*Psychophysiology, 52* (2015), 325–332. Wiley Periodicals, Inc. Printed in the USA.
Copyright © 2014 Society for Psychophysiological Research
DOI: 10.1111/psyp.12359

# From threat to safety: Instructed reversal of defensive reactions

VINCENT D. COSTA,[a,b] MARGARET M. BRADLEY,[a] AND PETER J. LANG[a]

[a]Center for the Study of Emotion and Attention, University of Florida, Gainesville, Florida, USA
[b]National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland, USA

## Abstract

Cues that signal the possibility of receiving an electric shock reliably induce defensive activation. To determine whether cues can also easily reverse defensive reactions, a threat reversal paradigm was developed in which a cue signaling threat of shock reversed its meaning across the course of the study. This allowed us to contrast defensive reactions to threat cues that became safe cues, with responses to cues that continued to signal threat or safety. Results showed that, when participants were instructed that a previously threatening cue now signaled safety, there was an immediate and complete attenuation of defensive reactions compared to threat cues that maintained their meaning. These findings highlight the role that language can play both in instantiating and attenuating defensive reactions, with implications for understanding emotion regulation, social communication, and clinical phenomena.

**Descriptors:** Fear inhibition, Emotion regulation, Instructed fear, Threat of shock, Anticipation, Reversal learning

Humans use language to both elicit and attenuate fear. For example, media coverage of aversive events can heighten or attenuate fear reactions, as well as change perceptions about the likelihood of dangerous events (Slovic, 2000). Affective learning through language is also of clinical interest for its potential role in both the etiology and treatment of specific phobias (Askew & Field, 2008; Rachman, 2002). It is clear from psychophysiological studies that cues that merely signal the possibility of an aversive experience not only affect fear reports, but also elicit defensive physiological reactions. For instance, simply threatening a person with the possibility of receiving an electric shock or suffocation is sufficient to elicit an array of autonomic and somatic reactions indicative of defensive activation (Bradley, Moulder, & Lang, 2005; Bradley, Silakowski, & Lang, 2008; Cook & Harris, 1937; Grillon, Ameli, Woods, Merikangas, & Davis, 1991; Lang et al., 2011).

It is of particular interest to determine whether language is equally effective in attenuating defensive reactions. In prior studies exploring whether instructions can reduce conditioned fear reactions, the shock electrode was removed when instructions were given (Hugdahl, 1978; Hugdahl & Öhman, 1977; Lipp & Edwards, 2002; Mallan, Sax, & Lipp, 2009; Notterman, Schoenfeld, & Bersh, 1952; Soares & Öhman, 1993). Because simply attaching a shock electrode heightens defensive responding (Greenwald, Bradley, Cuthbert, & Lang, 1998; Grillon & Ameli, 1998), effects due to instruction cannot be easily disentangled from the effect of removing the shock delivery device. Moreover, some evidence suggests that instructions are less effective than when extinction is induced by reinforcer omission or vicarious observation (Olsson & Phelps, 2004).

Despite mixed evidence that instructions may facilitate extinction of fear reactions, language cues have proved effective in reversing electrodermal responses in a simple cue reversal situation (Grings, Schell, & Carey, 1973; McNally, 1981; Wilson, 1968). In these studies, one cue is initially associated with shock while a second cue is not. When participants are told that the shock contingency is reversed, there is an immediate reversal in the size of skin conductance responses elicited by each cue. Despite a clear decrease in electrodermal reactivity to the previously threatening cue, it remains unclear whether there is any residual defensive engagement when that cue now signals safety. The most appropriate comparison is to assess reactions to a cue when it newly signals safety to reactions elicited by a cue that consistently signals safety: this method controls for differences in habituation, sensitization, or other time-related changes in physiology.

In the current study, we developed a threat reversal paradigm that allowed us to assess reactions to threat cues that became safe cues, to cues that always signaled safety, and vice versa. Moreover, in addition to electrodermal responses, we examined a broad array of defensive reactions that included measures that might more specifically index changes in hedonic valence, including the startle blink reflex, facial frowning, and heart rate. In our novel adaptation of the threat of shock paradigm, cues that signaled threat and safety were compound cues that varied in both shape and color (Figure 1). Participants were initially told that one feature of the cue (e.g., color) signaled the possibility of receiving an electric shock (e.g., red) or safety (e.g., blue). Halfway through the experiment, new

This research was supported by grants from the National Institute of Dental and Craniofacial Research (R01 DE 13956) awarded to MMB, and from the National Institute of Mental Health awarded to PJL and MMB (P50 MH 72850) and to VDC (F31 MH 080551).

Address correspondence to: Vincent Costa, Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Building 49 Room B1EE17, Bethesda, MD 20892-4415, USA. E-mail: atsocvd@gmail.com

