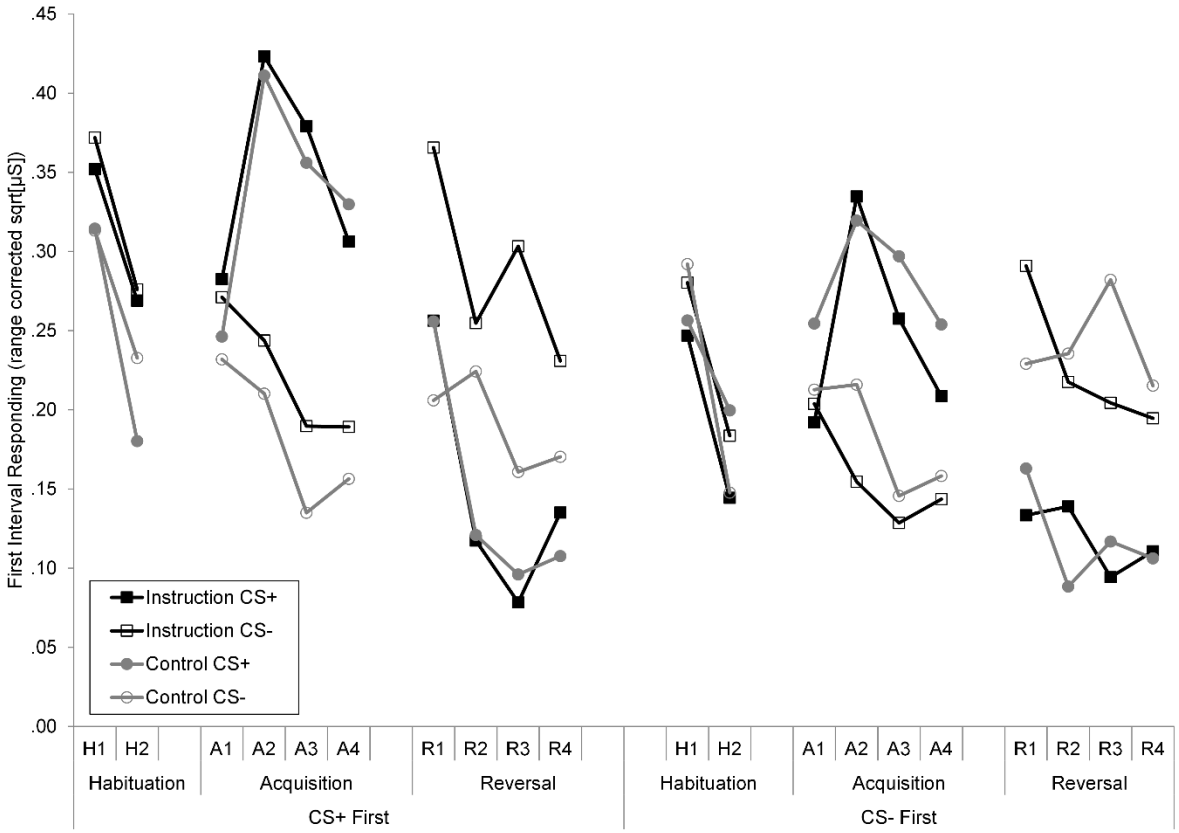

Figure 2. First interval electrodermal responding recorded throughout habituation, acquisition, and reversal.

