Manipulating expectancy violations to strengthen the efficacy of human fear extinction

Stemerding, L.E.; van Ast, V.A.; Kindt, M.

DOI
10.1016/j.brat.2023.104319

Publication date
2023

Document Version
Final published version

Published in
Behaviour Research and Therapy

Behaviour Research and Therapy 165 (2023) 104319

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Manipulating expectancy violations to strengthen the efficacy of human fear extinction

Lotte E. Stemerding *, Vanessa A. van Ast, Merel Kindt **

Department of Clinical Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT, Amsterdam, the Netherlands

A B S T R A C T

Recent theoretical and clinical articles have emphasized a role for expectancy violations in improving the effectiveness of exposure therapy. Expectancy violations are critical to extinction learning and strengthening these violations has been suggested to improve the formation and retention of extinction memories, which should result in lasting symptom reductions after treatment. However, more detailed mechanistic insights in this process are needed to better inform clinical interventions. In two separate fear-conditioning experiments, we investigated whether stronger expectancy violations (Exp1) or fostering awareness of expectancy violations (Exp2) during extinction could reduce the subsequent return of fear. We measured fear potentiated startle (FPS) and skin conductance responses (SCR) as physiological indices of fear, and US expectancy ratings to assess our manipulations. While we successfully created stronger expectancy violations in Exp1, we found no evidence that these stronger violations reduced the return of fear at test. Interestingly, fostering awareness of violations (Exp2) reduced differential SCRs, but not FPS responses. These findings provide novel insights into the effect of US expectancies on fear extinction in the lab, but they also illustrate the complexity of capturing clinically relevant processes of change with fear-conditioning studies.

1. Introduction

Exposure therapy is one of the most widely used and effective therapies for fear and anxiety disorders (Deacon & Abramowitz, 2004; Norton & Price, 2007), yet not all patients are responsive and some experience relapse after initial reduction of symptoms (Loerinc et al., 2015; Springer et al., 2018). During exposure, patients experience a disconfirmation of their maladaptive beliefs, such as a confrontation with a feared object or situation in a safe setting. Despite some significant improvements in the procedural understandings of exposure therapy, it remains relatively unknown how learning during treatment can be strengthened, which should subsequently result in improved retention of what has been learned. In the past decades there have been various hypotheses regarding the working mechanism of exposure therapy, from learning a competing response that interferes with the anxiety response (Wolpe, 1968) to acquiring corrective information through habituation of the fear response (Foa & Kozak, 1986). The most recent and currently dominant theory – the inhibitory retrieval model of exposure – is based on the assumption that during exposure a new inhibitory memory is created that competes with the original fear memory for behavioural expression (Bouton, 1993). As a consequence, the inhibitory retrieval model posits that exposure outcomes can be improved by strengthening this inhibitory memory and its retrievability (Craske et al., 2014, 2022). The formation of inhibitory memories is driven by expectancy violations that occur when predictions about outcomes are violated, and it is therefore suggested that optimizing these violations results in more effective treatments (Craske et al., 2014, 2022). While the role of expectancy violations in learning during treatment is unquestioned, there is limited empirical evidence for a causal relation between expectancy violations and symptom reduction. To better understand how exposure treatment can be improved, we need more detailed insight into whether optimizing expectancy violations indeed results in lasting reductions in fear.

The idea that expectancy violations play an important role in behavioural change has been extensively featured in various cognitive models of mental disorders and their treatment, but it has been formalized in models of associative learning. These models maintain that learning is driven by prediction errors, and that the strength of learning is – among other factors – governed by the size of the prediction error (Pearce & Hall, 1980; Rescorla & Wagner, 1972). Prediction errors are defined as the discrepancy between what is actually happening and what is predicted to happen (based on memory), and if this discrepancy is larger, more can be learned. For lack of a better measure or manipulation in humans, expectancy violations are often used as a proxy for prediction errors in exposure. It remains unclear, however, whether these violations of threat expectancy govern reductions in fear

* Corresponding author.
** Corresponding author.
E-mail addresses: l.e.stemerding@uva.nl (L.E. Stemerding), m.kindt@uva.nl (M. Kindt).

https://doi.org/10.1016/j.brat.2023.104319
Received 9 November 2022; Received in revised form 6 April 2023; Accepted 14 April 2023
Available online 14 April 2023

behaviour in the same fashion as prediction errors are deemed to govern associative learning in animal models. While learning models describe only one latent variable of learning known as associative strength, we know that threat expectancies and fear responses may reflect entirely different entities of learning – one may *know* that a threat is not likely to occur, but still *feel* very fearful. Furthermore, while associative learning models posit that larger expectancy violations result in faster updating, they make no predictions about the long-term benefits of what has been learned. Thus, while the proposition that increasing expectancy violations strengthens learning is originally rooted in associative learning theory, it is not evident that optimizing expectancy violations indeed results in stronger learning and better retention of what was learned.

Clinical studies investigating the effect of expectancy violations on exposure outcomes are scarce and have utilized widely varying methods to measure or manipulate expectancy violation. For example, an exposure intervention that continued until expectancies were below 5% was found to perform better than regular exposure (Deacon et al., 2013). It should be noted, however, that participants in the experimental group also received substantially more exposure trials, complicating an equal comparison between these groups. Other studies have investigated a range of alternative manipulations during exposure that aimed to optimize expectancy violations as well, such as the removal of safety signals or starting exposure with the most feared exercise, but show mixed effects (Kircanski et al., 2012; Lang & Craske, 2000; Meulders et al., 2016). Given the diversity in manipulations, it is not evident that these effects unambiguously result from the manipulation of expectancy violations (e.g., removing safety signals may also promote learning through other mechanisms). On an individual level, absolute expectancy violations during exposure were not found to predict treatment outcome (de Kleine et al., 2017; Pittig et al., 2022), yet patients who were more likely to update their outcome expectations showed a stronger reduction of symptoms (Pittig et al., 2022). Expectation updates were calculated as the difference between the outcome expectation before a given exposure exercise, and the outcome expectation if the participant were to repeat the exercise. Hence, the success of exposure may be better predicted by the degree to which threat expectations change rather than absolute expectancy violations. In line with these findings, it was recently recommended that boosting awareness of expectancy violations during and after treatment may also strengthen learning and improve retention (Craske et al., 2022; McGlade & Craske, 2021). In sum, there is initial support for the hypothesis that expectancy violations benefit exposure treatment, but the use of different manipulations complicates determining whether the observed changes are the causal effect of expectancy violations.

In contrast to clinical research, the use of fear-conditioning studies can provide more controlled investigations of the effect of expectancy violations on learning, as it allows for full control over learning history and trial-to-trial tracking of the updating of expectancy ratings and fear responses. In fear conditioning, participants learn that one conditioned stimulus (the CS+) is followed by an aversive outcome (the unconditioned stimulus; US) whereas another control stimulus (the CS-) is not followed by the US. Fear extinction is then procedurally similar to exposure and consists of a phase in which the CS+ is no longer followed by the US, typically resulting in a reduction of the conditioned response. Expectancy violations during extinction can be manipulated by increasing the expected probability that the US occurs. Because expectancy violations reflect the difference between the expectation of the outcome and the actual experience of the outcome, a higher expectation that the US will occur should result in a stronger expectancy violation when the US is omitted. Initial support for a role of expectancy violations in extinction learning came from studies on compound stimulus presentation. The rationale behind these studies is that the expectation of the US, and therefore the expectancy violation during extinction, can be increased when two different CSs first undergo extinction individually, and are then combined to again increase the expectation of the US (Rescorla, 2006). Indeed, individuals that received compound extinction showed better extinction retention compared to a group that received an equal number of single extinction trials (Coelho et al., 2015; Culver et al., 2015). In another investigation of the effect of expectancy violations, Gromer et al. (2022) found some support for an effect of expectancy violations on extinction retention and generalisation in an online study. During acquisition, participants viewed eight different CSs (spheres that differ in size) with different probabilities of being followed by the US. Participants then underwent extinction training either viewing only the four CSs with a low US probability (low expectancy group) or viewing only the four CSs with high US probability (high expectancy group). Confirming a role for expectancy violations in extinction learning, expectancy ratings at test were reduced in the high expectancy group, but only for the CSs with a high US probability. Yet critically, participants in the low expectancy group never viewed these high probability stimuli during extinction. The results can therefore also be interpreted as the low expectancy group showing poor generalisation to CSs with a higher reinforcement rate. Furthermore, the groups showed no differences on threat ratings, suggesting that stronger expectancy violations do not influence more affective components of learning.

The examples above illustrate that, even using a fear-conditioning paradigm, isolating the effect of expectancy violations on extinction learning is challenging. Unless extinction is instructed, the participants' threat expectancies are dependent on what was learned during acquisition. In aiming to manipulate threat expectancies between groups, the reinforcement of the CS + should thus already differ between the groups during acquisition. This, however, also affects the strength of the original memory, which complicates a comparison at test. The fear response at test is believed to reflect a competition between the strength of the original fear memory and the strength of the extinction memory. To compare manipulations of extinction learning, the strength of the original fear memory should thus be equal between groups. One way to circumvent this problem is to work with multiple CSs with different reinforcement rates during acquisition, and to only show a selection of these stimuli (high or low threat expectancy) during extinction (see also Gromer et al., 2022). However, a limitation of this approach is that participants may consider the various CSs during acquisition as unrelated to each other. If participants then only view a selection of these stimuli during extinction (e.g., the ones with high reinforcement), nothing is learned about the other set of stimuli (e.g., the ones with low reinforcement) and the test phase essentially becomes a generalisation test. To truly isolate the effects of expectancy violations during extinction without affecting the strength of the fear memory trace, we need to ensure that the acquisition and test phases are the same between groups, and that all participants view the same CS during extinction.

