Received: 5 February 2021 | Revised: 24 October 2021 | Accepted: 3 November 2021

DOI: 10.1111/psyp.13979

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Rating expectations can slow aversive reversal learning

Lauren Y. Atlas[1,2,3] | **Christina F. Sandman**[4] | **Elizabeth A. Phelps**[5]

[1]National Center for Complementary and Integrative Health, National Institutes of Health, Bethesda, Maryland, USA

[2]National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland, USA

[3]National Institutes on Drug Abuse, National Institutes of Health, Baltimore, Maryland, USA

[4]Department of Psychology, University of California, Los Angeles, California, USA

[5]Department of Psychology, Harvard University, Cambridge, Massachusetts, USA

Correspondence
Lauren Y. Atlas, National Center for Complementary and Integrative Health, National Institutes of Health, 10 Center Drive, Rm 4-1741, Bethesda, MD 20892, USA.
Email: lauren.atlas@nih.gov

Funding information
National Center for Complementary and Integrative Health, Grant/Award Number: ZIA-AT000030; National Institute of Mental Health, Grant/Award Number: RO1MH097085

Abstract

The process of learning allows organisms to develop predictions about outcomes in the environment, and learning is sensitive to both simple associations and higher order knowledge. However, it is unknown whether consciously attending to expectations shapes the learning process itself. Here, we directly tested whether rating expectations shapes arousal during classical conditioning. Participants performed an aversive learning paradigm wherein one image (CS+) was paired with shock on 50% of trials, while a second image (CS−) was never paired with shock. Halfway through the task, contingencies reversed. One group of participants rated the probability of upcoming shock on each trial, while the other group made no online ratings. We measured skin conductance response (SCR) evoked in response to the CS and used traditional analyses as well as quantitative models of reinforcement learning to test whether rating expectations influenced arousal and aversive reversal learning. Participants who provided online expectancy ratings displayed slower learning based on a hybrid model of adaptive learning and reduced reversal of SCR relative to those who did not rate expectations. Mediation analysis revealed that the effect of associative learning on SCR could be fully explained through its effects on subjective expectancy within the group who provided ratings. This suggests that the act of rating expectations reduces the speed of learning, likely through changes in attention, and that expectations directly influence arousal. Our findings indicate that higher order expectancy judgments can alter associative learning.

KEYWORDS

aversive learning, conditioning, defensive, expectancy, fear, reinforcement learning, skin conductance, threat

1 | INTRODUCTION

The study of threat-induced defensive reactions has proliferated in the last century, focusing in large part on Pavlovian classical conditioning, the critical process by which an organism learns the structure of the world and develops predictions or expectations about salient outcomes in its environment (Rescorla, 1988). Many of the defensive reactions that animals display can also be measured in humans, providing the opportunity for fundamental insights

[Correction added on December 5, 2021 after first online publication: The author's middle name was inadvertently updated as Christina A. Sandman. Now, it has been corrected. The copyright line was changed.]

through translational work. Of course, humans can also verbalize their expectations and feelings, and thus many human threat conditioning studies ask participants to provide expectancy ratings, or estimates of the probability of shock, as an additional measure of learning. People report increased shock expectancy in response to associations acquired through classical conditioning (Boddez et al., 2013; Dunsmoor et al., 2014), verbal instructions (Mertens, Braem, et al., 2018; Mertens & De Houwer, 2016), and vicarious (social) learning (Selbing & Olsson, 2019). Expectancy ratings are also clinically relevant, as they are elevated in patients with pathological anxiety (Britton et al., 2013; Chan & Lovibond, 1996) and post-traumatic stress disorder (Blechert et al., 2007), and meta-analyses of fear conditioning in anxiety indicate that of subjective ratings differ between patients and controls during both acquisition and extinction, driven in part by differences in response to the safety cue (CS−) (Duits et al., 2015). For these reasons, expectancy ratings have been argued to be a valid measure of the aversive learning process (Boddez et al., 2013), similar to other defensive reactions. But does the act of providing expectancy ratings influence adaptive learning and the expression of defensive behaviors including physiological arousal?

Although expectancies are influenced by many of the same factors as other defensive reactions and behaviors, several studies indicate that they are dissociable, suggesting they may rely on distinct mechanisms. Studies have shown dissociations between expectancies and skin conductance responses (SCR; Ohman & Soares, 1998; Schell et al., 1991; Schultz et al., 2013; Schultz & Helmstetter, 2010), eyeblinks (Perruchet, 2015; Weidemann et al., 2009, 2012), fear-potentiated startle (Kindt et al., 2009), and reaction time (Perruchet et al., 2006). These dissociations have been implicated as support for a dual process model whereby independent mechanisms support: (a) processes that are sensitive to associative learning and can exist outside of conscious awareness (e.g., defensive reactions and physiological responses) and (b) higher order processes that support propositional learning and require awareness (e.g., expectancy, subjective feelings, and instructed knowledge; LeDoux, 2013; Mineka & Öhman, 2002). Consistent with this, studies of the "Perruchet effect" (Perruchet, 2015) indicate that in partial reinforcement studies, unconditioned stimulus (US) expectancy increases after a series of trials without reinforcement (consistent with a gambler's fallacy), whereas behavioral or physiological measures decrease when US presentations are less recent. These findings have been viewed as support for the idea that expectancy ratings depend on higher order beliefs whereas automatic responses including autonomic responses and defensive reactions are driven by associative learning (but cf. Weidemann et al., 2009, Weidemann, McAndrew, et al., 2016; Weidemann, Satkunarajah, et al., 2016).

Importantly, a substantial body of work calls the dual process model into question and indicates that higher order knowledge can shape learning-related responses (for reviews, see Lovibond & Shanks, 2002; Mertens & Engelhard, 2020; Mitchell et al., 2009). Critiques of the dual process framework point to methodological problems in both study design (Lovibond & Shanks, 2002; Singh et al., 2013; Weidemann, McAndrew, et al., 2016) and analysis (Shanks, 2017) as well as publication bias (Mertens & Engelhard, 2020; Vadillo et al., 2016). Mechanistic studies of instructed fear indicate that higher order knowledge can modulate autonomic responses (Atlas et al., 2016; Atlas & Phelps, 2018; Costa et al., 2015; Grings et al., 1973; Mertens, Boddez, et al., 2018) and responses in learning-related systems (Atlas et al., 2016; Doll et al., 2009; Li, Delgado, et al., 2011). Interactions between higher order knowledge and dynamic learning suggest that propositional knowledge can shape learning through a single process. However, it is unknown whether meta-cognitive predictions in the form of expectancy ratings have similar effects as instructed knowledge on learning and defensive responses. To understand whether expectancies influence learning-related responses, we must not only compare expectancy ratings with physiological responses, but also measure physiological responses in the absence of ratings.

Our goal was to determine whether the act of providing online expectancy ratings during threat conditioning alters dynamic learning and physiological responses, relative to conditioning in the absence of ratings. We focused on SCR, which might be particularly sensitive to higher order knowledge. For example, prior work in patients with amygdala lesions suggests that expectancy ratings may indeed modulate learning as measured by SCR, consistent with an interactive model. Patients with amygdala lesions did not express conditioned responses during standard threat conditioning, as measured by differential SCR, but displayed differential SCR when they were asked to provide expectancy ratings (Coppens et al., 2009). This indicates that the act of providing expectancy ratings engaged circuits outside of the amygdala, which in turn influenced autonomic responses. Similarly, in a study of reconsolidation, a process that depends on the amygdala (Monfils et al., 2009), participants who provided expectancy ratings exhibited enhanced fear-potentiated startle during fear acquisition, but less evidence of reconsolidation, relative to participants who did not provide ratings (Warren et al., 2014). Ryan and colleagues (2021) also observed enhanced acquisition and extinction learning based on differential SCR when participants provided expectancy ratings relative evaluative ratings or learning without any ratings. Together, these studies suggest that the act of rating expectations can alter autonomic responses and engage different neural circuits to drive physiological responses, perhaps by increasing the

contribution of higher order processes such as attention or explicit beliefs. However, it is unknown whether expectancy ratings altered dynamic learning because all studies to date measured the effects of providing expectancy ratings in conditions without any stochasticity (i.e., 100% reinforcement during conditioning or 0% reinforcement during extinction). To determine whether expectancy ratings alter learning per se, we must measure the impact of expectancy ratings on dynamic associative learning. This can only be accomplished in environments that contain some volatility, such as partial reinforcement schedules and/or contingency reversals. In both partial reinforcement and reversal learning, individuals must continually update learning as a function of experience, and this requires flexible, adaptive learning, whereby individuals increase learning rates when environments become less stable (Behrens et al., 2007; Li, Schiller, et al., 2011; Mackintosh, 1975; Schiller et al., 2008). Previous work indicates that reversal learning can be modulated through instruction (Atlas, 2019; Atlas et al., 2016; Costa et al., 2015; Grings et al., 1973; Mertens & De Houwer, 2016). How do expectancy ratings shape the dynamic adaptive learning process?

We conducted a human behavioral experiment to examine whether the act of rating expectancy alters adaptive reversal learning. One group of healthy human subjects was assigned to an Expectancy Rating Group ("Rating Group"), who provided online expectancy ratings during aversive learning, and another set of participants was assigned to a Viewing Group, who learned from experience in the absence of online ratings. Both groups underwent the same aversive reversal learning task using partial reinforcement (50% reinforcement of the CS+; see Figure 1). We combined quantitative models of adaptive learning with traditional statistical analyses and multilevel mediation analyses to ask whether online expectancy ratings influence dynamic aversive learning.