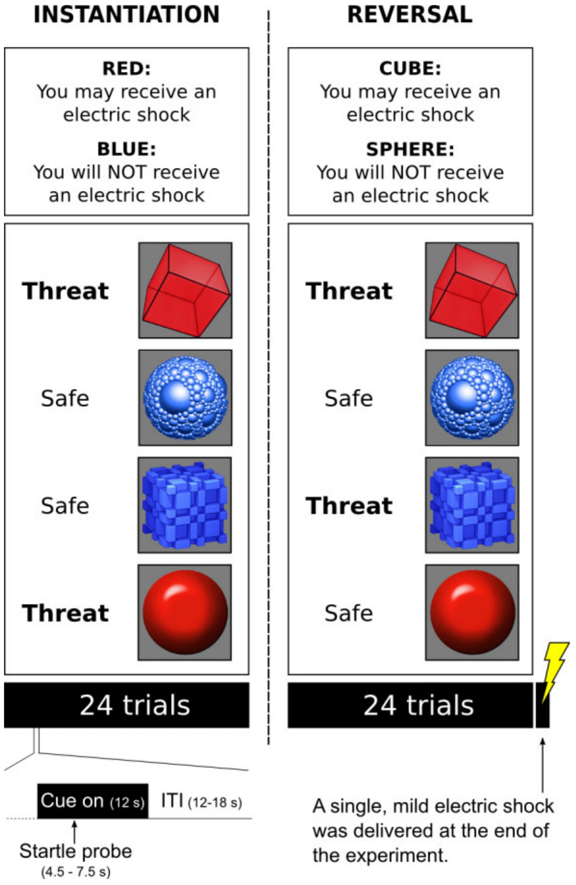
INSTANTIATION
RED:
You may receive an electric shock
BLUE:
You will NOT receive an electric shock
Threat
Safe
Safe
Threat
24 trials
REVERSAL
CUBE:
You may receive an electric shock
SPHERE:
You will NOT receive an electric shock
Threat
Safe
Threat
Safe
24 trials
Cue on (12 s)
ITI (12-18 s)
Startle probe
(4.5 - 7.5 s)
A single, mild electric shock was delivered at the end of the experiment.

**Figure 1.** Illustration of the instructed fear reversal paradigm.

instructions indicated that the other perceptual feature (e.g., shape) now indicated whether a cue signaled threat of shock (e.g., cube) or safety (e.g., sphere). As a result of these instructions, half of the cues reversed their prior associations with threat and safety, while the remaining half retained their affective meaning. By using multiple perceptual features to define threat and safe cues, we were able to compare—within participants—defensive reactions to threat and safe cues that reversed in meaning with those elicited by cues that consistently signaled threat or safety.

## Method

### Participants

Seventy-one students (46 female; *M* age = 19.6, *SD* = 2.1) from general psychology courses at the University of Florida participated for course credit, as approved by the University of Florida Institutional Review Board. All participants provided informed consent prior to their participation. Due to equipment or experimenter error, complete data were missing for two male participants and heart rate data were missing for one female participant.

### Materials and Design

Color (red/blue) and shape (cube/sphere) were crossed to form four sets of distinct perceptual cues that included 128 unique, abstract objects. In the instantiation phase, features of one cue—either color or shape—signaled the possibility of receiving threat of shock or safety. In the reversal phase (halfway through the experiment), a second set of instructions shifted attention to the other perceptual dimension to identify whether a cue signaled threat of shock or safety. The resulting four cues included (1) those that always signaled threat of shock, (2) those that always signaled safety, (3) safety cues that reversed in meaning to signal threat of shock, and (4) threat cues that reversed in meaning to signal safety. The specific colors and shapes associated with threat and safety in each phase were counterbalanced across participants.

A total of 48 trials were divided into two phases—instantiation and reversal. Each cue was presented for 12 s followed by an interstimulus interval (ITI) varying between 12 and 18 s in length. PC-compatible computers running Presentation (Neurobehavioral Systems, Inc., Albany, CA) and VPM software (Cook, 2001) were used to control stimulus presentation and collect all physiological measures. All visual cues and text instructions were projected onto a canvas screen measuring 121 cm × 182 cm using an LCD projector.

Startle probes were 96 db, 50-ms bursts of white noise generated by a Coulbourn S81-02 noise generator, gated by a Coulbourn S82-24 (Coulbourn Instruments, Whitehall, PA) audio-mixer amplifier and delivered through Telephonics TDH-96 (Farmingdale, NY) earphones. Startle probes were presented on 32 of the 48 trials between 4.5 and 7.5 s following cue onset.

A single very mild electric shock (20-ms duration, 1.6 mA) was delivered to the posterior surface of the right wrist at the end of the experiment to avoid deception. No additional shocks were delivered, and there was no shock workup conducted.

### Procedure

Participants first learned they would participate in a study involving electric shock when they read the informed consent form. No specific information was given about the amount or intensity of possible shock stimulation except for a statement that the shock would be perceptually equivalent to a needle prick or bee sting. Whenever participants asked about the shock during the consent process, the experimenter repeated verbatim what was stated in the consent form and emphasized that participation could be discontinued at any time without penalty.

After the sensors were attached, the experimenter instructed participants about the initial set of shock contingencies. For example, if the relevant feature dimension was color, participants were told that whenever a red cue appeared there was a possibility of receiving a shock, whereas during a blue cue there was absolutely no possibility of receiving a shock (or vice versa). No reference was made to the other feature dimension or that there would later be a reversal in the feature dimension cueing the shock contingencies.

The experimenter then showed the participant the stimulating bar electrode and attached it to the participant's right wrist. The bar electrode was connected to two braided wire cables, which were secured together at one end of the electrode with a sticker depicting the international safety symbol for high voltage. The sticker and cables were arranged so that they remained visible to the participant after attaching the bar electrode. In addition, the experimenter explained how a current would be passed between the two contacts in order to stimulate nerve fibers, and that the amount of current would result in a shock equivalent to a needle prick or a bee sting. Participants were given no additional information about the shock (e.g., when a shock would occur during threat periods or how many total shocks they would receive). Before leaving the room, the experimenter asked the participant to verbally reiterate the shock contingencies to verify that they understood the instructions.