In Experiment 1 of the current study, we aimed to address these potential concerns and isolated the effect of expectancy violations by using two contexts to signal the reinforcement rate of the CS+ (100% or 50%) during acquisition. These contexts thus acted as occasion setters, which were not alone predictive of the US but modulated the probability that the US would follow the CS+ (e.g., Craske et al., 2022; Zbozinek et al., 2021). In the extinction phase, the context was manipulated between two groups with the aim of creating high (100%) and low (50%) expected US probabilities. This should subsequently result in strong and weak expectancy violations upon the non-occurrence of the US. Importantly, all participants viewed the same CS during extinction, only the context differed. To investigate the return of fear, the CS was presented in a novel context during a test phase one day later. We hypothesized that participants in the strong expectancy violation group would show improved extinction retention. In Experiment 2 we tested a more clinically feasible manipulation by explicitly fostering awareness of expectancy violations during extinction, which has also been suggested to improve extinction retention. In both experiments, we measured fear-potentiated startle (FPS) and skin conductance responses (SCR) as indices of fear, and threat expectancy ratings to test the effectiveness of our manipulations. Although we were specifically

interested in the effect of expectancy violations on durable reductions in physiological responses, we also investigated whether stronger expectancy violations affected the expectation of the US one day later.

2. Methods

2.1. General measures and materials

2.1.1. Conditioned stimuli

The conditioned stimuli (CSs) were two female avatars that were presented in the middle of the computer screen. Both CSs were always presented on a screen-filling context (a forest or a riverside, counterbalanced). In Experiment 1 we used these contexts to signal the differential reinforcement rate of the CS during acquisition and extinction (50% versus 100%). A third context (a castle) was presented during the test phase. In Experiment 2 the contexts no longer played a role in signaling the reinforcement rate, and were a fixed part of the CS, such that one avatar was always presented on the forest context and the other always on the river context, in all phases of the experiment. The CSs were counterbalanced across conditions (CS+, CS-).

2.1.2. Unconditioned stimulus

The unconditioned stimulus (US) consisted of an aversive electrical stimulus delivered to the top of the left wrist. The US was delivered by a Digitimer DS71 (Welwyn Garden City, UK) through two 20 × 25 mm Ag/AgCl electrodes. The intensity was individually determined at the start of the experiment to be uncomfortable but not painful (see *Procedure*) and was set to a minimum of 2 mA and a maximum of 70 mA.

2.1.3. Fear-potentiated startle

Fear-potentiated startle (FPS) responses were measured with three 7 mm electromyography (EMG) electrodes filled with electrode gel. Two electrodes were placed on the left orbicularis oculi muscle, and a third ground electrode was placed on the forehead. A short loud noise (50 ms, 105 dB) was administered binaurally through headphones to trigger startle responses. The EMG electrodes were connected to a custom-made amplifier with a 5–1000 Hz bandwidth and an input resistance of 1 GΩ. The EMG signal was recorded and pre-processed using the in-house software VSRRP98 (due to a change of labs, data from 8 participants were collected using the in-house software FysioRecorder, with the same settings). The raw signal was sampled at 1000 Hz and then digitized, notch filtered at 50 Hz and bandpass filtered between 28 and 500 Hz with a 4th order Butterworth filter. Startle response values were calculated as a peak EMG value in a 0–200 ms window after startle probe onset. Missing FPS responses were linearly interpolated within stimulus type and day (0.34% in Exp1, 0.16% in Exp2). Raw FPS responses were standardized within participant, but across day and stimulus type, to aid comparison between the days.

2.1.4. Skin conductance

Skin conductance responses (SCR) were measured using two 16 × 20 mm Ag/AgCl electrodes that were attached to the medial phalanx surfaces of the index and middle finger. The skin conductance signal was recorded and digitized using the in-house software VSRRP98 and sampled at 1000 Hz (due to a change of labs, data from 8 participants were collected using the in-house software FysioRecorder, with the same settings). The raw skin conductance signal was filtered using a 1st order Butterworth filter with a cut-off frequency of 1 Hz (Boucsein et al., 2012) in MATLAB version 2022a (The MathWorks Inc., 2022). SCRs were scored manually (and blind to condition) by identifying the first SCR onset in a 900–4000 ms window post CS onset. SCRs were calculated as the difference between this first onset and the first subsequent peak. All responses smaller than 0.02 μS were scored as zero and included in the analyses. SCRs were log-transformed to improve normality and standardized within participant and across day and stimulus type.

2.1.5. Expectancy ratings

Expectancy ratings were measured on each trial on a scale from 0 ("I will certainly not receive an electrical stimulus") to 100 ("I will certainly receive an electrical stimulus"), with 50 as middle point ("I am uncertain or there is a 50% chance"). Expectancy ratings were provided by moving a cursor over the scale, and to click to confirm. Participants were instructed they had approximately 5 s to rate their expectancy. The last location of the cursor was saved and used for data analysis if participants did not click to confirm their expectancy.

2.1.6. Preregistrations

Both experiments were preregistered at the Open Science Framework (Experiment 1: https://osf.io/w2hys, Experiment 2: https://osf.io/fc79p).

2.2. Methods Experiment 1

2.2.1. Participants

Participants were sixty healthy university students aged between 18 and 35, with thirty participants in each group. Groups were balanced on gender, age, and trait anxiety scores. Exclusion criteria were diagnosed mental disorders, epilepsy, pregnancy, colour-blindness, hearing problems, and previous participation in fear-conditioning studies. All exclusion criteria were checked before booking the appointments, based on self-report. Participants were further excluded if they rated the maximum US intensity (70 mA) as less than 7 (see *Procedure*). All participants were randomly assigned to one of the two groups, and all received research credits or a financial reward (€15/hour) for the time spent in the study. The study was approved by the ethics board of the University of Amsterdam and all participants signed informed consent before taking part.

2.2.2. Bayesian sequential updating

We used an adapted Bayesian sequential updating paradigm (Schönbrodt & Wagenmakers, 2018), where we tested a minimum of 40 participants (20 per group) and a maximum of 60 participants (30 per group). Because acquisition of fear responses is essential to investigate differences in extinction learning, we first specified a stopping rule for a strong effect of fear acquisition in both groups. We further determined stopping rules for evidence in favour of and against our effect of interest (a Stimulus × Group interaction on the first test trials), with a weaker requirement against the effect, given that it is more difficult to find evidence for the null. The maximum sample size of 60 participants was based on financial and practical limitations. This resulted in the following stopping rules:

1. To test for fear acquisition, we performed a Bayesian *t*-test on the acquisition data (day 1), comparing the average of the last two CS + trials to the average of the last two CS- trials, separately for each group. We continued testing until we observed a $BF_{10} > 7$ for a difference between CS+ and CS- responding in both groups.
2. If the acquisition data implied stopping, we performed a Stimulus (CS+, CS-) x Group (Strong EV, Weak EV) Bayesian repeated measures ANOVA on the first trial of the renewal phase (day 3). We would continue testing until we observed a $BF_{inc} > 7$ or $< 1/3$ for the Stimulus × Group interaction.
3. If the renewal data implied stopping, we performed the same analyses on the first reinstatement trial and would continue testing until we observed a $BF_{inc} > 7$ or $< 1/3$ for the Stimulus × Group interaction.

We performed interim checks at N = 40 and N = 50 but did not reach the threshold for acquisition at either check. We therefore continued testing until our maximum sample size was reached. The results of the interim analyses can be found in the supplementary material.

2.2.3. Design and experimental task

We employed a within-between subjects' design and included two groups, a strong expectancy violation and weak expectancy violation group (see Fig. 1). The experiment consisted of three sessions on three consecutive days. The first (acquisition) and third (test) day were the same for both groups. We opted for a design using two different contexts within subjects to signal the reinforcement rate of the CS + because it enabled us to manipulate the outcome expectancies during extinction between two groups while keeping acquisition the same. This is essential for a just comparison of return of fear between the groups, as differences in acquisition may affect the return of fear. During fear acquisition the CS+ and CS- were both presented in two different contexts. The CS- was never followed by the US in either context, whereas the CS+ was 100% reinforced in one context and 50% reinforced in the other context. The CS+ and CS- were presented 8 times each, 4 times per context. The presentations of CS+ and CS- in each context were semi-randomized, such that all four combinations of CS and context were shown at least once before showing the same combination again. A noise alone (NA) startle probe was presented 8 times randomly during the ITI to match the number of CS+ and CS- trials. To manipulate the expectation of the US between the groups during the extinction phase, extinction took place in either the 100% context (strong EV group) or the 50% context (weak EV group). Each CS was presented 12 times without reinforcement during extinction. All participants were tested for return of fear on the third day. Both CSs were presented in the novel context for 8 times as a renewal test and re-extinction procedure. To induce reinstatement of the fear response, three unexpected USs were presented with 16–20 s in between the last re-extinction trial and the first US and 10–14 s in between each US. After reinstatement, both CSs were presented 3 more times in the test context and were not reinforced. On all days, the CSs were presented on the screen for 8 s. The startle probe was presented 7 s after CS onset and the electrical stimulus, if presented, occurred 7.5 s after CS onset. To ensure that the context would not become part of the CS but merely acted as a signal of the reinforcement probability, the context was presented 4–6 s before and after CS presentation. Intertrial intervals (black screen with fixation cross in the middle) were 16–20 s. Each experimental session started with 10 startle sound presentations to reduce habituation effects.

2.2.4. Procedure

During the first session, all participants were informed about the procedures in the study and signed informed consent. To be able to exploratorily investigate individual differences in extinction learning in the future, as well as to control for effects of personality traits between the groups, all participants completed a series of questionnaires on the computer at the start of the first session. Then, FPS, SCR and stimulus electrodes were attached. The intensity of the electrical stimulus was determined individually with the use of a work-up procedure, where participants received increasingly strong presentations of the US and were asked to say stop when the stimulus was clearly uncomfortable, and they did not want to continue any higher. Participants were then asked to rate the intensity of the stimulus on a scale from 0 ("I barely felt anything") to 10 ("This is the most uncomfortable stimulus I can imagine to receive through these electrodes"). If participants rated the stimulus lower than 7, they were asked to try the next intensity level, but were free to choose either stimulus intensity. The task started immediately afterwards. On the second and third day, participants came back to the lab, completed a state anxiety questionnaire, and all electrodes were attached again before the experiment continued.