We focused on a hybrid model of adaptive learning that incorporates associability, or the extent to which attention gates learning in response to changes in the environment (Mackintosh, 1975; Pearce & Hall, 1980). Consistent with previous work (Homan et al., 2019; Li, Schiller, et al., 2011; Tzovara et al., 2018; Zhang et al., 2016), we focused on SCRs as a measure of physiological arousal that is sensitive to adaptive learning, orienting, uncertainty, and attention (Ojala & Bach, 2020). We hypothesized that the timecourse of associative learning would differ between participants who provide online expectancy ratings versus those who undergo aversive learning without rating shock probability. We did not have directional hypotheses, as ratings might increase attention to contingencies (Mackintosh,

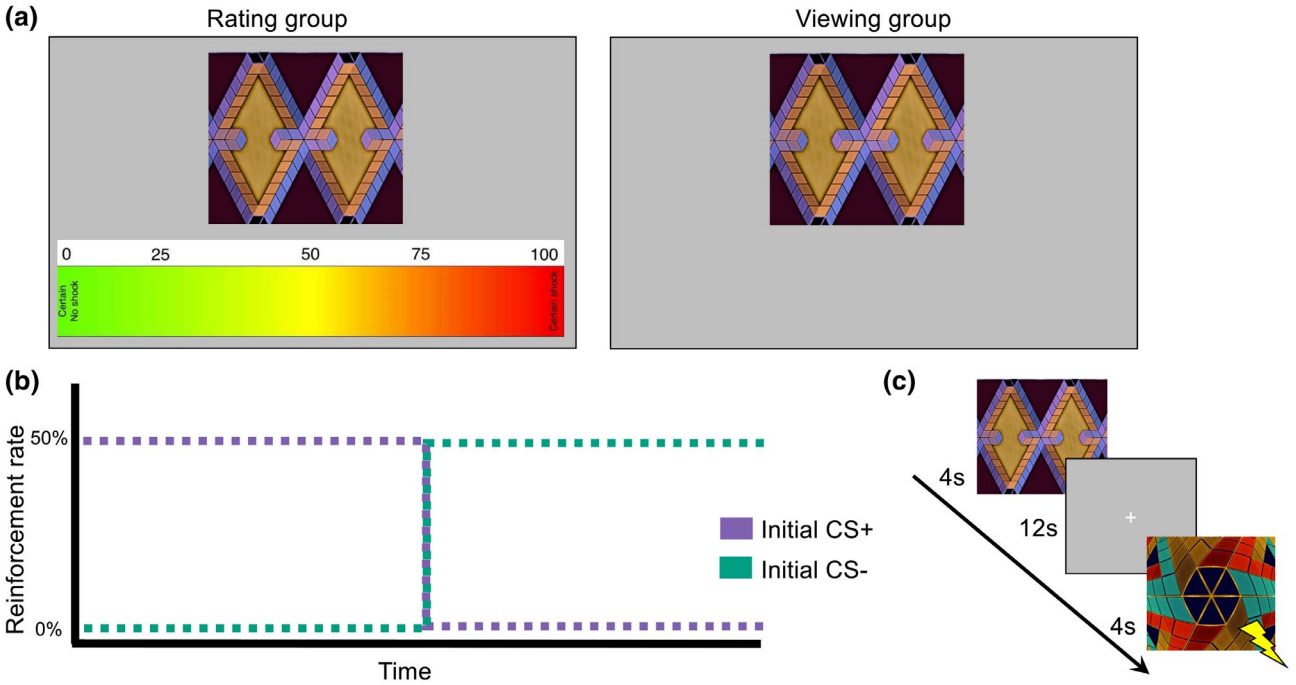

FIGURE 1 Task design. (a) Participants assigned to a Rating Group made expectancy ratings during aversive reversal learning, while Viewing Group participants viewed images and received shocks in the same task without making ratings. (b) There was a 50% reinforcement rate for the CS+ (i.e., 14 unreinforced presentations and 14 reinforced presentations). Halfway through the task, contingencies reversed and the initial CS− became the new CS+ and was reinforced with a 50% reinforcement rate for the duration of the task. (c) On each trial, the CS was presented for 4 s followed by a 12 s inter-stimulus interval. CS+ presentation coterminated with a 200 ms shock. Two stimuli were used (purple/yellow fractal or red/green fractal) and initial CS assignment was counterbalanced across participants

1975), consistent with enhanced differential responses during acquisition in previous work (Warren et al., 2014), or they might reduce overall arousal if they serve to enhance elaborative processing and reduce the threat value of the conditioned stimulus (e.g., through distraction).

2 | METHOD

2.1 | Participants

Participants were recruited from New York University and the surrounding community. All participants were required to be right-handed and fluent in English. Participants were ineligible if they had participated in an experiment using shocks within the prior six months, if they were taking medication for anxiety or depression, if they might be pregnant, or if they had a history of heart or blood pressure problems. All participants provided informed consent in accordance with the Declaration of Helsinki and as approved by the New York University Institutional Review Board (IRB # 13-9582). Eighty-nine participants provided informed consent. Seven participants did not show measurable SCR in response to the breath-hold task and were dismissed prior to the threat conditioning experiment. Technical difficulties prevented us from acquiring SCR or behavioral data from two additional participants. Eighty participants completed the task (47 Female, Mean Age = 22.28, *SD* = 3.09, missing age from 6 participants).

2.2 | Stimuli and materials

Participants underwent a threat conditioning paradigm with a 50% reinforcement rate and one reversal of CS-US contingencies (Figure 1b). Two fractal images served as conditioned stimuli (Figure 1c). Image assignment (initial CS+, initial CS−) was counterbalanced across subjects. A Grass Medical Instruments SD9 Stimulator delivered shocks to participants' right forearms (200 ms duration) through a bar electrode (BIOPAC Systems, Inc., Goleta, CA) filled with standard NaCl electrolyte gel (Signagel from Parker Laboratories, Fairfield, NJ). Shocks coterminated with the CS presentation and consisted of a 200 ms duration train of pulses at 40 Hz.

SCR were measured through shielded Ag-AgCl electrodes filled with 0.5% NaCl isotonic electrolyte gel (EL507; BIOPAC Systems, Inc., Goleta, CA) attached to the left palm. Data were recorded at a sample rate of 200 Hz using an MP-150 BIOPAC system with the AcqKnowledge software (BIOPAC Systems, Inc., Goleta, CA). Acknowledge software was used for analysis. Participants also completed the Spielberger State-Trait Anxiety Inventory STAI (form X); (Gaudry et al., 1975) and the Intolerance of Uncertainty Scale (IUS; Buhr & Dugas, 2002). The present study focuses on between-groups differences in SCR without respect to anxiety or intolerance of uncertainty.

2.3 | Procedure

2.3.1 | Skin conductance eligibility and shock calibration

Participants provided informed consent "to take part in a research study to learn more about emotion and cognition." Following consent, the participant completed questionnaires and was invited to a behavioral testing room where they were affixed with SCR and shock electrodes and completed the task on a computer. The experimenter remained in the room for the duration of the task. As all participants were right handed, participants used a mouse to record ratings with their right hand and skin conductance was recorded from the index and middle fingers of the left hand.

Prior to the main experiment, participants performed a breath holding task to ensure that they showed measurable SCR. Participants were asked to take a deep breath and hold it for three seconds. Participants whose skin conductance increased in response to the breath hold were eligible to continue. Seven participants were dismissed at this point.

Following the breath holding task, an electric shock stimulator was attached to the participant's right wrist. We calibrated the shock intensity using an ascending staircase procedure, in which intensity was increased incrementally from 20 V in 5-V increments until it reached a level that participants considered "highly annoying but not painful". Once this level was achieved, the shock remained at this intensity throughout the conditioning task ($M = 36.78$ V, $SD = 9.69$).

2.3.2 | Experimental design

Participants were randomly assigned to the Rating Group or the Viewing Group. Participants in the Rating Group were asked to rate the expected likelihood of shock during each CS presentation, using a continuous visual analogue scale ranging from "0%: sure of no shock" to "100%: sure of shock" (Figure 1a). This type of online US-expectancy rating is thought to be one of the most valid measures of contingency awareness (Boddez et al., 2013). Participants were asked to record expectancy ratings using the mouse within the 4-second CS presentation. If a participant failed to record a response during the 4-s period, the final location of the

mouse was used as the rating. On average, subjects missed fewer than 1 rating ($M = 0.23$, $SD = 0.53$). Participants in the Viewing Group viewed the CS images while making no overt responses. Participants received general contingency instructions but were not informed about the specific relationships between the CSs and outcomes. All participants were told to "try to figure out the relationship between the stimuli you see and the shocks you feel."

All participants underwent the same aversive reversal learning task with a single reversal and 50% reinforcement rate (see Figure 1). During the first 42 trials, the original CS+ coterminated with a shock (US) on 50% of CS+ trials (i.e., 14 pairings), while the CS− was never paired with the US. Thus, the pre-reversal phase included 14 CS− trials, 14 unreinforced CS+ trials, and 14 trials in which the CS+ coterminated with a shock US. Halfway through the task the contingencies reversed, such that the former CS+ became the CS− and vice versa for the last 42 trials. Reinforcement rates were the same, that is, the new CS+ (original CS−) had a 50% reinforcement rate (i.e., 14 unpaired trials, 14 trials paired with a shock) and there were 14 new CS− (original CS+) presentations. Participants were not instructed upon reversal. We used two trial orders, which were each pseudorandomized within the constraints that there were never three of the same CS image sequentially or two shocks in a row. All participants saw the same total number of each CS type (42 original CS+ trials, 42 original CS− trials) and received 28 shocks over the course of the experiment. Each CS was displayed for 4 s, followed by a 12-s inter-stimulus interval (ISI) during which a fixation cross was displayed. CS images were counterbalanced across participants.