Prior to presenting to the first cue, participants viewed a text slide restating the initial set of instructed contingencies (15 s). Midway through the experiment, a second set of text instructions appeared on the screen for 15 s and informed participants that a different perceptual feature now indicated whether a cue signaled threat of shock or safety.

After the experiment, all sensors were removed and participants used a 9-point Likert-type scale to rate the unpleasantness-pleasantness (anchored at 1 and 9, respectively) of their anticipation of shock during threat periods as well as the experience of receiving the shock. Participants were then debriefed and thanked for their participation. The entire procedure lasted approximately 2 h.

## Data Acquisition

**Startle blink magnitude.** The startle reflex was recorded using two small Ag/Ag-Cl electrodes placed over the left orbicularis oculi muscle. Raw orbicularis activity was acquired at 8–1000 Hz using a Coulbourn S75-01 bioamplifier and sampled at 2000 Hz from 100 ms prior to probe onset to 250 ms after probe onset. Offline, the digitized signal was filtered from 28–500 Hz (Blumenthal et al., 2005) using a Hamming windowed, nonrecursive bandpass filter, and then rectified and smoothed using a Butterworth filter with a 20-ms time constant.

**Corrugator EMG.** Activity over the corrugator supercilli muscle above the left eye was measured with small Ag/Ag-CL electrodes. The raw electromyographic (EMG) signal was bandpass filtered from 90–1000 Hz using a Coulbourn S75-01 bioamplifier, rectified and integrated using a Coulbourn S76-01 contour following integrator with a 500-ms time constant, sampled at 20 Hz, and half-second bins of mean corrugator EMG activity calculated offline.

**Skin conductance.** Skin conductance was measured using two large Ag/Ag-Cl electrodes filled with 0.05 NaCl paste (TD-246; Mansfield R & D, St. Albans, VT), placed adjacently over the hypothenar eminence of the left palm. A constant current (.5 V) was generated between the electrodes using a Coulbourn S71-22 coupler. Activity was sampled and digitized at 20 Hz, and half-second bins of mean skin conductance were calculated offline.

**Heart rate.** The electrocardiogram was recorded from the left and right forearms, using large Ag/Ag-CL electrodes and a Coulbourn S75-01 bioamplifier with a bandpass filter of 8–40 Hz. Raw electrocardiogram activity was sampled at 500 Hz. Offline R-wave spikes were registered and interbeat intervals calculated for conversion to, which were then converted into heart rate in beats per minute (bpm) in half-second averages with each interval weighted proportionally to the amount of time it occupied (Graham, 1978).

## Data Analysis

Startle blink magnitude and onset latency were scored offline using a peak scoring algorithm (Globisch, Hamm, Schneider, & Vaitl, 1993) implemented in MATLAB. Automated detection of startle reflex onset and peak magnitude was verified by manually screening each trial after running the algorithm. Trials with an onset of less than 20 ms or excessive baseline activity were omitted from the analysis (< 1%). Otherwise trials with no discernable response were scored as zero magnitude.

To assess reactions during the threat and safe periods, each half-second bin of corrugator, skin conductance, and heart rate activity was deviated from a 1-s baseline prior to cue onset. This resulted in change scores for each measure that reflected increases (or decreases) from baseline during the cue presentation period. Startle probe presentation is known to cause increases in skin conductance and heart rate (Bradley et al., 2008). To avoid mixing responses elicited by the cue or the later startle probe, skin conductance change was averaged between 1–4 s after cue onset. Heart rate was scored in terms of the maximum initial deceleration from baseline (D1) and subsequent maximum midinterval acceleration (A1) in the first 6-s postcue onset (Bradley, Codispoti, Sabatinelli, & Lang, 2001). Corrugator activity was unaffected by the startle probe and was averaged across the entire 12-s cue period. Cue-elicited skin conductance and heart rate responses, when plotted, represent averages using trials that did not include startle probe presentations; however, statistical analyses included all trials.

For each measure, a mixed effects analysis of variance (ANOVA) was computed separately for each phase (instantiation and reversal) and included repeated measures of cue (threat, safe), contingency (maintain, reverse), and a between-participant factor of gender (male, female). In the reversal phase, cues were coded as either threat or safe based on the current set of contingencies. Coded in this manner, a main effect of cue in the reversal phase indicates a significant difference in reactivity during threat and safety, regardless of whether the cue contingencies from the instantiation phase were maintained or reversed. On the other hand, an interaction of contingency and cue indicated that reactions to threat and safe cues differed based on whether a cue had maintained or reversed its prior association with threat or safety.

# Results

Table 1 lists mean blink magnitude, skin conductance activity, corrugator EMG activity, and heart rate change during threat and safe periods as a function of phase (instantiation, reversal), contingency (maintain, reverse), and cue (safe, threat). There were no main effects or interactions involving gender for any measure other than corrugator activity.

## Startle Blink Reflex

In the instantiation phase, blink magnitude was larger for startle probes presented in the context of cues that signaled threat of shock, compared to those that signaled safe periods, $F(1,67) = 49.73$, $p < .001$, $d = 0.84$ (Figure 2). As expected, startle magnitude during threat and safety in the instantiation phase did not differ based on whether meaning of a cue would be maintained or reversed in the second phase (Cue × Contingency, $F(1,67) = 1.1$, *ns*; $d = 0.12$).