2.2.4.1. Experimental instructions. On the first day, participants were instructed that they would see two persons on the screen, one by one, and that the persons were always presented in an environment. We then instructed them that one of the two persons would never be followed by the electrical stimulus, in none of the environments, whereas the other person would always be followed by the stimulus in one environment, and sometimes in the other. We provided these detailed instructions to ensure that our manipulation would be successful, because we can only test our hypothesis during the extinction phase if the acquisition manipulation is successful. On day two participants were instructed that the experiment would continue, and that they had to think back about what they learned the day before, and on day 3 participants were simply told that the experiment would continue.

2.2.5. Statistical analyses

All hypotheses were tested with both Bayesian and frequentist statistics in JASP (JASP Team, 2020). We first investigated the effectivity of our manipulation by performing a Stimulus (CS+, CS-) × Context (100%, 50%) × Trial (1–4) × Group repeated measures ANOVA on the expectancy data during acquisition and expected to find that CS + responses in the 100% context were larger than in the 50% context, indicating that participants learned the different contingencies. We further checked whether ratings were higher in the strong EV group than in the weak EV group at the start of extinction by performing a one-sided independent *t*-test on the CS + expectancies on the first extinction trial. Fear-potentiated startle and skin conductance data were analysed in similar fashion. As test of acquisition, we performed a Stimulus (CS+, CS-) × Context (100%, 50%) × Trial (1–4) × Group repeated measures ANOVA on all CS+/CS- trials. This is in contrast to our preregistration. We planned to only exploratively include context in the ANOVA, but the effects of interest are best tested in a single ANOVA including all factors. We expected to find a Stimulus × Trial effect, but no effects of Group. We

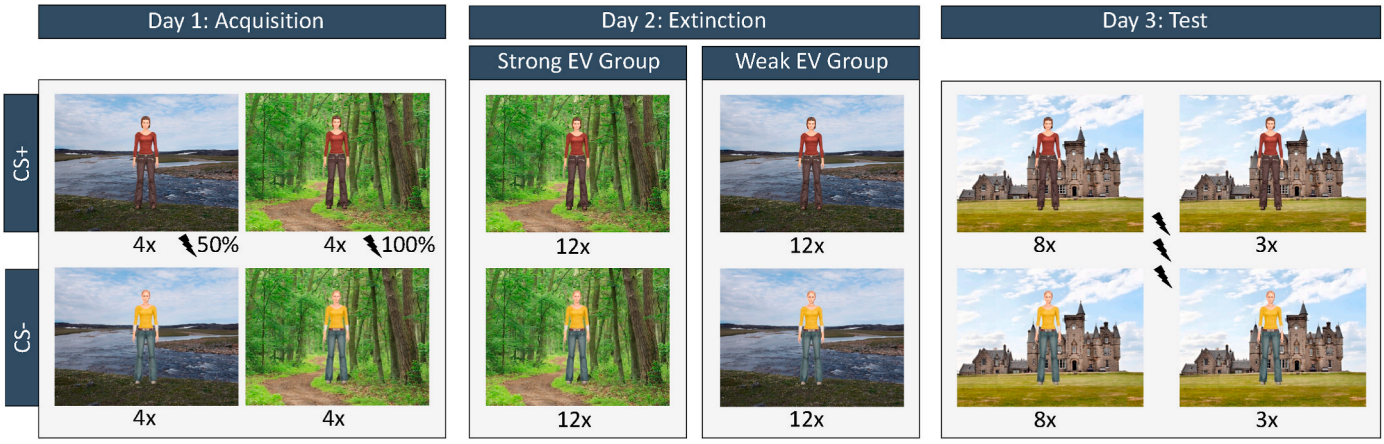

Fig. 1. Design of Experiment 1. Day 1 and day 3 are identical for all participants. The context in which extinction takes place on day 2 differs between the strong EV group and the weak EV group to create different outcome expectations.

were agnostic with regards to whether the responses would differ between the contexts. We then checked whether extinction occurred in both groups with a Stimulus (CS+, CS-) × Trial (1,2,11,12) × Group rm ANOVA on the first two and last two extinction trials. Lastly, we performed two Stimulus (CS+, CS-) × Group rm ANOVAs to compare differential responding on the first renewal and reinstatement trials as a test of our main hypothesis. We expected to observe a Stimulus × Group interaction for both ANOVAs, indicating less differential responding at the renewal and reinstatement tests in the strong EV group.

3. Results Experiment 1

3.1. Participant characteristics

Baseline participant characteristics are shown in Table 1. We found a significant difference between the groups on STAI-S scores on the first day, with higher state anxiety for the weak EV group.

3.2. Manipulation check expectancy ratings

We first checked whether participants in both groups learned that the US probability differed between the two contexts in a Stimulus (CS+, CS-) × Context (100%, 50%) × Trial (1–4) x Group rm ANOVA, and found a significant Stimulus × Context × Trial interaction (BF_{inc} = 8.46e13, F(2.0,118.2) = 34.6, $p < .001$, $\eta_p^2 = 0.37$), showing that in both groups participants learned that the reinforcement rate of the CS+ was higher in the 100% context than in the 50% context (see Fig. 2). A one-sided independent *t*-test comparing the CS + expectancy on the first extinction trial between the two groups further showed that US expectancy was higher in the strong EV group than in the weak EV group (BF_{+0} = 5.81e7, t(58) = 7.64, $p < .001$, d = 2.0). Higher threat expectancy in the strong EV group should result in stronger expectancy violations upon the non-occurrence of the US. The extinction data also showed a strong Stimulus (CS+, CS-) × Trial (1,2,11,12) × Group interaction (BF_{inc} = 3.15e3, F(2.0,114.0) = 10.1, $p < .001$, $\eta_p^2 = 0.15$), showing that extinction of the expectancy ratings differed between groups. However, in both groups we found a significant Stimulus × Trial interaction (Strong EV: BF_{inc} = 1.04e31, F(1.9,56.3) = 80.3, $p < .001$, $\eta_p^2 = 0.74$, Weak EV: BF_{inc} = 5.18e23, F(1.2,33.7) = 64.5, $p < .001$, $\eta_p^2 = 0.69$) showing that the differential expectancy ratings decreased over time. Importantly, we found a Stimulus × Group interaction (BF_{inc} = 7.64, F(1,58) = 6.98, $p = .011$, $\eta_p^2 = 0.11$) when comparing differential expectancies at the last two extinction trials, showing that the difference between CS+ and CS- expectancies was still larger in the weak EV group. Because these larger differential expectancies at the end of extinction may explain potential differences in the return of fear on day 3, we checked whether expectancy ratings differed between the groups on the first renewal trial. We found a main effect of Stimulus (BF_{inc} = 7.9e13, F(1,58) = 98.0, $p < .001$, $\eta_p^2 = 0.63$), but no evidence for a Stimulus × Group interaction (BF_{inc} = 0.80, F(1,58) = 2.66, $p = .108$, $\eta_p^2 = 0.04$), indicating that expectancy ratings at the start of day 3 were higher for the CS+, but that differential outcome expectancies were approximately similar between groups.

Table 1
Baseline characteristics of all participants from Experiment 1. Baseline characteristics were compared between groups with a Bayesian chi-square test (gender) and a Bayesian one-way ANOVA (all others). Bayes Factors showing evidence for the existence of a difference between groups are displayed.

	All	Strong EV	Weak EV	BF_{10}	*p*
N	60	30	30	–	–
Female/male	48/12	24/6	24/6	0.25	1.00
Age (years)	20.3 (2.1)	20.4 (2.0)	20.1 (2.2)	0.31	.530
US intensity (mA)	13.4 (8.9)	12.3 (6.9) (22.5)	14.5 (10.6)	0.38	.360
US rating	6.7 (1.1)	6.7 (1.2)	6.8 (1.0)	0.30	.597
ASI	21.4 (10.5)	21.4 (11.2)	20.4 (10.2)	0.26	.727
STAI-T	45.9 (9.6)	46.2 (10.9) (10.2)	45.5 (8.2)	0.27	.769
STAI-S D1	36.7 (7.6)	34.2 (7.1)	39.3 (7.3)	6.12	.007
STAI-S D2	33.8 (8.4)	33.4 (7.3)	34.1 (8.4)	0.28	.728
STAI-S D3	32.6 (8.1)	32.7 (8.4)	32.6 (7.7)	0.26	.992