Following the experimental task, participants answered a series of post-task questions assessing declarative knowledge of the CS-US contingencies, subjective emotional reactions to each of the CS images, and a free response item regarding any patterns or relationships observed during the study. Free responses were not included in the current analyses. Participants were then debriefed and dismissed.

2.3.3 | Skin conductance data processing

SCR data was processed in AcqKnowledge (BIOPAC Systems, Inc., Goleta, CA) and filtered with a 25-Hz low-pass FIR filter and smoothed with a Gaussian kernel of 10 samples. SCRs were measured as the base-to-peak amplitude difference for each trial during the 0.5–4.5 s window after stimulus onset. SCR amplitudes that were less than 0.02 microSiemens were considered non-responses and scored as 0. Amplitude estimates were square root transformed (Schlosberg & Stanley, 1953) and normalized relative to each participant's mean square-root-transformed US response (Ben-Shakhar, 1985; Fowles et al., 1981).

2.4 | Statistical analyses

We used ANOVAs to analyze post-task ratings as a function of Group (Rating vs. Viewing Only) and Stimulus type (Original CS+ vs. Original CS−) and used linear mixed models to analyze SCR outcomes on unreinforced trials (i.e., trials that were not paired with a US) as a function of Group, Stimulus, and Phase (Pre- vs. Post-reversal). ANOVAs were implemented using Matlab's "anovan.m" program and we modeled subject as random and nested in Group.

To account for the fact that conditioning and reversal learning are dynamic processes that occur over time as a function of reinforcement and experience, we used linear mixed effects models that model outcomes trial-by-trial and can account for effects that vary over time, as well as computational models of learning (see next section). Linear mixed models are advantageous relative to ANOVAs, which use summary statistics, average responses across trials, and cannot capture the dynamic nature of nature of learning. All linear mixed models were analyzed in R (R Core Team, 2014) using the nlme package (Bates et al., 2015). We modeled fixed effects of Group, Stimulus Type, Phase, and Trial, and all possible interactions. Slopes and intercepts were treated as random, and we modeled autoregression (AR(1)). Post-hoc pairwise comparisons were evaluated using the R package "emmeans" (Lenth, 2020).

2.4.1 | Computational modeling

While our linear mixed models provide tests of whether responses emerge gradually as a function of time (i.e., linear effect of Trial), computational models of reinforcement learning provide further insight on the dynamic learning process by testing how responses update not only as a function of trial, but also in response to specific outcomes. More specifically, they test whether expected value (EV) updates in response to a given reinforcement (r), which depends on the learning rate (α), or the speed of updating, and the prediction error (δ) which is the deviation between EV and reinforcement (r) on a given trial.

$$\delta_t = r_t - \mathrm{EV}_{(\mathrm{CS})t}$$

The learning rate governs the extent to which a prediction error causes EV to update on the next trial. Thus reinforcement learning models extend insights from mixed models by not only testing whether the magnitude of the differential response increases over time (e.g., a Cue × Trial interaction), but also how learning takes place on a trial-by-trial basis, both in response to unexpected reinforcements (e.g., shocks) and unexpected omissions of reinforcement (e.g., reversals).

Because reversals engage contextual shifts that have been previously shown to engage adaptive learning (Behrens et al., 2007; Li, Schiller, et al., 2011), we evaluated a "hybrid model" of adaptive learning (Li, Schiller, et al., 2011) to test whether act of rating expectations modulates value-based learning. In contrast to a standard Rescorla-Wagner model (Rescorla & Wagner, 1972) which assumes that learning is stable over time, hybrid models are based on the Pearce-Hall model and allow learning rates vary dynamically as a function of associability (Mackintosh, 1975; Pearce & Hall, 1980), which is inversely related to the stability of the environment. In other words, learning rates are higher in volatile environments and lower in stable environments, and depend on the recent history of prediction errors. The model is referred to as a "hybrid" model because it integrates the concept of prediction error from the Rescorla-Wagner model with associability from the Pearce-Hall model (Li, Schiller, et al., 2011). Prior work using hybrid models in aversive learning (Atlas et al., 2019; Li, Schiller, et al., 2011; Zhang et al., 2016) indicates that SCR reflects the joint combination of associability (i.e., the current dynamic learning rate (α)) and expected value (EV). These parameters update dynamically based on two free parameters, κ and η, which control the rate at which prediction errors influence outcomes:

$$\mathrm{EV}_{(\mathrm{CS})t+1} = \mathrm{EV}_{(\mathrm{CS})t} + * \alpha_{(\mathrm{CS})t} \times \delta_t$$

$$\alpha_{(\mathrm{CS})t+1} = \eta \left|\delta_t\right| + (1-\eta)\, \alpha_{(\mathrm{CS})t}$$

We fit models using Matlab's "fmincon.m" function and minimized the sum squared errors between EV or EV and associability and each participant's SCR on unreinforced trials. We also included a linear effect of time in all models, and fit models separately to participants in each group. Consistent with previous work (Atlas & Phelps, 2018; Miller et al., 1998; Wu, 1986), we used a "jack-knife" approach to model fitting and iteratively left out one subject on each iteration and fit to the remaining subjects to estimate parameters across the group. This provides a distribution of scores for statistical comparison, while reducing the noise associated with individual model fits. We compared three models: (1) a hybrid model with four free parameters (κ, η, initial EV, initial α); (2) a model that assumed an initial EV of 0.5 and an initial α of 1.0 (i.e., only κ and η were modeled as free parameters); (3) a standard Rescorla-Wagner model that included a constant learning rate which was assumed to be stable over time, and the learning rate (α) and initial EV were modeled as free parameters.

We fit models to skin conductance on unreinforced trials throughout the entire task (i.e., pre- and post-reversal) and computed Aikake's Information Criterion for each model (Akaike, 1974), which penalizes models for extra parameters. We used Bayesian model selection implemented with SPM_bms (Stephan et al., 2009) to compare models. The hybrid model with four free parameters was determined to be the best fit across participants (see Results) and we therefore use this model for inference.

We compared group differences in each parameter using two-sample *t*-tests in Matlab based on fitted parameters from jack-knife approaches. We also evaluated models fit to individuals and across the entire group for completeness.

2.4.2 | Multilevel mediation analysis

To understand the relationship between expectancy ratings and observed differential responses in autonomic arousal within the Rating Group, we tested whether expectancy ratings formally mediated conditioned responses (i.e., differential effects of CS on SCR). Multilevel mediation analyses were implemented in Matlab using the Mediation-Moderation Toolbox (Atlas et al., 2010; Wager et al., 2009). We modeled CS type (current CS+ vs. current CS−) as the input variable (*X*), SCR on unreinforced trials as the outcome variable (*Y*), and tested for mediation by expectancy ratings. On trials on which subjects did not provide an expectancy rating ($M = 1.55$, $SD = 2.36$), we used the mouse position at the end of the 4 s CS period as a measure of expectancy. We used bootstrapping to test the significance of mediation to account for non-normality of the indirect path (Shrout & Bolger, 2002).

Our mediation analysis differs from the multilevel models and quantitative learning models in that it tests responses only within the Rating Group. Furthermore, the mediation model measures the contribution of subjective expectancy ratings themselves both as a function of CS type and in relationship to evoked SCR, whereas the multilevel models and computational models examine SCR as a function of whether or not ratings were collected, rather than considering the trial-by-trial ratings themselves.

3 | RESULTS

3.1 | Post-task ratings

There were no differences between Groups in the number of perceived reversals during the task ($M_{\mathrm{RG}} = 1.79$, $SD_{\mathrm{RG}} = 0.95$; $M_{\mathrm{PG}} = 1.66$, $SD_{\mathrm{PG}} = 1.05$; $p > .5$; see Figure 2). Retrospective probability ratings indicated that participants associated a higher likelihood of shock with the original CS+ during the beginning of the study and a higher likelihood of shock with the original CS− at the end of the study

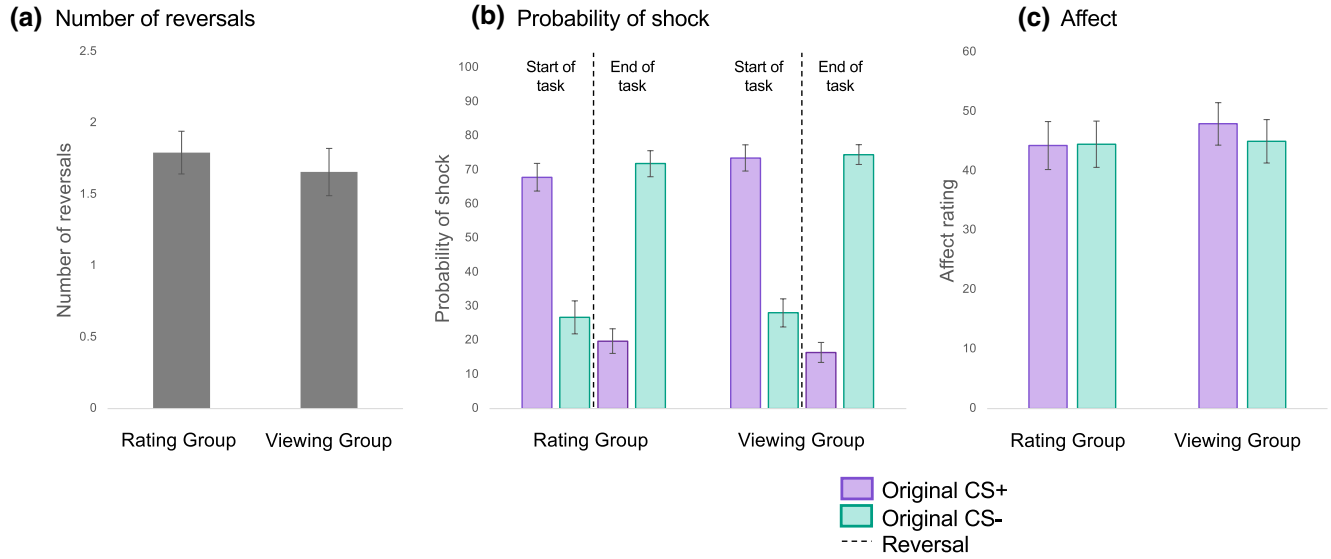

FIGURE 2 Retrospective ratings. Upon task completion, participants retrospectively rated (a) number of perceived reversals, (b) probability of shock associated with each stimulus at the beginning and at the end of the study; and (c) affect in response to each stimulus. Groups did not differ in any retrospective ratings. Error bars denote SEM