In the reversal phase, blink magnitude was again larger for probes presented in the context of threat compared to safe cues, Cue: $F(1,67) = 21.65$, $p < .001$, $d = 0.55$. Importantly, there was no interaction of cue and contingency, $F(1,67) < 1$, *ns*. Not surprisingly, when cues retained their original meaning, blink reflexes remained potentiated during threat compared to safe periods in the reversal phase, $t(68) = 3.53$, $p < .001$, $d = 0.42$. More importantly, startle reflexes were larger for threat cues that had previously signaled safety, compared to safe cues that previously signaled threat of shock ($t = 3.49$, $p < .001$, $d = 0.41$), and blink magnitude for probes presented in the context of safe cues that had initially been threatening was equivalent to blinks elicited during cues that

**Table 1.** *Defensive Reactions During Threat and Safe Periods Before and After the Instructed Contingency Reversal for Cues that Maintained or Reversed in Meaning*

| | Instantiation | | Reversal | | Overall | |
|---|---|---|---|---|---|---|
| | Threat | Safe | Threat | Safe | Threat | Safe |
| Startle magnitude (μV) | | | | | | |
| Maintain | 22.9 (3.0) | 11.9 (1.9) | 13.8 (2.2) | 8.8 (1.5) | 18.3 (2.5) | 10.4 (1.6) |
| Reverse | 22.7 (3.1) | 13.2 (2.1) | 14.3 (2.4) | 9.2 (1.7) | 18.5 (2.6) | 11.2 (1.7) |
| Skin conductance (μS) | | | | | | |
| Maintain | .14 (.03) | −.01 (.01) | .04 (.02) | −.01 (.01) | .09 (.02) | −.01 (.01) |
| Reverse | .15 (.03) | −.02 (.01) | .04 (.02) | −.01 (.01) | .10 (.02) | −.01 (.01) |
| Heart rate: D1 (bpm) | | | | | | |
| Maintain | −7.8 (.42) | −6.54 (.42) | −6.64 (.35) | −6.11 (.34) | −7.26 (.33) | −6.33 (.32) |
| Reverse | −8.0 (.42) | −6.56 (.41) | −6.78 (.35) | −5.67 (.34) | −7.42 (.33) | −6.11 (.32) |
| Heart rate: A1 (bpm) | | | | | | |
| Maintain | 4.98 (.37) | 6.49 (.36) | 5.46 (.34) | 6.47 (.34) | 5.22 (.29) | 6.48 (.29) |
| Reverse | 4.76 (.37) | 6.55 (.37) | 5.18 (.34) | 6.46 (.33) | 4.96 (.29) | 6.51 (.30) |
| Corrugator EMG (μV) | | | | | | |
| Maintain | .49 (.09) | .19 (.08) | .60 (.10) | .33 (.09) | .55 (.08) | .26 (.07) |
| Reverse | .50 (.11) | .25 (.09) | .63 (.09) | .35 (.09) | .57 (.08) | .30 (.07) |

*Note*. Values in parentheses indicate standard error of the mean (*SEM*).

always signaled safety. Similarly, startle magnitude was equally potentiated for probes presented in the context of threat cues that had initially signaled safety and those that had always signaled shock threat.

To assess how long it took for startle magnitude to decrease when a cue reversed in meaning and no longer signaled threat of shock, blink magnitude was compared for individual startle probe presentations in the reversal phase. Blink magnitude during cues that had previously signaled threat of shock was attenuated compared to cues that always signaled threat of shock on the initial trial in the reversal phase, all $t$s(68) < −2.47, $p$s < .008, $d$s > .27, and was equivalent to blink responses elicited in the context of cues that consistently signaled safety.

### Skin Conductance, Corrugator EMG, and Heart Rate

During instantiation, defensive reactions were heightened in the context of threat (compared to safe) cues, including increased skin conductance, $F(1,67) = 63.27$, $p < .001$, $d = 0.9$, heightened corrugator EMG activity, $F(1,67) = 5.83$, $p = .018$, $d = 0.27$, as

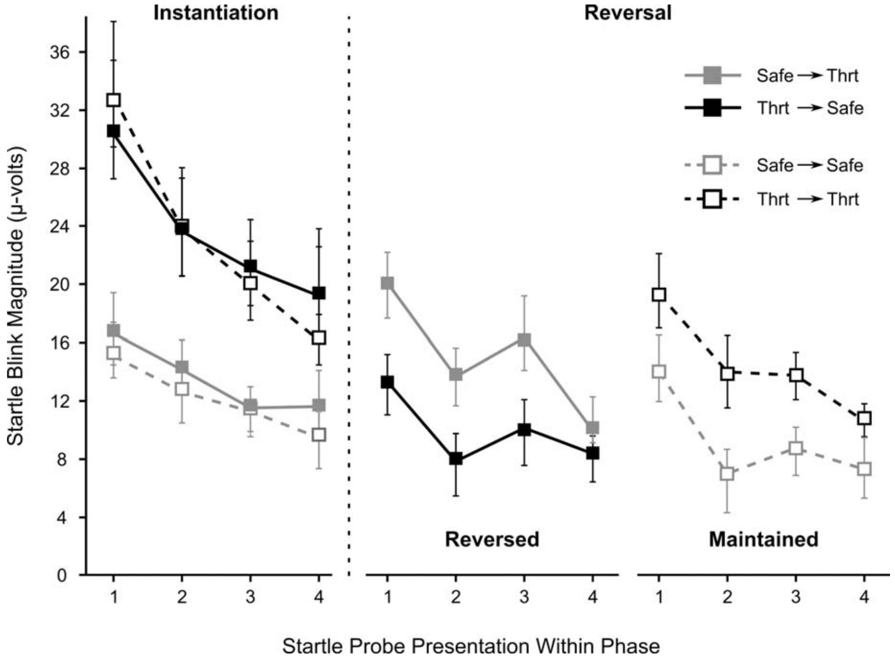

**Figure 2.** Startle blink magnitude is plotted for acoustic startle probes presented in the instantiation (left) and reversal (right) phases. In the reversal phase, blink magnitude is separately plotted for cues whose initial associations with threat or safety were either reversed or maintained. Error bars represent standard error of the mean.