3.3. Acquisition of FPS responses and SCRs

We performed a Stimulus (CS+, CS-) × Context (50%, 100%) × Trial (1–8) × Group (Strong EV, Weak EV) repeated measures ANOVA to test whether fear acquisition took place on day 1 for both FPS responses and SCRs. For FPS responses we found a main effect of Stimulus (BF_{inc} = 7.0e4, F(1,58) = 34.6, $p < .001$, $\eta_p^2 = 0.37$), showing that responding to the CS+ is larger than to the CS-, but no Stimulus × Trial interaction (BF_{inc} = 0.08, F(3,174) = 1.81, $p = .148$, $\eta_p^2 = 0.03$). We further found weak, but significant, evidence for an interaction between Stimulus × Group (BF_{inc} = 0.62, F(1,58) = 4.41, $p = .040$, $\eta_p^2 = 0.07$), which, based on the graphs, appears to be driven by stronger differential responding at the start of acquisition in the weak EV group. We further found no evidence for a Stimulus × Context effect (BF_{inc} = 0.11, F(1,58) = 0.17, $p = .681$, $\eta_p^2 < 0.01$), showing that FPS responding did not statistically differ between the CS50 and the CS100 context. The SCR data showed a main effect of Stimulus (BF_{inc} = 4.41e9, F(1,58) = 31.7, $p < .001$, $\eta_p^2 = 0.35$), and a Stimulus × Trial interaction (BF_{inc} = 1.17, F(3,174) = 3.93, $p = .010$, $\eta_p^2 = 0.06$). While we found no evidence for an interaction between Stimulus × Context (BF_{inc} = 0.10, F(1,58) = 0.01, $p = .920$, $\eta_p^2 < 0.01$), we did find a significant Stimulus × Context × Group interaction (BF_{inc} = 1.05, F(1,58) = 5.48, $p = .023$, $\eta_p^2 = 0.09$). When checking separately per group, we found no evidence for a Stimulus × Context interaction in the strong EV group (BF_{inc} = 0.32, F(1,29) = 1.95, $p = .173$, $\eta_p^2 = 0.06$), but a trend-wise significant effect in the weak EV group (BF_{inc} = 0.32, F(1,29) = 4.17, $p = .050$, $\eta_p^2 = 0.13$). This latter effect appears to be driven by larger differential responding in the 50% context. We thus found fear acquisition in both groups, with no overall effects of context, although SCRs to the CS+ in the 50% context were overall lower than in the 100% context in the strong EV group. The absence of a Stimulus x Trial effect in the FPS data is likely due to the inclusion of all trials. When analysing only the first versus the last trial, as is common in fear-conditioning studies, we did find a significant though small interaction (BF_{inc} = 1.13, F(1,58) = 4.41, $p = .040$, $\eta_p^2 = 0.07$).

3.4. Extinction of FPS responses and SCRs

We checked whether fear extinction took place in both groups with a Stimulus (CS+, CS-) × Trial (1,2,11,12) × Group repeated measures ANOVA. We found a main effect of Stimulus (BF_{inc} = 75.7, F(1,58) = 12.8, $p < .001$, $\eta_p^2 = 0.18$) and Trial (BF_{inc} = 1.68e23, F(3,174) = 47.1, $p < .001$, $\eta_p^2 = 0.45$), but not of Stimulus × Trial (BF_{inc} = 0.24, F(3,174) = 1.62, $p = .187$, $\eta_p^2 = 0.03$) for the FPS data, indicating that extinction did not fully occur. However, we found no significant difference between CS+ and CS- responding at the end of extinction across groups (last two trials; BF_{inc} = 0.45, F(1,58) = 1.77, $p = .188$, $\eta_p^2 = 0.03$). For SCRs we found a main effect of Stimulus (BF_{inc} = 73.26, F(1,58) = 13.1, $p < .001$, $\eta_p^2 = 0.18$) and Trial (BF_{inc} = 1.53e23, F(2.6,153.1) = 12.7, $p < .001$, $\eta_p^2 = 0.18$), but no Stimulus × Trial interaction (BF_{inc} = 0.30, F(3,174) = 1.40, $p = .246$, $\eta_p^2 = 0.02$). We found inconclusive, but significant, evidence for a Stimulus × Trial × Group interaction (BF_{inc} = 1.21, F(3,174) = 2.76, $p = .044$, $\eta_p^2 = 0.05$), which appears to be driven by the fact that in the strong EV group differential responding is already reduced on the second extinction trial. Again, we found no evidence for differential responding at the end of extinction (BF_{inc} = 0.57, F(1,58) = 2.14, $p = .149$, $\eta_p^2 = 0.04$).

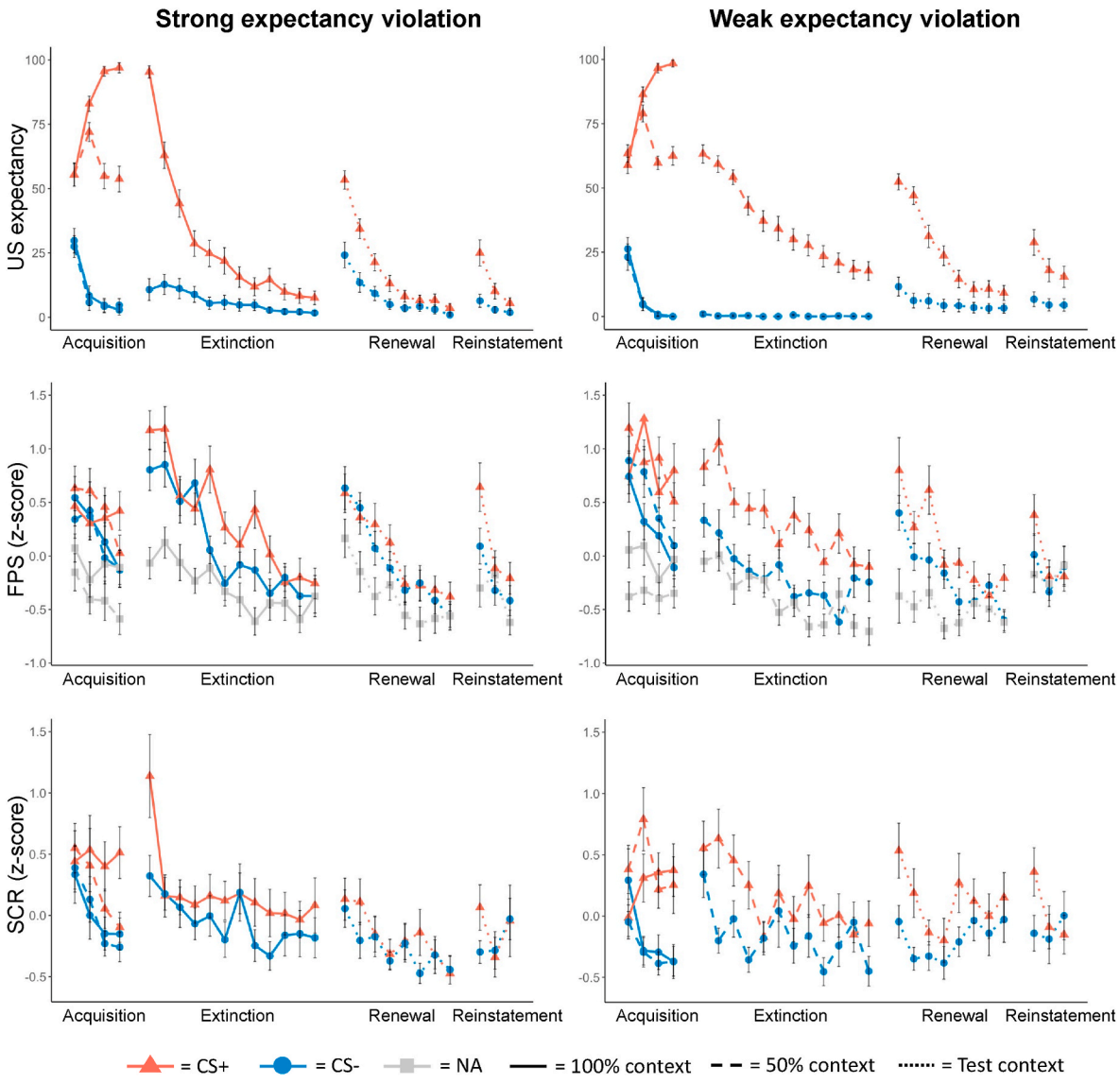

Fig. 2. Trial-by-trial US expectancy ratings, standardized fear-potentiated startle responses, and standardized skin conductance responses during Experiment 1 per group. Error bars represent 1 standard error.

3.5. FPS responses and SCRs at test

Because we found no differences between the groups during extinction, we did not include trial in the analyses and performed a Stimulus (CS+, CS-) × Group rm ANOVA to test for the return of fear on Day 3. Against our expectations, we found no evidence for a main effect of Stimulus (BF_{inc} = 0.27, F(1,58) = 0.68, p = .413, η_p^2 = 0.01) or a Stimulus × Group interaction (BF_{inc} = 0.46, F(1,58) = 1.11, p = .296, η_p^2 = 0.02) for FPS responses, showing that conditioned responding did not differentially return, and that this did not differ between the groups. After a re-extinction phase (8 trials) we induced reinstatement of the fear response by presenting three unannounced USs. We tested whether fear reinstatement differed between the groups in a Stimulus (CS+, CS-) × Group rm ANOVA and found a main effect of Stimulus (BF_{inc} = 9.2, F(1,58) = 8.75, p = .004, η_p^2 = 0.13), but again no Stimulus × Group interaction (BF_{inc} = 0.31, F(1,58) = 0.34, p = .563, η_p^2 < 0.01). These results indicate that the return of fear does not differ between the strong EV and weak EV groups. For SCR data, the renewal test showed a weak main effect of Stimulus (BF_{inc} = 1.5, F(1,58) = 4.36, p = .041, η_p^2 = 0.07). Again, we found no Stimulus × Group interaction (BF_{inc} = 0.8, F(1,58) = 2.55, p = .116, η_p^2 = 0.04), showing that differential SCRs at the start of day 3 did not significantly differ between groups. The reinstatement test also showed a main effect of Stimulus (BF_{inc} = 9.6, F(1,58) = 7.57, p = .008, η_p^2 = 0.12), but no Stimulus × Group interaction (BF_{inc} = 0.33, F(1,58) = 0.19, p = .661, η_p^2 < 0.01). In sum, our findings for FPS and SCR data are virtually comparable, and both showed that, in contrast with our hypotheses, differential conditioned responding did not differ between the groups during the test phase.