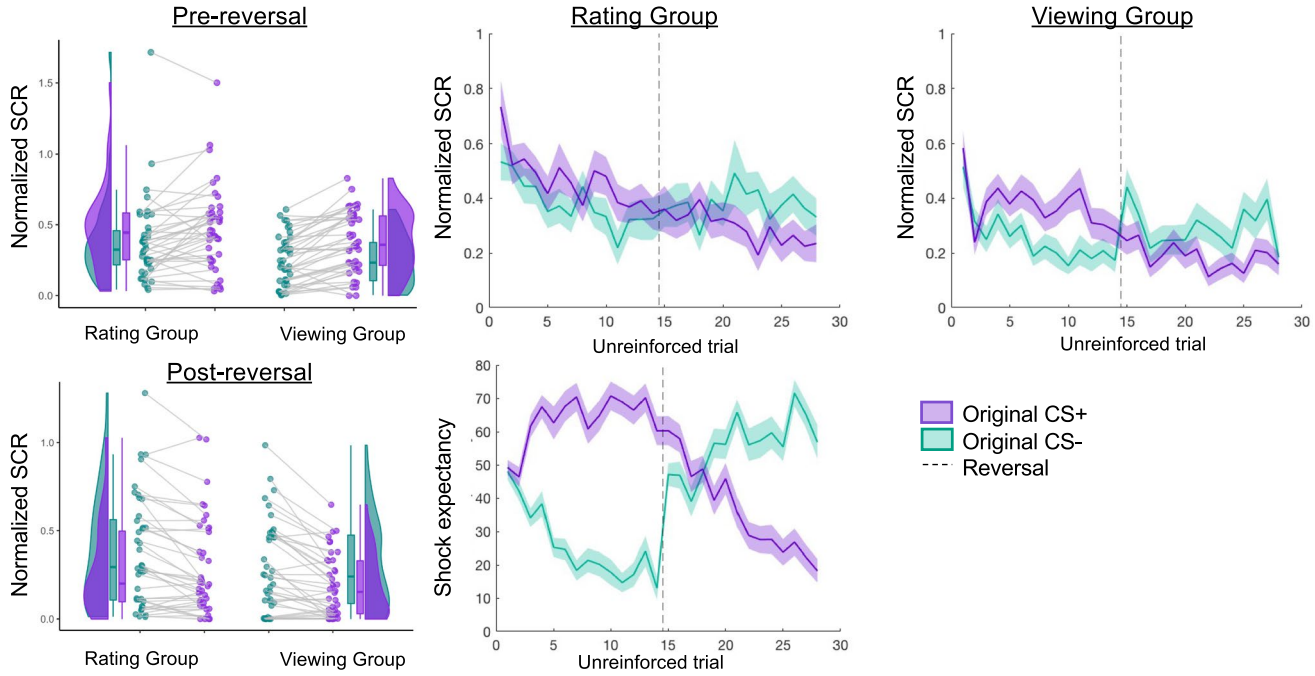

FIGURE 3 Skin conductance as a function of Group and Phase. Left: This figure illustrates skin conductance responses as a function of Stimulus prior to reversal (top) and following contingency reversals (bottom). Only the Viewing Group showed significant differences prior to reversal and a complete reversal of the differential response when contingencies changed. Middle: SCR and expectancy ratings show a similar timecourse on unreinforced trials within Rating Group participants, where responses do not reverse until several trials after the reversals. Top right: SCR in the Viewing Group reverses immediately upon contingency reversal. Raincloud plots are visualized using the R package raincloudplots (Allen et al., 2021)

(*Stimulus* × *Phase*, $F(1,291) = 136.07$, $p < .001$; $\eta^2 = 0.508$). Participants also reported higher likelihood of shock at the start of the task versus the end ($F(1,291) = 7.41$, $p = .008$; $\eta^2 = 0.007$). Probability estimates did not differ by group (all p's > .3). Retrospective affect ratings also did not differ as a function of CS Type or Group (all p's > .2).

3.2 | Participants who make expectancy ratings show sustained SCRs based on original contingencies

We used linear mixed models to examine the effects of Group, Phase (Pre- vs. Post-reversal), Stimulus (Original

CS+ vs. Original CS−), and Trial on SCR to unreinforced trials to measure whether rating expectations alters aversive learning (Figure 3). A fully specified linear mixed model failed to converge using the lme4 package in R with various optimization factors, so we used the lme package from nlme which allowed us to control for autoregression and evaluate the full model. We included fixed factors for Stimulus, Phase, Trial, and Stimulus*Phase interactions, and included random intercepts per subject and random factors for all effects except Trial. Results revealed significant main effects of Stimulus ($p = .014$), Phase ($p = .003$), and Trial ($p = .001$), as well as significant interactions between Stimulus and Phase ($p < .001$), Phase and Trial ($p < .001$), Group × Stimulus × Phase ($p = .021$), Group × Stimulus × Trial ($p = .039$), and Stimulus × Phase × Trial ($p < .001$). We were most interested in the Group × Stimulus × Phase interaction. Posthoc analyses separated by Group indicated that the Viewing Group showed a significant Stimulus × Phase interaction ($B = 0.067$, $p < .001$) whereas the Stimulus × Phase interaction was not significant in the Rating Group ($p > .1$). For complete results and results separated by group, please see Tables S1–S3. We also conducted pair-wise comparisons post-hoc between all factors using the R package emmeans (Lenth, 2020). Pairwise comparisons indicated that the key interaction was driven by the fact that only the Viewing Group showed significant differences between the CS+ and CS− prior to reversal ($p < .001$) and that responses to the original CS− increased after the reversal within the Viewing Group ($p < .001$). There were no significant differences as a function of Stimulus or Phase within the Rating Group in pairwise post-hoc tests based on adjusted p-values. For complete results of pair-wise post-hoc tests, see Table S4.

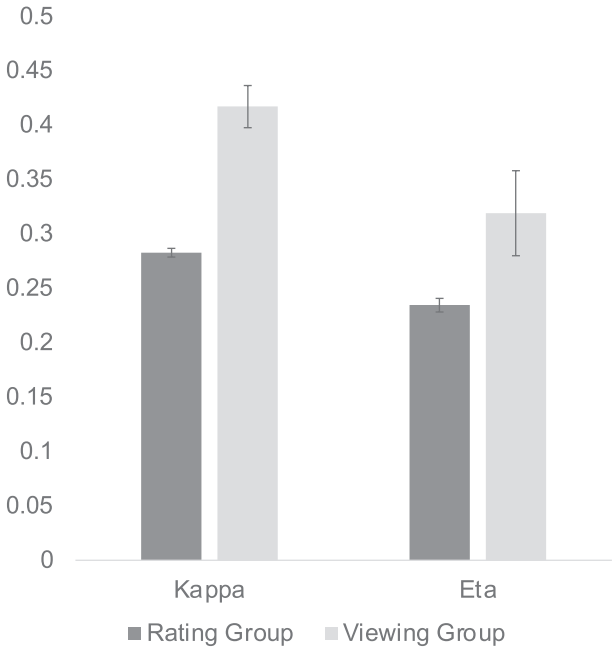

FIGURE 4 Hybrid model learning parameters differ by group. Fitting a Rescorla-Wagner model of reinforcement learning to SCR on unreinforced trials revealed higher learning rates in Viewing Group participants than participants in the Rating Group, whether fit to individuals or jack-knife estimation to iteratively fit across each group using cross validation. Learning rates depicted here are from jack-knife estimation. See Figure S1 for complete results of jack-knife estimation and fits to individuals

3.3 | Rating expectancy slows associative learning

Next, we fit dynamic learning models to SCR on unreinforced trials throughout the entire task (i.e., both pre- and post-reversal) to test whether rating expectations alters dynamic value-based learning. We fit models separately to each individual and used jack-knife estimation to iteratively leave out one subject and fit estimates to remaining participants in each group, which provides a distribution of estimates for between-group comparison that is less sensitive to noise than individual subject-level fits. We compared a standard Rescorla-Wagner model (Rescorla & Wagner, 1972) which assumes a stable learning rate across time with two variations of a hybrid model (Atlas et al., 2019; Li, Delgado, et al., 2011; Mackintosh, 1975; Pearce & Hall, 1980) that assumes that learning rates vary as a function of associability (see Method for model details). Model comparison using SPM_bms (Stephan et al., 2009) revealed that a hybrid model that included all parameters as free provided the best fit for our data, based on an exceedance probability of 0.76, versus the hybrid model that assumed starting parameters for learning rate and expected value (exceedance probability = .24) and the Rescorla-Wagner model (exceedance probability = 0). We therefore make inferences based on the hybrid model that included four free parameters.