well as greater initial cardiac deceleration, D1: $F(1,66) = 20.18$, $p < .001$, $d = 0.55$, followed by reduced midinterval acceleration, A1: $F(1,66) = 45.81$, $p < .001$, $d = 0.62$. Similar to the startle reflex, there were no interactions involving cue and contingency in the reversal phase for skin conductance or corrugator EMG activity. Rather, despite their initial meaning, threat compared to safe periods were associated with heightened skin conductance, $F(1,67) = 8.6$, $p = .004$, $d = 0.35$, and corrugator EMG activity, $F(1,67) = 5.85$, $p = .018$, $d = 0.29$ (Figure 3A, 3B).

Whether a cue maintained or reversed its meaning did impact threat- and safety-related changes for each heart rate component (Figure 3C; D1: Cue × Contingency, $F(1,66) = 6.05$, $p = .016$, $d = 0.29$; A1: Cue × Contingency, $F(1,66) = 4.34$, $p = .041$, $d = 0.25$). For cues whose meaning was reversed, threat cues elicited greater initial cardiac deceleration while safety cues prompted heightened midinterval acceleration. In contrast, there were no differences in either heart rate component among cues whose initial meaning remained unchanged (all *p*s > .16).

To determine how quickly the instructed reversal attenuated threat-related changes in these measures, we compared trial-by-trial reactions in the reversal phase to threat cues that became safe cues, and to safe cues that maintained their hedonic meaning. Threat-related changes in skin conductance, $t(63) = -2.26$, $p = .027$, $d = 0.27$; corrugator EMG, $t(63) = -2.6$, $p = .011$, $d = 0.31$; and heart rate, D1: $t(63) = 2.02$, $p = .047$, $d = 0.24$, were significantly attenuated during the first presentation of a threat cue that reversed its meaning, compared to reactions elicited by the most proximal presentation of a threat cue that maintained its meaning. More importantly, responses during the first presentation of a cue that had reversed in meaning from threat to safety did not differ from responses elicited during cues that consistently signaled safety (all *p*s > .241). Thus, as for the startle reflex, there was an immediate reversal in defensive reactions elicited by threat and safety cues following instructions that reversed their meaning.

Replicating previous findings (Bradley et al., 2001), women were more facially expressive than men, with greater overall corrugator activity both in the instantiation, $F(1,67) = 10.22$, $p = .002$, $d = 0.38$, and reversal phase, $F(1,67) = 4.24$, $p = .043$, $d = 0.24$.

### Aversiveness Ratings

Not surprisingly, shock anticipation was rated as somewhat unpleasant by all participants ($M = 4.19$, $SE = 0.29$) and differed significantly from the neutral point on the rating scale, $t(67) = -2.71$, $p = .0085$, $d = .32$. There was no difference in ratings between men and women, $F(1,67) = 1.94$, $p = .168$.

## Discussion

Language was effective not only in eliciting defensive reactions when a threatening cue was present, but also in quickly attenuating defensive reactions when the same cue no longer signaled the possibility of shock. Thus, when cues that had initially signaled threat of shock reversed in meaning and signaled safety, there was no evidence of an initial association with threat of shock. Instead, there were marked decreases in the startle reflex, skin conductance, frowning activity, and heart rate deceleration. This broad attentuation of defensive reactions is consistent with prior studies that used instructions to extinguish conditioned autonomic and somatic reactions (e.g., Dawson & Schell, 1985; Grings et al., 1973; Lipp & Edwards, 2002). Importantly, reactions to cues that no longer signaled threat of shock were indistinguishable from responses to cues that had consistently signaled safety, which provides an important control for simple effects of time. Moreover, fear attenuation occurred in a context in which threat of shock remained possible and the shock electrode remained attached.

Learning conflicting hedonic associations has been shown to impede learning of contingency reversals and to delay appropriate changes in conditioned responding (Bouton, 1993; Peck & Bouton, 1990). For example, in Pavlovian conditioning, several trials are typically needed following an unsignaled reversal or termination of the shock contingencies before fear responses accurately reflect the new contingencies (Schiller, Levy, Niv, Ledoux, & Phelps, 2008). Straightforward extinction of fear-conditioned skin conductance and startle responses are similarly protracted in humans (Hamm & Vaitl, 1996; Norrholm et al., 2006). Here, there was no evidence of lingering defensive engagement when a threat cue now signaled safety—even in measures sensitive to changes in hedonic valence. The immediate reversal of defensive reactions found here therefore contrasts with the slower reversal of defensive reactions observed following Pavlovian conditioning. The most obvious difference in the two contexts is that conditioning includes the actual experience of a painful or aversive reinforcer, whereas instructed threat relies instead on memory or imagery of previous aversive experiences for defensive activation (Olsson & Phelps, 2007).