4. Experiment 2

Against our hypothesis, we found no evidence that stronger expectancy violations during extinction improve extinction retention. Importantly, the size of expectancy violations in clinical settings is dependent on the expectations that patients have about the occurrence or intensity of an outcome. While various methods exist to increase these expectations, other recommendations to strengthen inhibitory learning focus on enhancing awareness of violations that inherently occur during exposure (Craske et al., 2022; McGlade & Craske, 2021). For example, clinicians can ask patients immediately after a session whether their greatest worry came true and what they learned from the experience. While the working mechanisms of this manipulation are not explained

by associative learning theory, more awareness of expectancy violations may strengthen extinction retention through more cognitive mechanisms (Craske et al., 2022). In our second experiment, we therefore tested the hypothesis that fostering awareness of expectancy violations during extinction improves extinction retention. We aimed to increase awareness by explicitly asking participants after each extinction trial if the outcome they expected (i.e., the US) actually occurred in the experimental group, whereas in the control group extinction took place as normal.

4.1. Methods Experiment 2

All measures and materials were the same as in Experiment 1 (see *General measures and materials*).

4.1.1. Participants

Participants were sixty healthy university students aged between 18 and 35, with thirty participants in each group. Groups were balanced on gender. Exclusion criteria were as mentioned in Experiment 1. The study was approved by the ethics board of the University of Amsterdam and all participants signed informed consent before taking part.

4.1.2. Bayesian sequential updating

We preregistered a similar Bayesian sequential updating paradigm as in Experiment 1, where we planned to test a minimum of 40 participants and a maximum of 60 participants. We set the following stopping rules:

1. To test for fear acquisition, we performed a Bayesian *t*-test on the acquisition data, comparing the average of the last two CS + trials to the average of the last two CS- trials, separately for each group. We continued testing until we observed a $BF_{10} > 7$ for a difference between CS+ and CS- responding in both groups.
2. If the acquisition data implied stopping, we performed a Stimulus (CS+, CS-) x Group (Awareness, Control) Bayesian repeated measures ANOVA on the first trial of the test phase. We would continue testing until we observed a $BF_{inc} > 7$ or $< 1/3$ for the Stimulus × Group interaction.

Due to technical problems we were not able to access the processed data in time to perform the interim checks. We therefore decided to continue data collection until we reached the maximum sample size.

4.1.3. Design and experimental task

We employed a within-between subjects' design with two groups: An awareness group and a control group (see Fig. 3). The experiment consisted of three sessions on three consecutive days. The first (acquisition) and third (test) day were the same for both groups. During fear acquisition the CS+ and CS- were semi-randomly presented 8 times each. The CS+ was 50% reinforced and the CS- was never followed by the US. The reinforcement schedule of the CS+ was fixed and the first trial was always reinforced. The CS type of the first trial in the experiment (CS + or CS-) was randomized and balanced between groups. During extinction each CS was presented 12 times without reinforcement. In the awareness group, participants were asked after every trial "Did the outcome you expected occur?". This question was presented in the middle of the screen in white letters on a black background, and participants had to click on either "Yes" or "No". The question would remain on the screen for 3 s after answering, with a maximum of 10 s in total. The total time that the question was presented on the screen was deducted from the ITI time, to ensure the timing of trial presentation was consistent across groups. In the control group extinction proceeded as normal, and the ITI was presented immediately after each trial. All participants were tested for return of fear on the third day. The experiment started with three unexpected US presentations at the original US intensity to induce reinstatement of the fear response, with 10–14 s between each US. Then both CSs were presented 8 more times without reinforcement. Both the extinction and test phase started with a CS + presentation, as this allows for a better individual comparison of differential responding. On all days, the CSs were presented on the screen for 8 s. The startle probe is presented 7 s after CS onset and the electrical stimulus, if presented, occurs 7.5 s after CS onset. Intertrial intervals (black screen with fixation cross in the middle) are 16–20 s. In every session, a noise alone (NA) startle probe was presented randomly during the ITI to match the number of CS+ and CS- trials (i.e., 8 times on day 1). Each experimental session started with 10 startle sound presentations to reduce habituation effects.

4.1.4. Procedure

The procedure of Experiment 2 was identical to Experiment 1 with one exception: We learned from interviewing participants in the pilot phase that many participants appeared to believe that the equipment was not working properly during the extinction phase (i.e., that the absence of the electrical stimulus presentations was due to a technical failure), which may have interfered with safety learning. We therefore explicitly instructed them on day 1 that they should know that the equipment always works, and that not receiving an electrical stimulus can be part of the experiment. We further tested the electrical stimulus on both day 2 and 3 by giving participants a very low intensity US (2 mA) after attaching the electrode (thus not as part of the experiment itself) to indirectly show that the equipment was still working.

4.1.4.1. Experimental instructions. On the first day, participants were instructed that one of the persons would never be followed by the electrical stimulus and that this person only served for baseline measurements. They were also told that the other person could be followed by the electrical stimulus and that they had to learn what the probability is that they would receive the electrical stimulus after seeing that person. We chose to employ these detailed instructions based on piloting results in which participants showed high expectancy ratings and fear responses to the CS- during fear extinction. On the second day, participants were instructed that the experiment would continue and that they had to think about what the learned the previous day, and on day 3 we instructed participants only that the experiment would continue.

4.1.5. Statistical analyses

All hypotheses were tested with both Bayesian and frequentist statistics in JASP. We first investigated the effect of fear acquisition by performing a Stimulus (CS+, CS-) x Trial (1–8) x Group repeated measures ANOVA on the CS+/CS- trials for FPS, SCR, and threat expectancies. We then checked whether extinction occurred in both groups with a Stimulus (CS+, CS-) x Trial (1,2,11,12) x Group rm ANOVA on the first two and last two extinction trials. Again we expected to find a Stimulus x Trial effect showing a reduction of differential fear responding. To test our main hypothesis, we lastly performed a Stimulus (CS+, CS-) x Group rm ANOVA on the first reinstatement test trial. If fostering awareness of expectancy violations during extinction strengthens extinction learning, we would expect to observe a significant Stimulus x Group effect showing that the differential return of fear was higher in the control group.

5. Results Experiment 2

5.1. Participant characteristics

Baseline participant characteristics for Experiment 2 are shown in Table 2. We found a significant difference between the groups on age, with a slightly higher average age in the awareness group.

5.2. Manipulation check expectancy ratings

We checked whether participants in both groups learned to expect

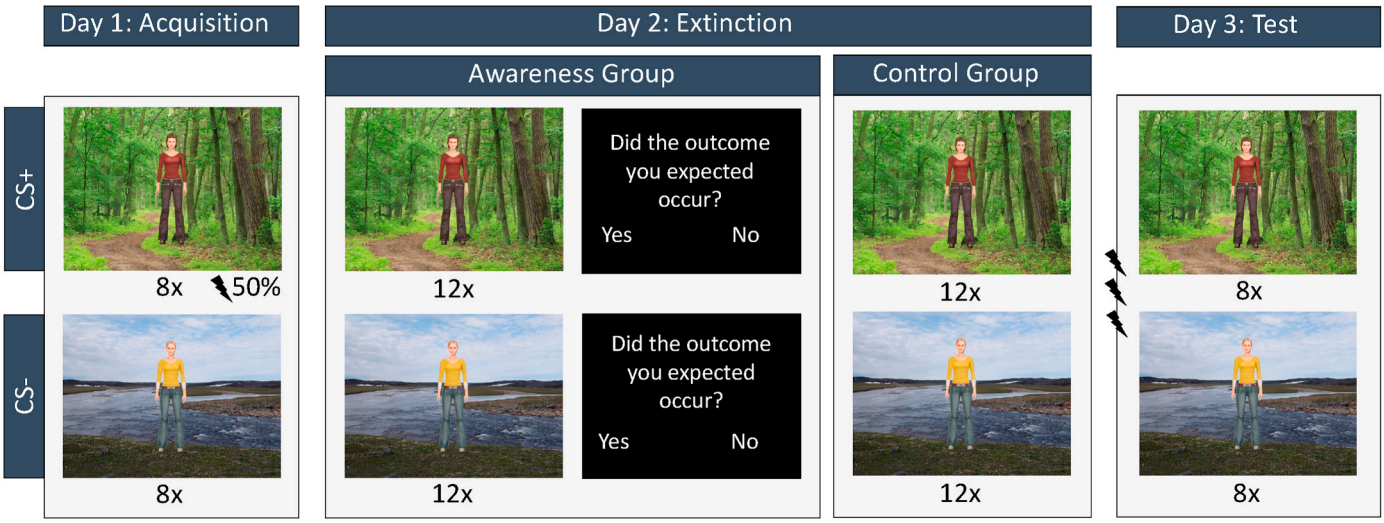

Fig. 3. Design Experiment 2. Again, day 1 and day 2 were identical for all participants, but during extinction only participants in one group were asked in the outcome they expected occurred.

Table 2
Baseline characteristics of all participants in Experiment 2. Baseline characteristics were compared between groups with a Bayesian chi-square test (gender) and a Bayesian one-way ANOVA (all other tests). Bayes Factors showing evidence for the existence of a difference between groups are displayed.