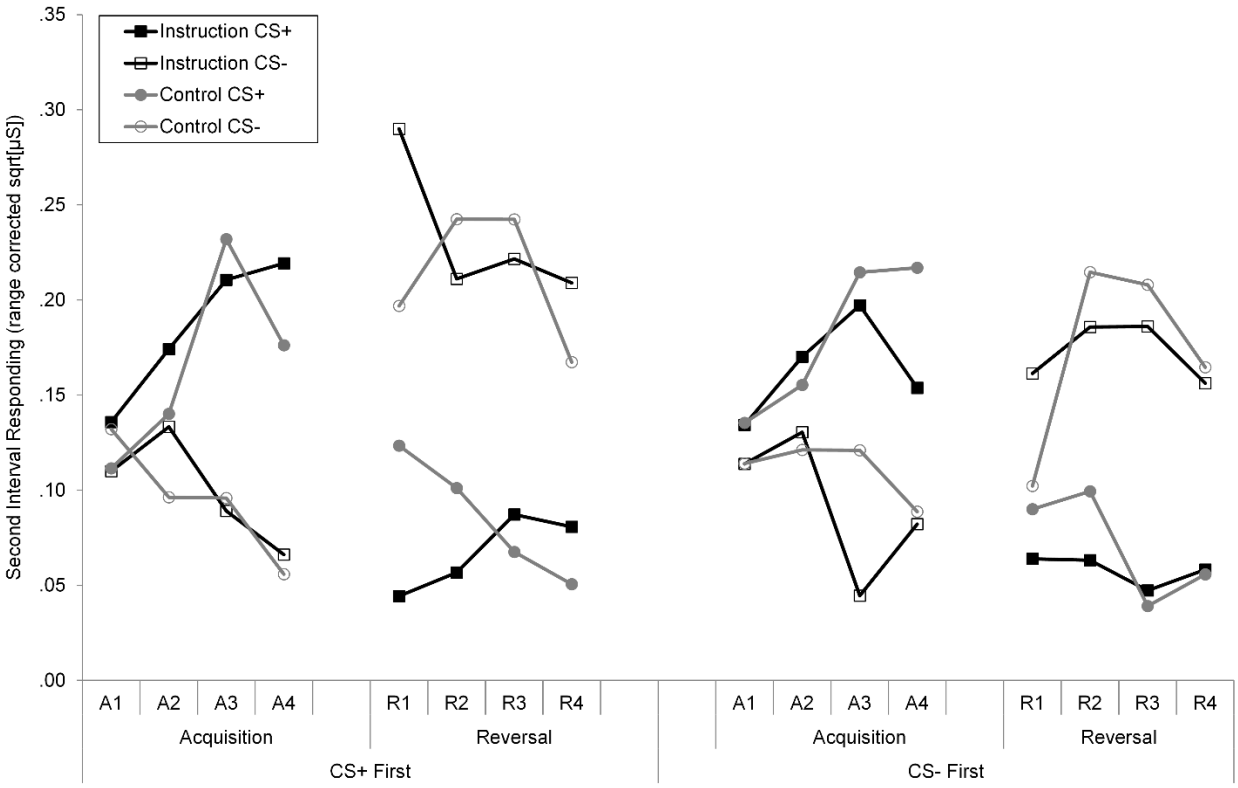

Figure 3. Second interval electrodermal responding recorded throughout acquisition and reversal.