Jack-knife estimation revealed that Viewing Group participants had significantly higher values than Rating Group participants for all parameters (Figure 4), including initial learning rate ($M_{PG} = 1.0$; $SE_{PG} = 0.0$; $M_{RG} = 0.66$; $SD_{RG} = 0.86$; $t(1,78) = -19.11$, $p < .001$), initial expected value ($M_{PG} = 0.29$; $SE_{PG} = 0.05$; $M_{RG} = 0.01$; $SE_{RG} = 0.004$; $t(1,78) = -5.49$, $p < .001$), kappa ($M_{PG} = 0.42$; $SE_{PG} = 0.02$; $M_{RG} = 0.28$; $SE_{RG} = 0.004$; $t(1,78) = -6.79$, $p < .001$), and eta ($M_{PG} = 0.42$; $SE_{PG} = 0.04$; $M_{RG} = 0.23$; $SE_{RG} = 0.006$; $t(1,78) = -2.14$, $p = .036$). Fitting separately to each individual revealed parameters that differed in the same direction although group differences were not significant in these fits, consistent with the fact that individual estimates are noisier (see Figure S1 and Table S5).

3.4 | Expectancies mediate differential responses in SCR

Finally, we tested whether subjective expectancies formally mediated the effects of conditioned cues on SCR within the Rating Group. Here, we analyzed differential effects, that is, effects of current contingencies including the reversal (Cue × Phase interactions), rather than original contingencies, and included all Rating Group participants. As shown in Figure 5, there was a significant differential effect (current CS+ > current CS−) on SCR on unreinforced trials (Path $c = 0.04(0.01)$, $p < .001$), consistent with analyses reported above. There was also a significant differential effect on expectancy (Path $a = 0.14(0.02)$, $p < .001$), such that shock expectancy was higher in response to the current CS+ than the current CS−, and an effect of expectancy on SCR, controlling for the differential effect (Path $b = 0.18(0.05)$, $p < .001$), such that skin conductance was higher when subjects expected a shock. Finally, we found that expectancy ratings fully mediated the relationship between CS and SCR (Path $a \times b = 0.02$ (0.0), $p < .001$), such that the differential effect on SCR was absent when controlling for expectancy rating (Path $c' = 0.01(0.01)$, $p > .16$; see Figure 3), indicating that subjective expectancy fully explains differential responses in SCR when individuals make expectancy ratings. We also examined the covariance between Paths *a* and *b*, since the mediation effect in multilevel mediation (i.e., $c–c'$, the difference between the direct and indirect effects) can be driven by both the product of the path coefficients and the covariance of the paths (i.e., $c–c' = a \times b + \text{cov}(a,b)$; (Kenny et al., 2003). In this case, we observed significant negative covariance (see Figure 3), suggesting that participants who showed stronger CS effects on expectancy showed weaker additional effects of expectancy on arousal. Finally, to evaluate directionality, we tested a reversed mediation, that is, whether physiological arousal mediates effects on expectancy, as proposed by models such as the somatic marker hypothesis (Poppa &

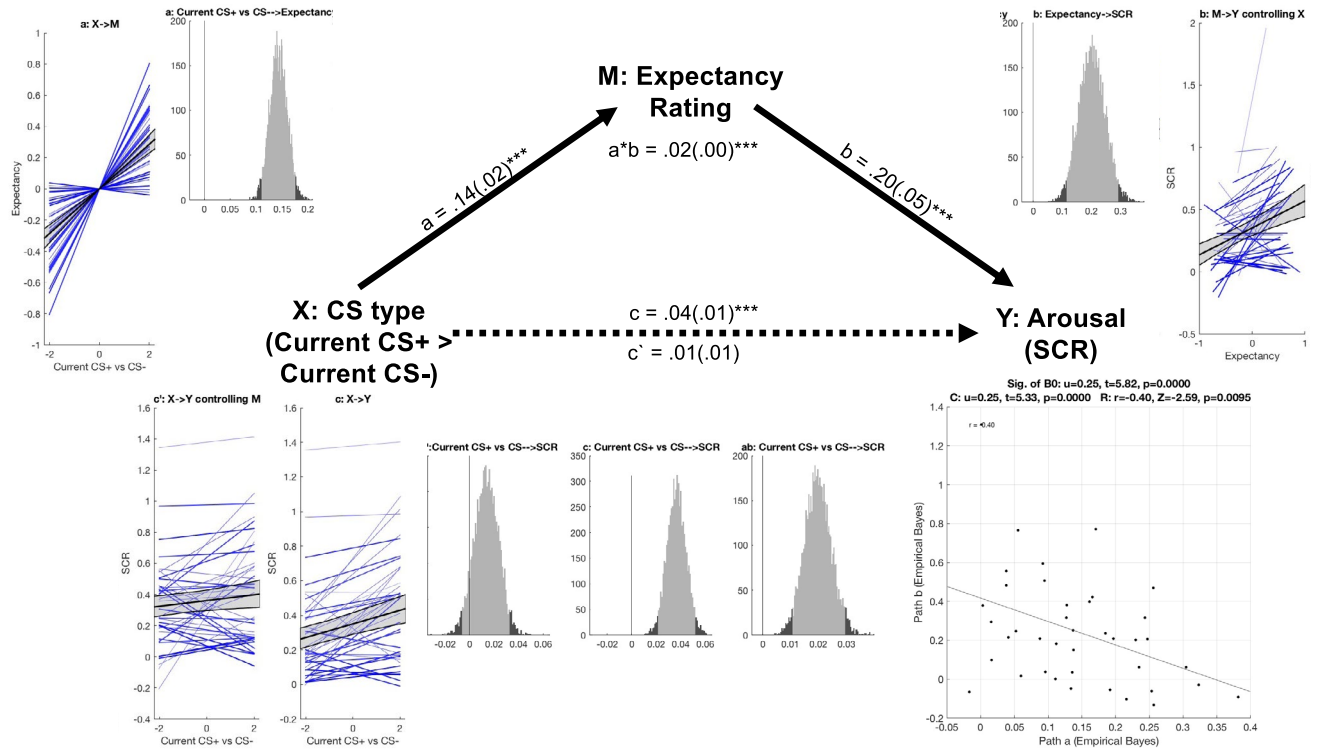

FIGURE 5 Expectancy fully mediates differential response within participants who make ratings. Multilevel mediation revealed that trial-by-trial expectancy ratings fully mediated effects of the current contingencies on SCR across participants. We used bootstrap estimation to determine the significance of the mediation effect (Shrout & Bolger, 2002). Slope plots depict individual estimates in blue lines, with the 95% confidence interval depicted in the gray shaded area that surrounds the overall group effect. Analyses were conducted in the Multilevel Mediation Moderation Toolbox (Atlas et al., 2010; Wager et al., 2009). *Upper left*: There was a significant effect of current CS contingencies on subjective expectancy (i.e., Path *a* in the mediation framework). *Upper right*: There was a significant effect of expectancy on SCR, controlling for current CS contingencies (i.e., Path *b* in the mediation framework). *Lower panel*: There was a significant direct effect of Current CS contingencies on SCR, which was non-significant when controlling for expectancy rating. There was a significant negative association between Path A and Path B coefficients, which suggests mediation was driven primarily by within-subjects effects

Bechara, 2017). When we tested whether CS effects on expectancy were mediated by SCR, we found that effect of cues on expectations was the same whether or not we control for cue effects on SCR ($c = 0.14$, STE = .02, $p < .001$; $c' = 0.14$, STE = .02, $p < .001$) and there was no evidence of mediation ($a \times b = 0$, $p > .08$). This provides further support for directionality of our model, that is, that cues affect expectations which in turn affect SCR.

3.5 | Comparing learners and non-learners

Our main analyses focus on quantitative models across all participants to assess how the act of providing expectancy ratings alters dynamic adaptive learning in response to reinforcement. However, classic approaches often average across responses regardless of time to discern whether individuals can be classified as learners (i.e., those who show elevated arousal in response to a CS+ relative to a CS−) or non-learners (those who show no differences). While this approach has known limitations (Lonsdorf et al., 2017), it can still provide convergent information regarding the overall effects of expectancy ratings on learning. We used a differential response cut-off of 0.05 μS during late acquisition to identify learners (see Figure S2). There were more learners in the Rating Group (26/40, or 65%) than the Viewing Group (18/40, or 45%), although differences were marginal based on a chi-square test ($\chi^2 = 3.23$, $p = .072$). There were no differences between Learners and Non-learners or interactions between Group and Learning in the number of perceived reversals or affect ratings for either CS, but we did observe significant interactions with Learning Status when evaluating post-task probability ratings (see Supplementary Results and Figure S3).

4 | DISCUSSION

Studies of threat conditioning have provided vast insights into the mechanisms that underlie learning and memory by measuring various defensive responses, many of which are largely conserved across human and animal models. Here, we asked whether the uniquely human act of rating expectations alters learning and threat expression in the form of autonomic arousal. We found that making expectancy ratings during aversive reversal learning slowed dynamic learning, as measured by SCR to conditioned cues. Participants who made online expectancy ratings were slower to reverse conditioned responses when contingencies changed, relative to participants who underwent the task without making ratings. Within the Rating Group, expectancy ratings fully mediated the differential response in SCR, suggesting that cues affect expectations which in turn affect SCR. In this section we discuss these findings, their relationship with previous work, and questions that should be addressed in future work.