	All	Awareness	Control	BF_{10}	*p*
N	60	30	30	–	–
Female/male	50/10	25/5	25/5	0.24	1.00
Age (years)	21 (3.1)	21.9 (3.8)	20.1 (1.9)	2.00	.031
US intensity (mA)	17.5 (9.9)	17.6 (10.5) (22.5)	17.5 (9.5)	0.26	.959
US rating	7.4 (0.8)	7.5 (0.7)	7.3 (0.8)	0.42	.293
ASI	20.9 (10.4)	20.5 (9.1)	21.2 (11.7)	0.27	.807
STAI-T	44.8 (10.3)	43.9 (10.2) (10.2)	45.7 (10.6)	0.32	.504
STAI-S D1	35.2 (9.5)	35.0 (10.3)	35.4 (8.7)	0.27	.861
STAI-S D2	38.0 (5.6)	37.6 (6.1)	38.3 (5.1)	0.29	.630
STAI-S D3	35.4 (10.1)	34.9 (12.2)	35.8 (7.6)	0.28	.723

the US during acquisition in a Stimulus (CS+, CS-) x Trial (1–8) x Group (Awareness, Control) repeated measures ANOVA. We found a clear Stimulus × Trial interaction for threat expectancy ratings (BF_{inc} = 2.7e28, F(5.1,294.0) = 32.12, $p < .001$, $\eta_p^2 = 0.36$) showing that all participants learned to expect the US after the CS+, and no effects of Group. During the extinction phase, CS + expectancy ratings did decrease (Stimulus x Trial (1,2,11,12): BF_{inc} = 5.3e34, F(1.6, 90.2) = 98.3, $p < .001$, $\eta_p^2 = 0.63$), but this reduction was not complete as evidenced by average CS + ratings of 15 at the end of extinction (see Fig. 4). To further understand how fostering awareness during extinction affects the updating of US expectancies we exploratively compared the course of extinction between groups. In line with Pittig et al. (2022), we used a Rescorla Wagner learning rule to calculate the optimal individual learning rate for each participant for the CS + ratings during the extinction and test phases (see supplementary material). We then compared the average learning rates between groups and found a significantly higher learning rate in the awareness group (BF_{10} = 1.85, t(58) = 2.18, $p = .034$, d = 0.56). Interestingly, while we found no differences on the return of CS + threat expectancy on the first trial between the groups (BF_{10} = 0.30, t(58) = 0.57, $p = .571$, d = 0.15), the higher CS + learning rate in the awareness group was maintained in the test phase (BF_{10} = 1.82, t(58) = 2.17, $p = .034$, d = 0.56), while we no longer manipulated awareness in this phase.

5.3. Acquisition of FPS responses and SCRs

Fear responses were successfully acquired for all measures in both groups. We performed a Stimulus (CS+, CS-) x Trial (1–8) x Group (Awareness, Control) repeated measures ANOVA on the acquisition trials. Analysis of the FPS responses showed a significant Stimulus × Trial interaction (BF_{inc} = 0.73, F(5.8,335.2) = 2.54, $p = .022$, $\eta_p^2 = 0.04$), showing that a differential response developed over time in both groups (all BF < 0.34 and *ps* > .180 for effects of group). For SCRs too, we found a significant Stimulus × Trial interaction (BF_{inc} = 1.7, F(4.1,236.9) = 3.17, $p = .014$, $\eta_p^2 = 0.05$). However, while the Stimulus x Trial × Group interaction was not significant (BF_{inc} = 0.04, F(4.1, 236.9) = 1.06, $p = .377$, $\eta_p^2 = 0.02$), we did observe a significant Stimulus × Group interaction (BF_{inc} = 2.08, F(1,58) = 5.34, $p = .024$, $\eta_p^2 = 0.08$) indicating overall differential responding only in the awareness group. Thus, while differential skin conductance responding in the control group was weak, in general participants acquired fear responses to the CS + for all measures.

5.4. Extinction of FPS responses and SCRs

To check whether extinction of the fear responses occurred we performed a Stimulus (CS+, CS-) x Trial (1,2,11,12) x Group (Awareness, Control) rm ANOVA on the responses during extinction. Further, for FPS responses we found no evidence for a Stimulus × Trial interaction (BF_{inc} = 0.09, F(2.8,149.2) = 1.28, $p = .283$, $\eta_p^2 = 0.02$), showing that extinction did not take place. Indeed, both groups still showed significant differential responding on the last two extinction trials (main effect of Stimulus BF_{inc} = 1.8e7, F(1,58) = 31.4, $p < .001$, $\eta_p^2 = 0.35$). For skin conductance data, extinction was successful as evidenced by a significant Stimulus × Trial interaction (BF_{inc} = 2.07, F(3,174) = 3.68, $p = .013$, $\eta_p^2 = 0.06$), with no differences between the groups. This shows that while SCRs were not successfully acquired in the control group, fear responses were retained and subsequently extinguished.

5.5. FPS responses and SCRs at test

To investigate the effect of fostering awareness of expectancy violations during extinction on the retention of extinction, we compared (the return of) differential responding on the first test trial between groups in a Stimulus (CS+, CS-) x Group rm ANOVA. Because the experiment started with a CS + trial for all participants, a higher CS + response in itself is not very meaningful as this can be driven by order

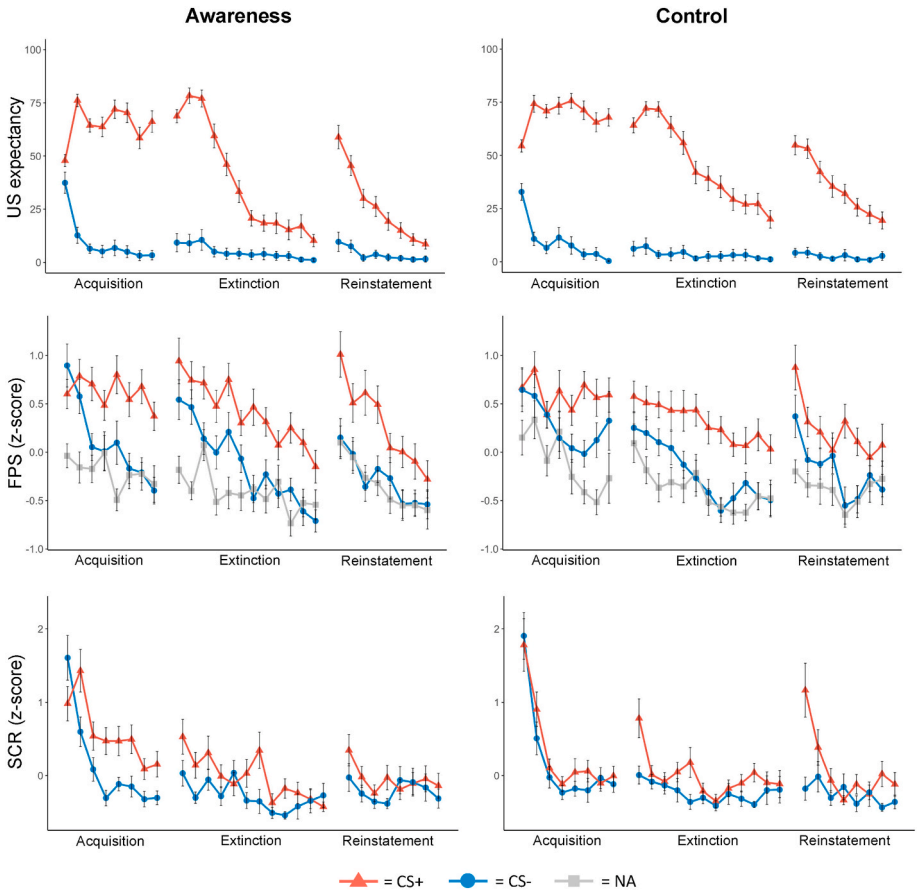

Fig. 4. Trial-by-trial US expectancy ratings, standardized fear-potentiated startle responses, and standardized skin conductance responses during Experiment 2 for both groups. Error bars represent 1 standard error.

effects, but differences in differential responding between the groups can be interpreted as differences in extinction retention. For FPS responses we found a strong effect of Stimulus ($BF_{inc} = 31.9$, $F(1,58) = 9.72$, $p = .003$, $\eta_p^2 = 0.14$), but no Stimulus × Group interaction ($BF_{inc} = 0.38$, $F(1,58) = 0.66$, $p = .420$, $\eta_p^2 = 0.01$), showing that for FPS responses, fostering awareness during extinction did not affect the return of fear. In contrast, for SCRs we did find a significant Stimulus × Group interaction[1] ($BF_{inc} = 2.49$, $F(1,58) = 5.06$, $p = .028$, $\eta_p^2 = 0.08$). A paired samples *t*-test within each group showed that differential responding only returned in the control group ($BF_{10} = 103.3$, $t(29) = 4.13$, $p < .001$, $d = 0.76$) and not in the awareness group ($BF_{10} = 0.42$, $t(29) = 1.31$, $p = .201$, $d = 0.24$). These latter findings support our hypothesis and suggest that fostering awareness during extinction may improve extinction retention and reduce the return SCRs one day later.

6. Discussion

In two fear-conditioning studies we investigated whether stronger expectancy violations or fostering awareness of expectancy violations can strengthen extinction learning and improve extinction outcomes. In the first experiment, we employed a fear-conditioning paradigm with two different contexts acting as occasion setters that modulated the probability of the US during extinction. This allowed us to isolate the effect of expectancy violations independent of the strength of prior learning. Our data show that the manipulation was effective as participants in the strong violation group had significantly higher US expectancies than in the weak violation group, which should result in stronger violations upon US omission. We indeed observed that stronger expectancy violations resulted in faster updating of threat expectancies. However, we found no differences between the groups in differential return of fear during the test phase, indicating that stronger expectancy violations did not improve extinction retention in our study. While we were predominantly interested in the effect of expectancy violations on physiological responses, we also examined differences in threat expectancies at test, but found no differences between the groups either. In the second experiment we tried to increase the awareness of expectancy violations by explicitly asking participants after each trial to what extent the outcome they expected actually occurred. This manipulation, too, led to faster updating of outcome expectancies during extinction, and although we found no differences between the groups on the return of FPS responses, the awareness group showed significantly lower differential SCRs at test. This suggests that our manipulation improved extinction retention to some extent.