Learning in the absence of a physical reinforcer could be weaker or more labile, making it easier to reverse reactions. On the other hand, though, threat cues elicit equivalent electrodermal reactions whether fear is instantiated through instructions or actual conditioning (Bridger & Mandel, 1964; Raes, De Houwer, De Schryver, Brass, & Kalisch, 2014), and instructions are also quite effective in quickly reversing explictly conditioned electrodermal responses (Grings et al., 1973; McNally, 1981; Wilson, 1968). Moreover, startle reflex potentiation remains elevated for up to 3 days to cues signaling threat of shock, even though threat cues were not followed by an actual shock reinforcer (Bublatzky, Gerdes, & Alpers, 2014). Finally, vividly imagining an aversive reinforcer during conditioning, without actually reexperiencing it, prolongs conditioned skin conductance responses (Jones & Davey, 1990). Despite these apparent similarities, it would be worthwhile to measure fear reversal in a version of the current paradigm that paired cue presentations with an actual aversive event.

The extinction or reversal of Pavlovian conditioned responses is suggested to rely on experiencing the aversive cue in an unreinforced context, which then generates a prediction error used to update the previous contingency (Schiller et al., 2008). If so, language may facilitate fear reversal because prior knowledge about the contingencies can be updated without actual experience. For example, when participants are instructed about a change in the intensity of the shock used during conditioning, skin conductance responses elicited by the next shock are smallest when the change in shock intensity matched the instructions (i.e., no prediction error), and largest when there was a discrepancy between the instructed and actual change in shock intensity (i.e., prediction error; Öhman, 1971). Thus, instructed reversal might be more similar to aversive learning under continuous reinforcement (Bouton, 2007), or following the development of a learning set (Wilson & Gaffan, 2008)—situations where prediction errors are highly informative, and both animals and humans quickly learn to shift behavioral responses.

The finding that instructed learning is effective in both potentiating and attenuating fear reactions suggests the threat reversal paradigm may prove useful in probing mechanisms of fear

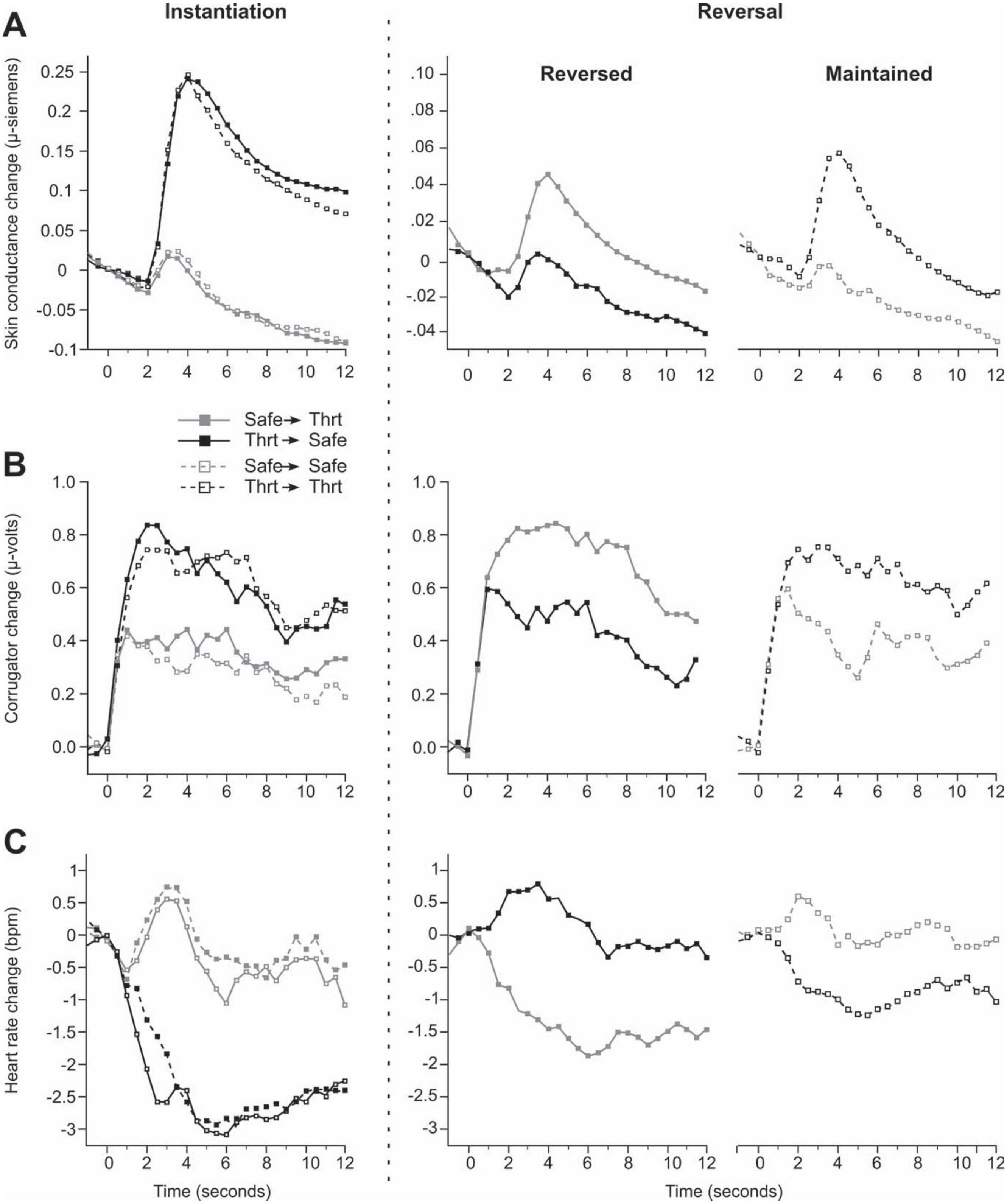

**Figure 3.** Threat of shock heightens defensive reactions both before and after instructions that reversed the threat and safety contingency changes: (A) skin conductance, (B) corrugator EMG activity, (C) heart rate. Skin conductance and heart rate are averaged over trials in which a startle probe was not presented.