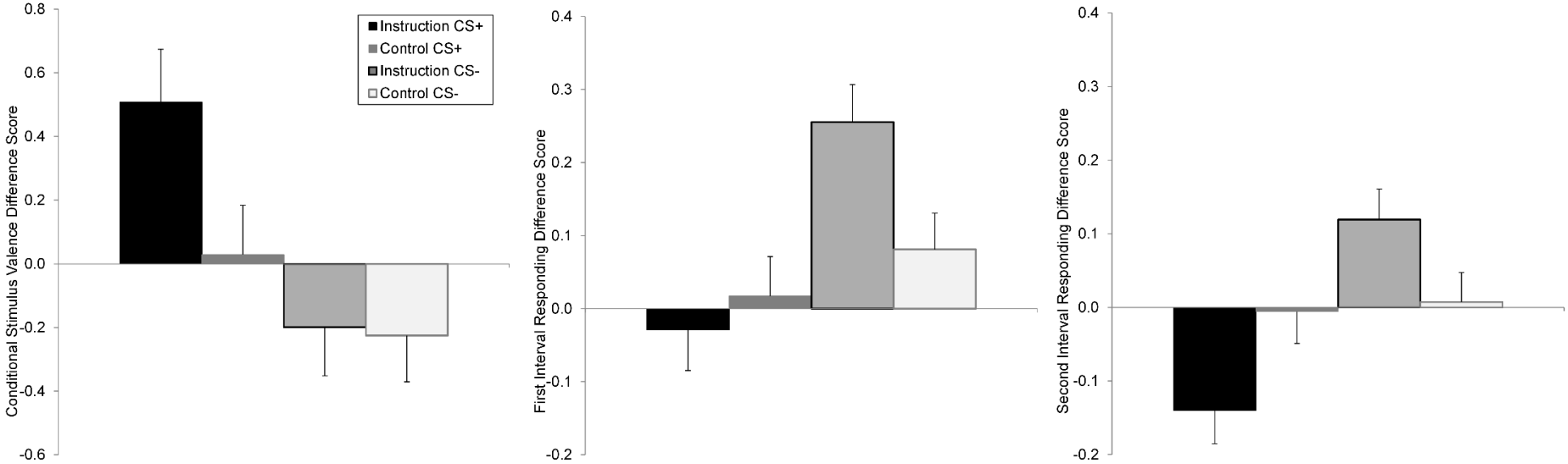

Figure 4. First trial difference scores (first reversal trial – last acquisition trial) for CS valence (left), first interval (middle), and second interval electrodermal responding (right). Positive values indicate that the stimulus is becoming more pleasant or that electrodermal responding is increasing. Negative values indicate that the stimulus is becoming less pleasant or that electrodermal responding is decreasing. (Error bars indicate standard errors of the mean).