Surprisingly, stronger expectancy violations did not result in better retention of extinction learning. These results contrast fear-conditioning studies that demonstrated a beneficial effect of increasing expectancy violations on extinction outcomes (Coelho et al., 2015; Culver et al., 2015; Gromer et al., 2022). In the current study, we aimed to isolate the effect of expectancy violations by ensuring that there were no initial differences in learning between the groups and that the same CS was used during extinction. While one may argue that in our paradigm participants only learned new information about the CS-US relationship

[1] Because differences in SCRs between the groups during fear acquisition influence the interpretability of the Stimulus × Group interaction on day 3, we performed the same analyses on a subset of participants with good acquisition (larger absolute average CS + response on day 1). The Stimulus × Group interaction on the first test trial is maintained in this subset indicating that our results are not dependent on differences in fear acquisition.

in one of the two contexts, the use of a novel context in the test phase should avoid spill over effects of the context on test responses. Any differences in return of fear between the groups can thus be interpreted as an effect of what participants could learn during extinction, and our data suggests that larger US expectancy violations do not strengthen extinction retention. We would like to stress, however, that this observation by no means implies that expectancy violations are trivial to learning. Participants in both groups experienced expectancy violations, but differences in the magnitude of these violations during extinction learning did not affect the return of fear. While these findings seem to challenge the idea that increasing outcome expectancies strengthens extinction retention, it is important to note that isolating the effect of expectancy violations is complex, and potential alternative explanations for our findings are discussed in the next paragraph. Interestingly, one important difference with the successful compound extinction procedure (Coelho et al., 2015; Culver et al., 2015) is the timing of the expectancy violation. In compound extinction, two CSs are fist extinguished separately, and only later combined to increase the expectation of the US again. This procedure may thus depend on sustained expectancy violations throughout extinction rather than increasing the US expectancy at the start of learning.

Our results illustrate the importance of a clearer and more unified understanding of what it means to optimize expectancy violations. We aimed for the (in our view) most direct manipulation by creating strong differences between the groups in the predicted probability of the US. The high violation group was expecting the CS to be 100% reinforced, resulting in large violations and faster extinction learning when the US is omitted during extinction. However, participants in this group also updated their US expectancy faster, actually causing lower expectancy violations toward the end of extinction. In contrast, in the low violation group, participants were more likely to still expect the US to occur at the end of extinction. The total expectancy violation occurring in the entire extinction phase may therefore in fact be larger in the low violation group. According to the inhibitory retrieval model, expectancy violations should be optimised throughout exposure (Craske et al., 2022). While expectancy violations were larger at the start of extinction in our design, this may not have resulted in the most efficient method to strengthen extinction learning, as violations were not continuously increased. Moreover, recent studies found that when expectancy violations were too large, the information that these violations provided (e.g., safety in our study) was considered less valid, which consequently led to less updating of future expectations (Kube et al., 2022; Spicer et al., 2020). Another potential limitation of the current design is the 50% reinforcement schedule during acquisition, which may have affected more than just the US expectancy during extinction. For example, participants may be more likely to discard information (i.e., the non-occurrence of the US) provided during extinction, as the US omission is not surprising given prior non-reinforcement of the CS (e.g., Harris et al., 2019). The partial reinforcement may also have resulted in more context-specific learning (Bouton & Sunsay, 2001; Pearce et al., 1997). Given the use of a novel test context, this difference could have affected the return of fear between the groups. These considerations show that manipulating the expectation of the US during extinction could affect more than just the expectancy violations, and that the most optimal method to increase expectancy violations remains unclear.

Another critical question with respect to the current design is whether the US expectancy violations in our study are comparable to those experienced in clinical situations. The proposed effect of expectancy violations on treatment assumes that these violations are a proxy for prediction errors (e.g., Craske et al., 2022). Associative learning models define prediction errors as the difference between the predicted outcome based on the affective strength of the memory, and the actual outcome (Rescorla & Wagner, 1972). In the current experiment, we were specifically interested in the effect of US expectancy violations during extinction learning on subsequent retention of extinction and therefore deliberately aimed to keep the affective strength of the original memories the same between the groups. Yet arguably, this manipulation in fact resulted in similar levels of prediction error in both groups – if it is truly prediction errors that drive the effect of expectancy violations. In clinical settings, outcome expectancies are more likely to capture at least part of the affective value of the memory, whereas US expectations in fear-conditioning reflect perhaps only propositional learning. Expectancy violations may therefore be a better proxy for prediction error in clinical settings than US expectancy violations are in conditioning experiments. This difference also illustrates how complex it is to mirror and investigate the proposed processes of change during exposure treatment in a fear-conditioning experiment. It should be noted, however, that there exists no direct behavioural read-out of prediction error. While using a proxy such as expectancy violation is a good option when aiming to identify the conditions that can strengthen extinction learning and retention, one can always argue that prediction errors were not actually manipulated, resulting in an unfalsifiable theory.

In contrast to manipulating the magnitude of expectancy violations, fostering the awareness of violations that inherently occur during extinction is probably a more feasible intervention in clinical practice. We found that differential SCRs returned during the test phase only in the control group, and not in the awareness group. To our knowledge, this is the first experimental study in which this approach has been tested, but our SCR findings support earlier observations from an exposure study (McGlade & Craske, 2021). Critically, however, FPS responses were not affected by the manipulation, indicating that the valence of, or the defensive reactions to, the CS did not differ between the groups (Ojala & Bach, 2020). That said, the interpretation of our FPS data is complicated by the complete absence of extinction of FPS responses. No extinction learning occurred as measured by FPS responses in either group which prohibits comparisons of the retention of what was (not) learned. Interestingly, we found no differences on expectancy ratings to the first test trial either, indicating that the SCRs during test did not directly reflect the expected outcome probability of the US. Taken together, fostering the awareness of expectancy violations may strengthen extinction retention, yet given the inconsistencies between measures, our findings need replication before drawing firm conclusions.

The working mechanisms of fostering awareness as a potential manipulation to strengthen extinction learning and retention are unclear. It was previously reported that patients with higher learning rates during exposure (i.e., more adaptive updating of threat expectancies) show better exposure outcomes (Pittig et al., 2022). We exploratorily investigated the effect of our manipulation on the CS + learning rate and found higher learning rates in the awareness group. Interestingly, this effect carried over to the test phase (in which awareness was no longer manipulated) suggesting that the manipulation may persistently incentivize faster updating of threat expectancies. These induced differences in extinction learning could potentially explain the lower return of SCRs, yet the relationship between faster updating and better extinction retention requires further investigation, as we did not observe this pattern in the first experiment where faster updating also occurred in the high violation group. Alternatively, our awareness manipulation may have increased attention paid to the CS during the task, which is suggested to improve extinction retention and generalisation (Barry et al., 2017; Howley & Waters, 2017; Klein et al., 2021; O'Malley & Waters, 2018). Lastly, given that extinction was entirely uninstructed, it may also simply be that fostering awareness decreased experimental uncertainty during extinction and that participants were more likely to trust that the US was no longer occurring.

Our psychophysiological data also contained some unexpected observations. For example, extinction of the FPS responses to the CS+ was compromised in both experiments. Although these effects can be partially explained by the observation that expectancy ratings did not fully decline either, there still remains a large discrepancy between US expectancy ratings and FPS responses. Interestingly, this is consistent with clinical observations that behavioural reactions and feelings to

stimuli may strongly differ from the expectations that people hold about these stimuli (Elsey & Kindt, 2021). Specifically in our data, it appears that participants never believed the CS + to be entirely safe, even when able to learn that the probability of the US occurring declined. Especially in the second experiment, the lack of extinction hinders interpretation of the FPS results at test. Increasing the number of extinction trials may have been necessary to increase the retrievability of the extinction memory and to observe potential differences between the groups. Another factor that may affect the interpretation of the physiological responses is the use of the startle probe (a 50ms 104 dB white noise burst) to induce FPS responses. The startle probe can also function as a US in conditioning experiments (Sperl et al., 2016), and especially arousal-driven responses such as SCRs may be affected by the anticipation of the startle probe (de Haan et al., 2018). That said, the startle probe was consistently presented on both CS+ and CS- trials and with some exceptions we found that both SCRs and FPS responses were stronger to the CS+, which can only be due to the expectation of the electrical stimulus. In sum, we believe that the use of FPS responses to index the affective value of CSs outweighs the potential drawbacks of using startle probes in conditioning.

Finally, it is important to note that our results do not invalidate the inhibitory retrieval model (Craske et al., 2022). The hypothesis that strengthening the formation and retrievability of inhibitory memories can improve exposure treatments has clear face validity and has neither been falsified nor convincingly proven. While fear-conditioning studies could help to further elucidate the working mechanisms of the proposed strategies to strengthen inhibitory memories, variations in US expectancy may not sufficiently mirror the changes in outcome expectancies that occur in clinical settings, as discussed above. The results from Experiment 2 instead suggest that it could be beneficial to employ manipulations that foster awareness of expectancy violations during exposure. While speculative, such manipulations may improve exposure outcomes by increasing patients' tendencies to update their beliefs when experiencing expectancy violations (e.g., Pittig et al., 2022). Many treatment suggestions of the inhibitory retrieval model already emphasize manipulations focusing on e.g., awareness or consolidation of expectancy violations (Craske et al., 2022; McGlade & Craske, 2021). Developing a better understanding of the working mechanisms that contribute to stronger or more readily retrievable memories is key to further optimize these processes.

In conclusion, our data show that creating stronger US expectancies during extinction does not result in better retention of the extinction memory in a fear-conditioning paradigm. In the second experiment, we found that fostering awareness can reduce the return of SCRs, but not FPS responses. The inconsistency between physiological measures makes it difficult to draw conclusions about the clinical validity of these latter results, but we believe that there is potential in boosting awareness of expectancy violations as means to improve exposure outcomes. Importantly, while our results shed some light on the potential factors that may affect extinction retention, there exist a multitude of methods to manipulate expectancy violations. Perhaps despite – or due to – the use of a well-controlled paradigm we were unable to successfully create meaningful differences in the type of expectancy violations that matter for learning (i.e., that are a better proxy of prediction error). That said, we hope that these studies do not only contribute to a better understanding of the role of expectancy violations in extinction learning and retention, but also illustrate the difficulties that are inherent to the investigation of these questions in the laboratory and clinical practice. To further improve exposure treatment and to better align theoretical predictions with experimental research and clinical applications, a more concise and unified understanding of the exact factors that contribute to long-lasting reductions in fear needs to be developed.