Expectancy ratings have traditionally been included in many threat conditioning studies as a measure of learning (Boddez et al., 2013). However, rating expectations engages cognitive processes that may not occur in the absence of ratings. In particular, providing subjective ratings requires attention, decision-making, and probability inference. Our results indicate that these concurrent processes act to reduce the rate of simple associative learning, in particular making individuals less sensitive to changes in contingencies, that is, reversals. The group differences we observed are quite surprising, given the relatively high 50% reinforcement rate used. In fact, during piloting, we found that subjects who made expectancy ratings did not show SCR reversals at all when we used a 30% reinforcement rate, whereas this reinforcement rate was sufficient to induce repeated SCR reversals when expectancy ratings were not incorporated (Atlas et al., 2016). One possibility is that explicitly rating probabilities causes people to make higher order predictions that in turn guide attention. Numerous studies of associability indicate that attention can gate learning (Atlas et al., 2019; Li, Schiller, et al., 2011; Mackintosh, 1975; Pearce & Hall, 1980; Roesch et al., 2012), and that attention and learning rates decrease as an environment becomes more stable (Behrens et al., 2007; Browning et al., 2015). Individuals who explicitly believe the environment has stabilized may pay less attention to individual outcomes, and therefore may be slower to react to the contingency reversal, consistent with over-learning. Relatedly, providing online expectancy ratings might have served as a distraction in Rating Group participants, reducing their attention to changes in the environment, or might have shifted individuals from automatic to elaborative processing. It is also possible that specific features of the task (e.g., stimulus duration, stimulus discriminability) moderate the extent to which providing online expectancy ratings alters attention. For example, our CS duration was rather short, although it was sufficient for subjects to provide expectancy ratings. Longer CS presentations might allow subjects to attend to subjective expectancy without time pressure or memory load, perhaps affording greater awareness to the contingencies. Future studies should directly measure the role of attention to differentiate between these alternatives and to understand how expectancy and attention interact to shape behavior and autonomic responses.

Interactions between expectancy and attention can also explain how our findings relate to previous work examining the relationships between expectancy ratings and physiological responses during threat conditioning (Perruchet,

2015; Ryan et al., 2021; Warren et al., 2014). In two studies using 100% reinforcement, participants who made expectancy ratings demonstrated enhanced fear acquisition and enhanced extinction retention based either on startle potentiation (Warren et al., 2014) or SCR (Ryan et al., 2021). In contrast, we combined partial reinforcement and reversal learning with a hybrid model of adaptive learning and found that the process of rating expectations reduced sensitivity to prediction errors and changes in context. We believe these findings can be easily reconciled based on differences in environmental volatility. If expectancy ratings reduced attention to individual outcomes and enhanced confidence in judgments, this would be beneficial in stable environments (e.g., tasks with 100% or 0% reinforcement during acquisition and extinction, respectively) but be deleterious in stochastic environments such as the partial reinforcement reversal task we used here (or a lower reinforcement rate with more reversals, as mentioned above). Future work should systematically manipulate reinforcement rate to formally measure the impact of expectancy ratings as a function of volatility.

While the between-groups aspect of our study highlights how the inclusion of expectancy ratings can shape autonomic responses relative to learning without online ratings, our within-subjects mediation approach also provides insight on the dynamic contribution of subjective expectancy ratings themselves. We found that subjective expectations fully mediated the differential SCR response within the Rating Group. These findings are relevant in light of previous work that compared the dynamics of expectancy ratings and autonomic responses within subjects and observed meaningful dissociations. In particular, studies of the so-called Perruchet effect (Perruchet, 2015) and gamblers' fallacy (Clark et al., 2002) indicate that the recent history of association has divergent effects on eye-blink conditioning versus subjective expectancy. Unconditioned stimulus (US) recency is positively associated with the magnitude of the conditioned response, yet negatively related to expectancy; in other words, the greater the time since US presentation, the more people expect reinforcement, but the weaker the magnitude of the conditioned eye-blink. This suggests expectancy ratings are sensitive to higher order beliefs such as the gamblers' fallacy, but that conditioned eyeblink is not, which might support a dual process model. In exploratory analyses, we analyzed Rating Group responses as a function of US recency, and found that expectancy ratings did show a pattern that might be consistent with gamblers' fallacy (i.e., higher expectancy following more unreinforced CS+ presentations; see Figure S4). However, consistent with our mediation analyses and a single process model, we did *not* see dissociations between SCR and expectancy ratings; both showed similar effects of US history, and in fact trial-by-trial expectancy ratings fully mediated the effect of conditioned cues on SCR. This suggests that SCR reflects subjective expectations (at least when ratings are made) and builds on other studies that have drawn into question the generalizability of the Perruchet effect (Weidemann, Satkunarajah, et al., 2016). Importantly, our study differed from previous work on the Perruchet effect (Perruchet, 2015; Perruchet et al., 2006; Weidemann et al., 2009) in many ways: We included CS− trials, we did not instruct participants about stimulus contingencies, and we measured SCR instead of eyeblink, although the Perruchet effect has been replicated with SCR in previous work (McAndrew et al., 2012). Future studies should use a fully balanced design with and without expectancy ratings to simultaneously evaluate the Perruchet effect within participants who provide expectancy ratings and to test whether the conditioned response differs as a function of whether ratings are collected. In addition, future studies should formally compare the impact of expectancy ratings on different measures of conditioning, as SCR, eyeblink, and startle have been shown to be differentially sensitive to expectancy and awareness in previous work (Clark et al., 2002; Hamm & Vaitl, 1996; Manns et al., 2002; Weike, 2005), although there is much debate in this area (Lovibond & Shanks, 2002; Mertens & Engelhard, 2020; Schultz & Helmstetter, 2010) and all measures are sensitive to instructed reversals in the absence of reinforcement, indicating that they can be shaped by higher order knowledge (Costa et al., 2015).

Our findings are consistent with a model whereby higher order processes like instructed knowledge, meta-cognition, and executive function shape associative learning, that is, a single process model of threat learning (Grings, 1973; Mitchell et al., 2009). Which neural systems are likely to mediate the effects of expectancy rating on learning and autonomic responses? We and others have shown that instructed knowledge influences responses in the dorsolateral prefrontal cortex, which in turn shapes learning-related responses in the striatum and ventromedial prefrontal cortex (Atlas et al., 2016; Li, Delgado, et al., 2011), although reversal learning in the amygdala depended on experiential learning rather than instruction (Atlas, 2019; Atlas et al., 2016, 2019). It is possible that explicit expectations, which engage meta-cognition, act as an internally generated instruction. This seems plausible, based on prior work on amygdala lesion patients. Patients with amygdala lesions do not show SCRs during passive threat conditioning, but do show differential SCR when they make expectancy ratings (Coppens et al., 2009), indicating that higher order systems bypass the amygdala to interact with subcortical arousal circuits. Future work should directly compare neural mechanisms of threat learning with and without expectancy ratings to test whether explicitly rating expectations alters the brain responses that

mediate aversive learning, or whether aversive learning systems respond similarly irrespective of whether subjects make ratings. In addition, in the present study both groups received general contingency instructions (i.e., that there would be a relationship between the stimuli and shocks) although they were not informed about the relationship. General contingency instructions differ from pure uninstructed learning, that is, when participants are not informed about relationships between CS stimuli and shock outcomes (Mertens et al., 2021). Future work should measure whether expectancy ratings interact with instructed knowledge and directly evaluate whether both types of higher order processes have similar mechanisms and downstream effects on adaptive learning.

While traditional approaches used behavioral measures as an index of emotion, affective scientists increasingly recognize the important distinction between defensive behaviors and subjective feelings (LeDoux, 2012; LeDoux, 2013). Our findings expand this conversation by demonstrating that explicitly reporting subjective predictions can actually alter physiological arousal. Importantly, it is not known whether our findings of reductions in adaptive learning are specific to expectancy ratings, or a consequence of providing any subjective rating during learning. For instance, does rating subjective fear alter learning in a different way from rating subjective expectancy? If the impact of expectancy ratings on learning is mediated by general cognitive processes such as decision making, meta-cognition, and divided attention, then the type of rating might not matter, and all concurrent decisions might lead to reductions in adaptive learning. Alternatively, different types of concurrent ratings may have different effects on the trajectory of learning. For example, the act of rating expectancy might cause individuals to use higher order knowledge to focus on probability and reduce the impact of anxiety or fear, whereas rating subjective fear might heighten a sense of threat and anxiety and increase physiological arousal. Consistent with potential dissociations, recent findings indicate that the effect of CS-US pairing on startle potentiation and amygdala activation is not mediated by subjective expectancy, but rather by subjective fear (Mertens, Braem, et al., 2018). Future work should use between-subjects designs to directly compare how different subjective ratings impact associative learning to resolve these possibilities. Studies should also include an attentional control condition to control for the extent to which making any type of judgment influences the learning process. As mentioned above, future work should systematically manipulate and measure impacts of reinforcement rate, contingency instructions, and CS duration to test how these factors impact learning, attention, and the impact of expectancy ratings on multiple autonomic measures and defensive reactions. If all studies that incorporate subjective ratings also include a Viewing group, we will discover the conditions under which ratings impair or enhance learning as a field.