processing in disorders that involve catastrophizing and worry (e.g., panic disorder or generalized anxiety disorder) and in specific phobias that develop in spite of actually experiencing the feared stimulus (Askew & Field, 2008; Rachman, 2002). Studying instructed fear reversal in these populations may assist in determining how instructions reconfigure perceptual and response biases to lessen fear, processes which form the foundation of empirically supported cognitive behavioral therapies. This would intersect with existing evidence that anxiety disorder patients show resistance in fear extinction (Lissek et al., 2005).

The use of language to reverse threat and safety contingencies and alter fear reactions can be considered a form of emotion regulation (Hartley & Phelps, 2010; Ochsner & Gross, 2008). As suggested (Hartley & Phelps, 2010; Schiller & Delgado, 2010), comparing results from fear conditioning and emotion regulation studies will be useful in formulating more mechanistic accounts of how emotion can be regulated. One important advantage of the threat reversal paradigm used here, in terms of studying emotion regulation, is that changes in emotion are directly linked to a specific experimental manipulation (e.g., detection of the relevant feature dimension), rather than relying on participant-generated mental strategies (Jackson, Malmstadt, Larson, & Davidson, 2000; Lissek et al., 2007; McRae et al., 2010). In addition, this paradigm allows for coincident comparisons to cues that have maintained their affective meaning while controlling for extraneous factors known to affect psychophysiological reactivity (e.g., response habituation, lapses in attention, time). Because the threat reversal paradigm employs language to instantiate and reverse defensive reactions, it could prove useful in examining the links between the reversal, extinction, and the cognitive regulation of fear.

## References

Askew, C., & Field, A. P. (2008). The vicarious learning pathway to fear 40 years on. *Clinical Psychology Review*, *28*, 1249–1265. doi: 10.1016/j.cpr.2008.05.003

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

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, *114*, 80–99.

Bouton, M. E. (2007). *Learning and behavior: A contemporary synthesis*. Sunderland, MA: Sinauer Associates.

Bradley, M. M., Codispoti, M., Sabatinelli, D., & Lang, P. J. (2001). Emotion and motivation II: Sex differences in picture processing. *Emotion*, *1*, 300–319. doi: 10.1037//1528-3542.1.3.300

Bradley, M. M., Moulder, B., & Lang, P. J. (2005). When good things go bad—The reflex physiology of defense. *Psychological Science*, *16*, 468–473.

Bradley, M. M., Silakowski, T., & Lang, P. J. (2008). Fear of pain and defensive activation. *Pain*, *137*, 156–163. doi: 10.1016/J.Pain.2007.08.027

Bridger, W. H., & Mandel, I. J. (1964). A comparison of GSR fear responses produced by threat and electric-shock. *Journal of Psychiatric Research*, *2*, 31–40. doi: 10.1016/0022-3956(64)90027-5

Bublatzky, F., Gerdes, A. B., & Alpers, G. W. (2014). The persistence of socially instructed threat: Two threat-of-shock studies. *Psychophysiology*. Advance online publication. doi: 10.1111/psyp.12251

Cook, E. W. (2001). VPM reference manual. Birmingham, AL: Author.

Cook, S. W., & Harris, R. E. (1937). The verbal conditioning of the galvanic skin reflex. *Journal of Experimental Psychology*, *21*, 202–210. doi: 10.1037/H0063197

Dawson, M. E., & Schell, A. M. (1985). Information processing and human autonomic conditioning. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles. *Advances in psychophysiology* (Vol. 1, pp. 89–165). Greenwich, CT: JAI Press.

Globisch, J., Hamm, A. O., Schneider, R., & Vaitl, D. (1993). A computer program for scoring reflex eyeblink and electrodermal responses written in Pascal. *Psychophysiology*, *39*, S30.

Graham, F. K. (1978). Constraints on measuring heart rate and period sequentially through real and cardiac time. *Psychophysiology*, *15*, 492–495.

Greenwald, M. K., Bradley, M. M., Cuthbert, B. N., & Lang, P. J. (1998). Startle potentiation: Shock sensitization, aversive learning, and affective picture modulation. *Behavioral Neuroscience*, *112*, 1069–1079. doi: 10.1037//0735-7044.112.5.1069

Grillon, C., & Ameli, R. (1998). Effects of threat of shock, shock electrode placement and darkness on startle. *International Journal of Psychophysiology*, *28*, 223–231.

Grillon, C., Ameli, R., Woods, S. W., Merikangas, K., & Davis, M. (1991). Fear-potentiated startle in humans: Effects of anticipatory anxiety on the acoustic blink reflex. *Psychophysiology*, *28*, 588–595.

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

Hamm, A. O., & Vaitl, D. (1996). Affective learning: Awareness and aversion. *Psychophysiology*, *33*, 698–710. doi: 10.1111/J.1469-8986.1996.Tb02366.X

Hartley, C. A., & Phelps, E. A. (2010). Changing fear: The neurocircuitry of emotion regulation. *Neuropsychopharmacology*, *35*, 136–146. doi: 10.1038/Npp.2009.121

Hugdahl, K. (1978). Electrodermal conditioning to potentially phobic stimuli: Effects of instructed extinction. *Behaviour Research and Therapy*, *16*, 315–321.