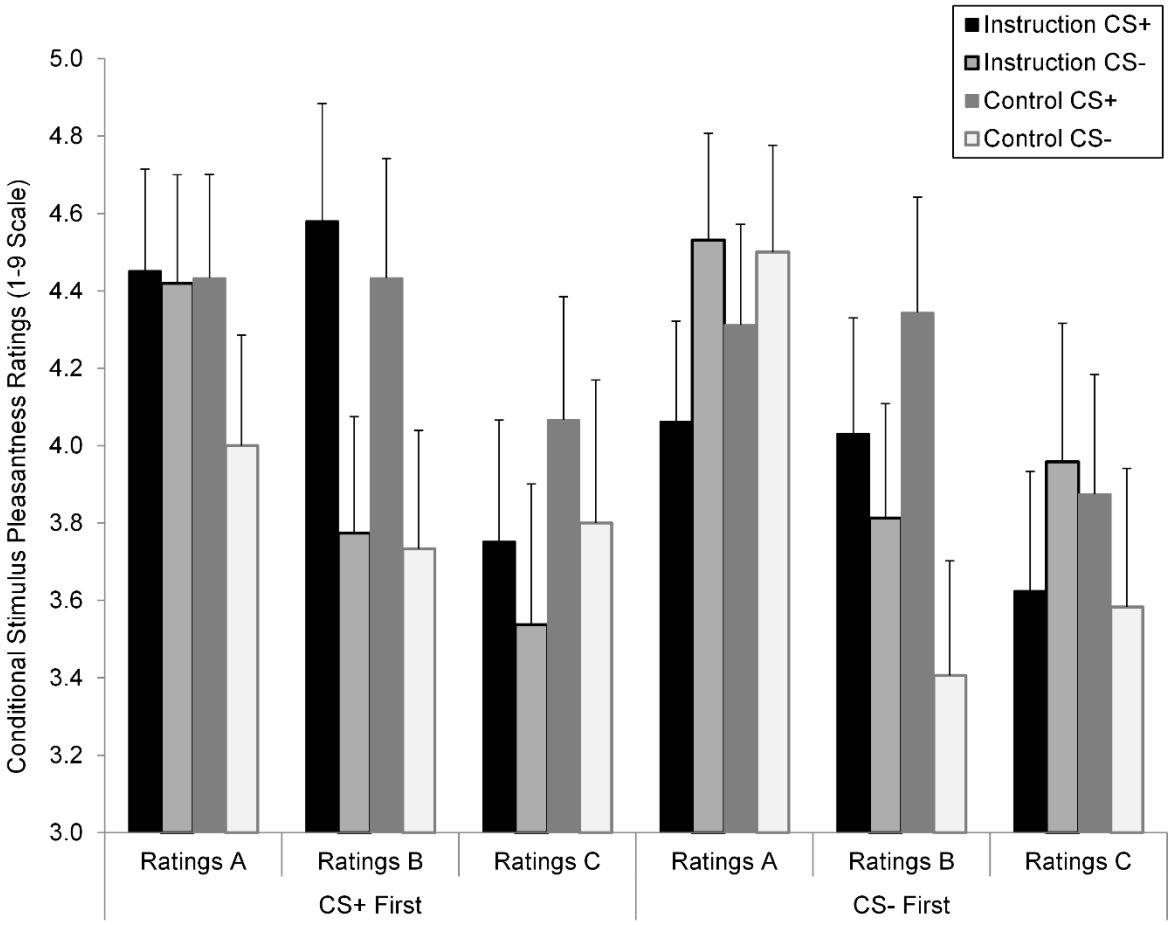

Figure 5. Conditional stimulus pleasantness ratings taken before conditioning (Ratings A), after reversal (Ratings B), and post-experimentally (Ratings C; Error bars indicate standard errors of the mean).

References

Baeyens, F., Eelen, P., Crombez, G., & van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour Research and Therapy, 30*, 133-142. doi: http://dx.doi.org/10.1016/0005-7967(92)90136-5

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

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*, 276-298. doi: 10.1037/1528-3542.1.3.276

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

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior, 37*, 1-20. doi: 10.3758/LB.37.1.1

De Houwer, J., Thomas, S., & Baeyens, F. (2001). Association learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin, 127*, 853-869. doi: 10.1037/0033-2909.127.6.853

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*, 116-124. doi: 10.3758/BF03195503

Grings, W. W., Schell, A. M., & Carey, C. A. (1973). Verbal control of an autonomic response in a cue reversal situation. *Journal of Experimental Psychology, 99*, 215-221. doi:10.1037/h0034653

Hermans, D., Dirikx, T., Vansteenwegenin, D., Baeyens, F., Van den Bergh, O., & Eelen, P. (2005). Reinstatement of fear responses in human aversive conditioning. *Behaviour Research and Therapy, 43,* 533-551. doi: 10.1016/j.brat.2004.03.013

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., 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

Lipp, O. V., & Purkis, H. M. (2006). The effects of assessment type on verbal ratings of conditional stimulus valence and contingency judgments: Implications for the extinction of evaluative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 32*, 431-440. doi: 10.1037/0097-7403.32.4.431

Luck, C. C., & Lipp, O. V. (2015). A potential pathway to the relapse of fear? Conditioned negative stimulus evaluation (but not physiological responding) resists instructed extinction. *Behaviour Research and Therapy, 66*, 18-31. doi: http://dx.doi.org/10.1016/j.brat.2015.01.001

Luck, C. C., & Lipp, O. V. (2015). To remove or not to remove? Removal of the unconditional stimulus electrode does not mediate instructed extinction effects. *Psychophysiology, 52*, 1248-1256. doi: 10.1111/psyp.12452

Luck, C. C., & Lipp, O. V. (2016). When orienting and anticipation dissociate — a case for scoring electrodermal responses in multiple latency windows in studies of human fear conditioning. *International Journal of Psychophysiology, 100*, 36-43. doi:http://dx.doi.org/10.1016/j.ijpsycho.2015.12.003

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.

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). U.S.A: Academic Press.

Rowles, M. E., Lipp, O. V., & Mallan, K. M. (2012). On the resistance to extinction of fear conditioned to angry faces. *Psychophysiology, 49*(3), 375-380. doi:10.1111/j.1469-8986.2011.01308.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: http://dx.doi.org/10.1016/j.psychres.2008.05.006

Zbozinek, T. D., Hermans, D., Prenoveau, J. M., Liao, B., & Craske, M. G. (2015). Post-extinction conditional stimulus valence predicts reinstatement fear: Relevance for long-term outcomes of exposure therapy. *Cognition and Emotion, 29*(4), 654-667. . doi: 0.1080/02699931.2014.930421