In conclusion, our study demonstrates that rating expectations alters learning in dynamic environments. We found that participants who provided expectancy ratings during aversive learning were slower to react when contingencies reversed, relative to participants who did not provide expectancy ratings. Furthermore, our mediation analyses indicated that subjective expectations directly shape autonomic responses within the group that provided ratings. Although the goal of learning is to generate predictions and expectations about outcomes in the environment, our work shows that making these expectations explicit can shape the dynamic process of learning itself, as measured by dynamic changes in anticipatory arousal. Studies of threat conditioning, threat related processing, and learning should quantify the extent to which concurrent measures like expectancy ratings may directly alter the behavior of interest.

ACKNOWLEDGEMENTS

The authors would like to thank Augustus Baker for assistance with subject recruitment, data collection, and SCR processing. This work was funded an NIH grant awarded to E.A.P. (RO1MH097085) and by the Intramural Research Program of the NIH's National Center of Complementary and Integrative Health's (PI LYA, ZIA-AT00030).

AUTHOR CONTRIBUTIONS

Lauren Y. Atlas: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Visualization; Writing—original draft; Writing—review & editing. **Christina F. Sandman:** Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Writing—original draft; Writing—review & editing. **Elizabeth A. Phelps:** Conceptualization; Funding acquisition; Project administration; Resources; Supervision; Writing—original draft; Writing—review & editing.

ORCID

Lauren Y. Atlas https://orcid.org/0000-0001-5693-4169

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., & Kievet, R. (2021). Raincloud plots: A multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Research*, *4*(63). https://doi.org/10.12688/wellcomeopenres.15191.2

Atlas, L. Y. (2019). How instructions shape aversive learning: Higher order knowledge, reversal learning, and the role of the

amygdala. *Current Opinion in Behavioral Sciences*, *26*, 121–129. https://doi.org/10.1016/j.cobeha.2018.12.008

Atlas, L. Y., Bolger, N., Lindquist, M. A., & Wager, T. D. (2010). Brain mediators of predictive cue effects on perceived pain. *Journal of Neuroscience*, *30*(39), 12964–12977. https://doi.org/10.1523/JNEUROSCI.0057-10.2010

Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D., & Phelps, E. A. (2016). Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *eLife*, *5*(MAY2016), e15192. https://doi.org/10.7554/eLife.15192

Atlas, L. Y., Doll, B. B., Li, J., Daw, N. D., & Phelps, E. A. (2019). How instructed knowledge shapes adaptive learning. *PsyArXiv*. https://doi.org/10.31234/osf.io/f4sh9

Atlas, L. Y., & Phelps, E. A. (2018). Prepared stimuli enhance aversive learning without weakening the impact of verbal instructions. *Learning and Memory*, *25*(2), 100–104. https://doi.org/10.1101/lm.046359.117

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Ben-Shakhar, G. (1985). Standardization within individuals: A simple method to neutralize individual differences in skin conductance. *Psychophysiology*, *22*(3), 292–299. https://doi.org/10.1111/j.1469-8986.1985.tb01603.x

Blechert, J., Michael, T., Vriends, N., Margraf, J., & Wilhelm, F. H. (2007). Fear conditioning in posttraumatic stress disorder: Evidence for delayed extinction of autonomic, experiential, and behavioural responses. *Behaviour Research and Therapy*, *45*(9), 2019–2033. https://doi.org/10.1016/j.brat.2007.02.012

Boddez, Y., Baeyens, F., Luyten, L., Vansteenwegen, D., Hermans, D., & Beckers, T. (2013). Rating data are underrated: Validity of US expectancy in human fear conditioning. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(2), 201–206. https://doi.org/10.1016/j.jbtep.2012.08.003

Britton, J. C., Grillon, C., Lissek, S., Norcross, M. A., Szuhany, K. L., Chen, G., Ernst, M., Nelson, E. E., Leibenluft, E., Shechner, T., & Pine, D. S. (2013). Response to learned threat: An FMRI study in adolescent and adult anxiety. *The American Journal of Psychiatry*, *170*(10), 1195–1204. https://doi.org/10.1176/appi.ajp.2013.12050651

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Publishing Group*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

Buhr, K., & Dugas, M. J. (2002). The intolerance of uncertainty scale: Psychometric properties of the English version. *Behaviour Research and Therapy*, *40*(8), 931–945. https://doi.org/10.1016/S0005-7967(01)00092-4

Chan, C. K. Y., & Lovibond, P. F. (1996). Expectancy bias in trait anxiety. *Journal of Abnormal Psychology*, *105*(4), 637–647. https://doi.org/10.1037/0021-843X.105.4.637

Clark, R. E., Manns, J. R., & Squire, L. R. (2002). Classical conditioning, awareness, and brain systems. *Trends in Cognitive Sciences*, *6*(12), 524–531. https://doi.org/10.1016/S1364-6613(02)02041-7

Coppens, E., Spruyt, A., Vandenbulcke, M., Van Paesschen, W., & Vansteenwegen, D. (2009). Classically conditioned fear responses are preserved following unilateral temporal lobectomy in humans when concurrent US-expectancy ratings are used. *Neuropsychologia*, *47*(12), 2496–2503. https://doi.org/10.1016/j.neuropsychologia.2009.04.021

Costa, V. D., Bradley, M. M., & Lang, P. J. (2015). From threat to safety: Instructed reversal of defensive reactions. *Psychophysiology*, *52*, 325–332. https://doi.org/10.1111/psyp.12359

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*(C), 74–94. https://doi.org/10.1016/j.brainres.2009.07.007

Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., Van Den Hout, M. A., & Baas, J. M. P. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. *Depression and Anxiety*, *32*(4), 239–253. https://doi.org/10.1002/da.22353

Dunsmoor, J. E., Kragel, P. A., Martin, A., & LaBar, K. S. (2014). Aversive learning modulates cortical representations of object categories. *(Cerebral Cortex New York, N.Y.: 1991)*, *24*(11), 2859–2872. https://doi.org/10.1093/cercor/bht138

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W. W., Lykken, D. T., & Venables, P. H. (1981). Committee report. Publication recommendations for electrodermal measurements. *Psychophysiology*, *18*(3), 232–239. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Gaudry, E., Vagg, P., & Spielberger, C. D. (1975). Validation of the state-trait distinction in anxiety research. *Multivariate Behavioral Research*, *10*(3), 331–341. https://doi.org/10.1207/s15327906mbr1003_6

Grings, W. W. (1973). Cognitive factors in electrodermal conditioning. *Psychological Bulletin*, *79*(3), 200–210. https://doi.org/10.1037/h0033883

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*(2), 215–221. https://doi.org/10.1037/h0034653

Hamm, A. O., & Vaitl, D. (1996). Affective learning: Awareness and aversion. *Psychophysiology*, *33*(6), 698–710. https://doi.org/10.1111/j.1469-8986.1996.tb02366.x

Homan, P., Levy, I., Feltham, E., Gordon, C., Hu, J., Li, J., Pietrzak, R. H., Southwick, S., Krystal, J. H., Harpaz-Rotem, I., & Schiller, D. (2019). Neural computations of threat in the aftermath of combat trauma. *Nature Neuroscience*, *22*(3), 470–476. https://doi.org/10.1038/s41593-018-0315-x

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, *8*(2), 115–128. https://doi.org/10.1037/1082-989X.8.2.115

Kindt, M., Soeter, M., & Vervliet, B. (2009). Beyond extinction: Erasing human fear responses and preventing the return of fear. *Nature Neuroscience*, *12*(3), 256–258. https://doi.org/10.1038/nn.2271

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron*, *73*(4), 653–676. https://doi.org/10.1016/j.neuron.2012.02.004

LeDoux, J. E. (2013). The slippery slope of fear. *Trends in Cognitive Sciences*, *7*(4), 155–156. https://doi.org/10.1016/j.tics.2013.02.004

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means.* (R package version 1.5.1) [Computer software]. https://CRAN.R-project.org/package=emmeans

Li, J., Delgado, M. R., & Phelps, E. (2011). How instructed knowledge modulates the neural systems of reward learning. *Proceedings*

of the National Academy of Sciences of the United States of America, *108*(1), 55–60. https://doi.org/10.1073/pnas.1014938108/-/DCSupplemental

Li, J., Schiller, D., Schoenbaum, G., Phelps, E., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, *14*(10), 1250–1252. https://doi.org/10.1038/nn.2904

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., Heitland, I., Hermann, A., Kuhn, M., Kruse, O., Meir Drexler, S., Meulders, A., Nees, F., Pittig, A., Richter, J., Römer, S., Shiban, Y., Schmitz, A., Straube, B., … Merz, C. J. (2017). Don't fear "fear conditioning": Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience and Biobehavioral Reviews*, *77*, 247–285. https://doi.org/10.1016/j.neubiorev.2017.02.026

Lovibond, P. F., & Shanks, D. R. (2002). The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 3–26. https://doi.org/10.1037//0097-7403.28.1.3

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. https://doi.org/10.1037/h0076778

Manns, J. R., Clark, R. E., & Squire, L. R. (2002). Standard delay eyeblink classical conditioning is independent of awareness. *Journal of Experimental Psychology: Animal Behavior Processes*, *28*(1), 32–37. https://doi.org/10.1037/0097-7403.28.1.32

McAndrew, A., Jones, F. W., McLaren, R. P., & McLaren, I. P. L. (2012). Dissociating expectancy of shock and changes in skin conductance: An investigation of the Perruchet effect using an electrodermal paradigm. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(2), 203–208. https://doi.org/10.1037/a0026718

Mertens, G., Boddez, Y., Krypotos, A.-M., & Engelhard, I. M. (2021). Human fear conditioning is moderated by stimulus contingency instructions. *Biological Psychology*, *158*, e107994. https://doi.org/10.1016/j.biopsycho.2020.107994