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*, 608–618. doi: 10.1037//0278-7393.3.5.608

Jackson, D. C., Malmstadt, J. R., Larson, C. L., & Davidson, R. J. (2000). Suppression and enhancement of emotional responses to unpleasant pictures. *Psychophysiology*, *37*, 515–522. doi: 10.1111/1469-8986.3740515

Jones, T., & Davey, G. C. L. (1990). The effects of cued ucs rehearsal on the retention of differential fear conditioning—An experimental analog of the worry process. *Behaviour Research and Therapy*, *28*, 159–164. doi: 10.1016/0005-7967(90)90028-H

Lang, P. J., Wangelin, B. C., Bradley, M. M., Versace, F., Davenport, P. W., & Costa, V. D. (2011). Threat of suffocation and defensive reflex activation. *Psychophysiology*, *48*, 393–396. doi: 10.1111/j.1469-8986.2010.01076.x

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

Lissek, S., Orme, K., McDowell, D. J., Johnson, L. L., Luckenbaugh, D. A., Baas, J. M., . . . Grillon, C. (2007). Emotion regulation and potentiated startle across affective picture and threat-of-shock paradigms. *Biological Psychology*, *76*, 124–133. doi: 10.1016/J.Biopsycho.2007.07.002

Lissek, S., Powers, A. S., McClure, E. B., Phelps, E. A., Woldehawariat, G., Grillon, C., & Pine, D. S. (2005). Classical fear conditioning in the anxiety disorders: A meta-analysis. *Behaviour Research and Therapy*, *43*, 1391–1424. doi: 10.1016/J.Brat.2004.10.007

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

McNally, R. J. (1981). Phobias and preparedness—Instructional reversal of electrodermal conditioning to fear-relevant stimuli. *Psychological Reports*, *48*, 175–180.

McRae, K., Hughes, B., Chopra, S., Gabrieli, J. D. E., Gross, J. J., & Ochsner, K. N. (2010). The neural bases of distraction and reappraisal. *Journal of Cognitive Neuroscience*, *22*, 248–262. doi: 10.1162/Jocn.2009.21243

Norrholm, S. D., Jovanovic, T., Vervliet, B., Myers, K. M., Davis, M., Rothbaum, B. O., & Duncan, E. J. (2006). Conditioned fear extinction and reinstatement in a human fear-potentiated startle paradigm. *Learning & Memory*, *13*, 681–685. doi: 10.1101/Lm.393906

Notterman, J. M., Schoenfeld, W. N., & Bersh, P. J. (1952). A comparison of three extinction procedures following heart rate conditioning. *Journal of Abnormal Psychology*, *47*, 675–677.

Ochsner, K. N., & Gross, J. J. (2008). Cognitive emotion regulation: Insights from social cognitive and affective neuroscience. *Current Directions in Psychological Science*, *17*, 153–158. doi: 10.1111/J.1467-8721.2008.00566.X

Öhman, A. (1971). Interaction between Instruction-induced expectancy and strength of unconditioned stimulus in GSR conditioning. *Journal of Experimental Psychology*, *88*, 384–390. doi: 10.1037/H0030897

Olsson, A., & Phelps, E. A. (2004). Learned fear of "unseen" faces after Pavlovian, observational, and instructed fear. *Psychological Science*, *15*, 822–828. doi: 10.1111/j.0956-7976.2004.00762.x

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience*, *10*, 1095–1102. doi: 10.1038/Nn1968

Peck, C. A., & Bouton, M. E. (1990). Context and performance in aversive-to-appetitive and appetitive-to-aversive transfer. *Learning and Motivation*, *21*, 1–31. doi: 10.1016/0023-9690(90)90002-6

Rachman, S. (2002). Fears born and bred: Non-associative fear acquisition? *Behaviour Research and Therapy*, *40*, 121–126.

Raes, A. K., De Houwer, J., De Schryver, M., Brass, M., & Kalisch, R. (2014). Do CS-US pairings actually matter? A within-subject comparison of instructed fear conditioning with and without actual CS-US pairings. *PLoS One*, *9*, e84888. doi: 10.1371/journal.pone.0084888

Schiller, D., & Delgado, M. R. (2010). Overlapping neural systems mediating extinction, reversal and regulation of fear. *Trends in Cognitive Sciences*, *14*, 268–276. doi: 10.1016/J.Tics.2010.04.002

Schiller, D., Levy, I., Niv, Y., Ledoux, J. E., & Phelps, E. A. (2008). From fear to safety and back: Reversal of fear in the human brain. *Journal of Neuroscience*, *28*, 11517–11525. doi: 10.1523/Jneurosci.2265-08.2008

Slovic, P. (2000). *The perception of risk*. Sterling, VA: Earthscan Publications.

Soares, J. J., & Öhman, A. (1993). Preattentive processing, preparedness and phobias: Effects of instruction on conditioned electrodermal responses to masked and non-masked fear-relevant stimuli. *Behaviour Research and Therapy*, *31*, 87–95.

Wilson, C. R. E., & Gaffan, D. (2008). Prefrontal-inferotemporal interaction is not always necessary for reversal learning. *Journal of Neuroscience*, *28*, 5529–5538. doi: 10.1523/Jneurosci.0952-08.2008

Wilson, G. D. (1968). Reversal of differential GSR conditioning by instructions. *Journal of Experimental Psychology*, *76*, 491–493.

(Received March 7, 2014; Accepted August 28, 2014)