CRediT authorship contribution statement

Lotte E. Stemerding: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing. **Vanessa A. van Ast:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Merel Kindt:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work rparaeported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by an ERC Advanced Grant (74326) rewarded to Merel Kindt. Vanessa van Ast is supported by a VENI NWO grant (451-16-021). We thank Coco de Waard for her help with data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2023.104319.

References

Barry, T. J., Vervliet, B., & Hermans, D. (2017). Feature specific attention and return of fear after extinction. *Journal of Experimental Psychopathology, 8*(1), 76–87. https://doi.org/10.5127/jep.051115

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, W. T., Dawson, M. E., & Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology, 49*(8), 1017–1034. https://doi.org/10.1111/j.1469-8986.2012.01384.x

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin, 114*(1), 80–99.

Bouton, M. E., & Sunsay, C. (2001). Contextual control of appetitive conditioning: Influence of a contextual stimulus generated by a partial reinforcement procedure. *Quarterly Journal of Experimental Psychology Section B Comparative and Physiological Psychology, 54*(2), 109–125. https://doi.org/10.1080/02724990042000083

Coelho, C. A. O., Dunsmoor, J. E., & Phelps, E. A. (2015). Compound stimulus extinction reduces spontaneous recovery in humans. *Learning & Memory, 22*(12), 589–593. https://doi.org/10.1101/lm.039479.115

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy, 58*, 10–23. https://doi.org/10.1016/j.brat.2014.04.006

Craske, M. G., Treanor, M., Zbozinek, T. D., & Vervliet, B. (2022). Optimizing exposure therapy with an inhibitory retrieval approach and the OptEx Nexus. *Behaviour Research and Therapy, 152*(March), Article 104069. https://doi.org/10.1016/j.brat.2022.104069

Culver, N. C., Vervliet, B., & Craske, M. G. (2015). Compound extinction: Using the Rescorla–Wagner model to maximize exposure therapy effects for anxiety disorders. *Clinical Psychological Science, 3*(3), 335–348. https://doi.org/10.1177/2167702614542103

Deacon, B. J., & Abramowitz, J. S. (2004). Cognitive and behavioral treatments for anxiety disorders: A review of meta-analytic findings. *Journal of Clinical Psychology, 60*(4), 429–441. https://doi.org/10.1002/jclp.10255

Deacon, B. J., Kemp, J. J., Dixon, L. J., Sy, J. T., Farrell, N. R., & Zhang, A. R. (2013). Maximizing the efficacy of interoceptive exposure by optimizing inhibitory learning: A randomized controlled trial. *Behaviour Research and Therapy, 51*(9), 588–596. https://doi.org/10.1016/j.brat.2013.06.006

Elsey, J. W. B., & Kindt, M. (2021). Expectations of objective threats and aversive feelings in specific fears. *Scientific Reports, 11*, Article 20778. https://doi.org/10.1038/s41598-021-00317-3, 1–15.

Foa, E. B., & Kozak, M. J. (1986). Emotional processing of fear: Exposure to corrective information. *Psychological Bulletin, 99*(1), 20–35.

Gromer, D., Hildebrandt, L., & Stegmann, Y. (2022). The role of expectancy violation in extinction learning: A two-day online fear conditioning study. *PrePrint* https://doi.org/10.31234/OSF.IO/FZSW5.

de Haan, M. I. C., van Well, S., Visser, R. M., Scholte, H. S., van Wingen, G. A., & Kindt, M. (2018). The influence of acoustic startle probes on fear learning in humans. *Scientific Reports, 8*(1), 1–11. https://doi.org/10.1038/s41598-018-32646-1

Harris, J. A., Kwok, D. W. S., & Gottlieb, D. A. (2019). The partial reinforcement extinction effect depends on learning about nonreinforced trials rather than

reinforcement rate. *Journal of Experimental Psychology: Animal Learning and Cognition*. https://doi.org/10.1037/xan0000220

Howley, J., & Waters, A. M. (2017). Overt verbalization of strategies to attend to and retain learning about the threat conditioned stimulus reduces US expectancy generalization during extinction. *Learning and Motivation, 59*(May), 19–26. https://doi.org/10.1016/j.lmot.2017.05.013

JASP Team. (2020), 0.14.1 *JASP*.

Kircanski, K., Mortazavi, A., Castriotta, N., Baker, A. S., Mystkowski, J. L., Yi, R., & Craske, M. G. (2012). Challenges to the traditional exposure paradigm: Variability in exposure therapy for contamination fears. *Journal of Behavior Therapy and Experimental Psychiatry, 43*(2), 745–751. https://doi.org/10.1016/j.jbtep.2011.10.010

de Kleine, R. A., Hendriks, L., Becker, E. S., Broekman, T. G., & van Minnen, A. (2017). Harm expectancy violation during exposure therapy for posttraumatic stress disorder. *Journal of Anxiety Disorders, 49*(March), 48–52. https://doi.org/10.1016/j.janxdis.2017.03.008

Klein, Z., Ginat-Frolich, R., Barry, T. J., & Shechner, T. (2021). Effects of increased attention allocation to threat and safety stimuli on fear extinction and its recall. *Journal of Behavior Therapy and Experimental Psychiatry, 72*(January), Article 101640. https://doi.org/10.1016/j.jbtep.2021.101640

Kube, T., Kirchner, L., Lemmer, G., & Glombiewski, J. A. (2022). How the discrepancy between prior expectations and new information influences expectation updating in depression—the greater, the better? *Clinical Psychological Science, 10*(3), 430–449. https://doi.org/10.1177/21677026211024644

Lang, A. J., & Craske, M. G. (2000). Manipulations of exposure-based therapy to reduce return of fear: A replication. *Behaviour Research and Therapy, 38*(1), 1–12. https://doi.org/10.1016/S0005-7967(99)00031-5

Loerinc, A. G., Meuret, A. E., Twohig, M. P., Rosenfield, D., Bluett, E. J., & Craske, M. G. (2015). Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clinical Psychology Review, 42*, 72–82. https://doi.org/10.1016/J.CPR.2015.08.004

McGlade, A. L., & Craske, M. G. (2021). Optimizing exposure: Between-session mental rehearsal as an augmentation strategy. *Behaviour Research and Therapy, 139* (November 2020), Article 103827. https://doi.org/10.1016/j.brat.2021.103827

Meulders, A., van Daele, T., Volders, S., & Vlaeyen, J. W. S. (2016). The use of safety-seeking behavior in exposure-based treatments for fear and anxiety: Benefit or burden? A meta-analytic review. *Clinical Psychology Review, 45*, 144–156. https://doi.org/10.1016/j.cpr.2016.02.002

Norton, P. J., & Price, E. C. (2007). A meta-analytic review of adult cognitive-behavioral treatment outcome across the anxiety disorders. *The Journal of Nervous and Mental Disease, 195*(6), 521–531. https://doi.org/10.1097/01.nmd.0000253843.70149.9a

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

O'Malley, K. R., & Waters, A. M. (2018). Attention avoidance of the threat conditioned stimulus during extinction increases physiological arousal generalisation and retention. *Behaviour Research and Therapy, 104*, 51–61. https://doi.org/10.1016/J.BRAT.2018.03.001

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

Pearce, J. M., Redhead, E. S., & Aydin, A. (1997). Partial reinforcement in appetitive Pavlovian conditioning with rats. *The Quarterly Jorunal of Experimental Psychology, 50B*(4), 273–294.

Pittig, A., Heinig, I., Goerigk, S., Richter, J., Hollandt, M., Lueken, U., Pauli, P., Deckert, J., Kircher, T., Straube, B., Neudeck, P., Koelkebeck, K., Dannlowski, U., Arolt, V., Fydrich, T., Fehm, L., Ströhle, A., Totzeck, C., Margraf, J., … Wittchen, H.-U. (2022). Change of threat expectancy as mechanism of exposure-based psychotherapy for anxiety disorders: Evidence from 8,484 exposure exercises of 605 patients. *Clinical Psychological Science, 0*(0), 1–19. https://doi.org/10.1177/21677026221101379

Rescorla, R. A. (2006). Deepened extinction from compound stimulus presentation. *Journal of Experimental Psychology: Animal Behavior Processes, 32*(2), 135–144. https://doi.org/10.1037/0097-7403.32.2.135

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Sperl, M. F. J., Panitz, C., Hermann, C., & Mueller, E. M. (2016). A pragmatic comparison of noise burst and electric shock unconditioned stimuli for fear conditioning research with many trials. *Psychophysiology, 53*, 1352–1365. https://doi.org/10.1111/psyp.12677

Spicer, S. G., Mitchell, C. J., Wills, A. J., & Jones, P. M. (2020). Theory protection in associative learning: Humans maintain certain beliefs in a manner that violates prediction error. *Journal of Experimental Psychology: Animal Learning and Cognition, 46*(2), 151–161. https://doi.org/10.1037/xan0000225

Springer, K. S., Levy, H. C., & Tolin, D. F. (2018). Remission in CBT for adult anxiety disorders: A meta-analysis. *Clinical Psychology Review, 61*, 1–8. https://doi.org/10.1016/J.CPR.2018.03.002

Wolpe, J. (1968). Psychotherapy by Reciprocal inhibition. *Conditional Reflex: A Pavlovian Journal of Research & Therapy, 3*(4), 234–240.

Zbozinek, T. D., Wise, T., Perez, O. D., Qi, S., Fanselow, M. S., & Mobbs, D. (2021). Pavlovian occasion setting in human fear and appetitive conditioning: Effects of trait anxiety and trait depression. *Behaviour Research and Therapy, 147*. https://doi.org/10.1016/J.BRAT.2021.103986