Mertens, G., Boddez, Y., Sevenster, D., Engelhard, I. M., & De Houwer, J. (2018). A review on the effects of verbal instructions in human fear conditioning: Empirical findings, theoretical considerations, and future directions. *Biological Psychology*, *137*(October, 2017), 49–64. https://doi.org/10.1016/j.biopsycho.2018.07.002

Mertens, G., Braem, S., Kuhn, M., Lonsdorf, T. B., van den Hout, M. A., & Engelhard, I. M. (2018). Does US expectancy mediate the additive effects of CS-US pairings on contingency instructions? Results from subjective, psychophysiological and neural measures. *Behaviour Research and Therapy*, *110*, 41–46. https://doi.org/10.1016/j.brat.2018.09.003

Mertens, G., & De Houwer, J. (2016). Potentiation of the startle reflex is in line with contingency reversal instructions rather than the conditioning history. *Biological Psychology*, *113*, 91–99. https://doi.org/10.1016/j.biopsycho.2015.11.014

Mertens, G., & Engelhard, I. M. (2020). A systematic review and meta-analysis of the evidence for unaware fear conditioning. *Neuroscience & Biobehavioral Reviews*, *108*, 254–268. https://doi.org/10.1016/j.neubiorev.2019.11.012

Miller, J., Patterson, T., & Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. *Psychophysiology*, *35*(1), 99–115. https://doi.org/10.1111/1469-8986.3510099

Mineka, S., & Öhman, A. (2002). Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological Psychiatry*, *52*(10), 927–937. https://doi.org/10.1016/S0006-3223(02)01669-4

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(2), 183–198. https://doi.org/10.1017/S0140525X09000855

Monfils, M.-H., Cowansage, K. K., Klann, E., & LeDoux, J. E. (2009). Extinction-reconsolidation boundaries: Key to persistent attenuation of fear memories. *Science*, *324*(5929), 951. https://doi.org/10.1126/science.1167975

Ohman, A., & Soares, J. J. (1998). Emotional conditioning to masked stimuli: Expectancies for aversive outcomes following nonrecognized fear-relevant stimuli. *Journal of Experimental Psychology: General*, *127*(1), 69–82. https://doi.org/10.1037/0096-3445.127.1.69

Ojala, K. E., & Bach, D. R. (2020). Measuring learning in human classical threat conditioning: Translational, cognitive and methodological considerations. *Neuroscience & Biobehavioral Reviews*, *114*, 96–112. https://doi.org/10.1016/j.neubiorev.2020.04.019

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Perruchet, P. (2015). Dissociating conscious expectancies from automatic link formation in associative learning: A review on the so-called Perruchet effect. *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*(2), 105–127. https://doi.org/10.1037/xan0000060

Perruchet, P., Cleeremans, A., & Destrebecqz, A. (2006). Dissociating the effects of automatic activation and explicit expectancy on reaction times in a simple associative learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(5), 955–965. https://doi.org/10.1037/0278-7393.32.5.955

Poppa, T., & Bechara, A. (2017). ScienceDirect The somatic marker hypothesis: Revisiting the role of the "body-loop" in decision-making. *Current Opinion in Behavioral Sciences*, *19*, 61–66. https://doi.org/10.1016/j.cobeha.2017.10.007

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rescorla, R. A. (1988). Pavlovian conditioning. It's not what you think it is. *The American Psychologist*, *43*(3), 151–160. https://doi.org/10.1037/0003-066X.43.3.151

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black, & W. Prokasky (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). Appleton-Century-Crofts.

Roesch, M. R., Esber, G. R., Li, J., Daw, N. D., & Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *European Journal of Neuroscience*, *35*(7), 1190–1200. https://doi.org/10.1111/j.1460-9568.2011.07986.x

Ryan, K. M., Neumann, D. L., & Waters, A. M. (2021). Does the assessment of different combinations of within-phase subjective measures influence electrodermal responding and between-phase subjective ratings during fear conditioning and extinction experiments? *Biological Psychology*, *162*, e108085. https://doi.org/10.1016/j.biopsycho.2021.108085

Schell, A. M., Dawson, M. E., & Marinkovic, K. (1991). Effects of potentially phobic conditioned stimuli on retention, reconditioning, and extinction of the conditioned skin conductance

response. *Psychophysiology*, *28*(2), 140–153. https://doi.org/10.1111/j.1469-8986.1991.tb00403.x

Schiller, D., Levy, I., Niv, Y., Ledoux, J. E., & Phelps, E. (2008). From fear to safety and back: Reversal of fear in the human brain. *Journal of Neuroscience*, *28*(45), 11517–11525. https://doi.org/10.1523/JNEUROSCI.2265-08.2008

Schlosberg, H., & Stanley, W. C. (1953). A simple test of the normality of twenty-four distributions of electrical skin conductance. *Science*, *117*(3028), 35–37. https://doi.org/10.1126/science.117.3028.35

Schultz, D. H., Balderston, N. L., Geiger, J. A., & Helmstetter, F. J. (2013). Dissociation between implicit and explicit responses in postconditioning UCS revaluation after fear conditioning in humans. *Behavioral Neuroscience*, *127*(3), 357–368. https://doi.org/10.1037/a0032742

Schultz, D. H., & Helmstetter, F. J. (2010). Classical conditioning of autonomic fear responses is independent of contingency awareness. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(4), 495–500. https://doi.org/10.1037/a0020263

Selbing, I., & Olsson, A. (2019). Anxious behaviour in a demonstrator affects observational learning. *Scientific Reports*, *9*, 9181. https://doi.org/10.1038/s41598-019-45613-1

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*(3), 752–775. https://doi.org/10.3758/s13423-016-1170-y

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*(4), 422–445. https://doi.org/10.1037//1082-989X.7.4.422

Singh, K., Dawson, M. E., Schell, A. M., Courtney, C. G., & Payne, A. F. H. (2013). Can human autonomic classical conditioning occur without contingency awareness? The critical importance of the trial sequence. *Biological Psychology*, *93*(1), 197–205. https://doi.org/10.1016/j.biopsycho.2013.02.007

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Tzovara, A., Korn, C. W., & Bach, D. R. (2018). Human Pavlovian fear conditioning conforms to probabilistic learning. *PLOS Computational Biology*, *14*(8), e1006243. https://doi.org/10.1371/journal.pcbi.1006243

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*(1), 87–102. https://doi.org/10.3758/s13423-015-0892-6

Wager, T. D., Waugh, C. E., Lindquist, M., Noll, D. C., Fredrickson, B. L., & Taylor, S. F. (2009). Brain mediators of cardiovascular responses to social threat. *NeuroImage*, *47*(3), 821–835. https://doi.org/10.1016/j.neuroimage.2009.05.043

Warren, V. T., Anderson, K. M., Kwon, C., Bosshardt, L., Jovanovic, T., Bradley, B., & Norrholm, S. D. (2014). Human fear extinction and return of fear using reconsolidation update mechanisms: The contribution of on-line expectancy ratings. *Neurobiology of Learning and Memory*, *113*, 165–173. https://doi.org/10.1016/j.nlm.2013.10.014

Weidemann, G., Broderick, J., Lovibond, P. F., & Mitchell, C. J. (2012). Both trace and delay conditioned eyeblink responding can be dissociated from outcome expectancy. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(1), 1–10. https://doi.org/10.1037/a0024411

Weidemann, G., McAndrew, A., Livesey, E. J., & McLaren, I. P. L. (2016). Evidence for multiple processes contributing to the Perruchet effect: Response priming and associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition*, *42*(4), 366–379. https://doi.org/10.1037/xan0000117

Weidemann, G., Satkunarajah, M., & Lovibond, P. F. (2016). I think, therefore eyeblink: The importance of contingency awareness in conditioning. *Psychological Science*, *27*(4), 467–475. https://doi.org/10.1177/0956797615625973

Weidemann, G., Tangen, J. M., Lovibond, P. F., & Mitchell, C. J. (2009). Is Perruchet's dissociation between eyeblink conditioned responding and outcome expectancy evidence for two learning systems? *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(2), 169–176. https://doi.org/10.1037/a0013294

Weike, A. I. (2005). Fear conditioning following unilateral temporal lobectomy: Dissociation of conditioned startle potentiation and autonomic learning. *Journal of Neuroscience*, *25*(48), 11117–11124. https://doi.org/10.1523/JNEUROSCI.2032-05.2005

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression. *Annals of Statistics*, *14*(4), 1343–1350. https://doi.org/10.1214/aos/1176350161

Zhang, S., Mano, H., Ganesh, G., Robbins, T., & Seymour, B. (2016). dissociable learning processes underlie human pain conditioning. *Current Biology*, *26*(1), 1–7. https://doi.org/10.1016/j.cub.2015.10.066

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

FIGURE S1 Learning parameters by estimation procedure
FIGURE S2 Responses during late acquisition used to define learners and non-learners
FIGURE S3 Post task ratings as a function of learning status
FIGURE S4 Perruchet effect analysis within Rating Group participants
TABLE S1 Multilevel model measuring the effects of expectancy rating on aversive reversal learning
TABLE S2 Multilevel model measuring aversive reversal learning within Rating Group participants
TABLE S3 Multilevel model measuring aversive reversal learning within Viewing Group participants
TABLE S4 Posthoc pairwise comparisons from model measuring the effects of expectancy rating on aversive reversal learning
TABLE S5 Results of hybrid model fit to individuals

How to cite this article: Atlas, L. Y., Sandman, C. F., & Phelps, E. A. (2022). Rating expectations can slow aversive reversal learning. *Psychophysiology*, 59, e13979. https://doi.org/10.1111/psyp.13979