Behaviour Research and Therapy 160 (2023) 104233

Contents lists available at ScienceDirect

Behaviour Research and Therapy

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

Reduction of costly safety behaviors after extinction with a generalization stimulus is determined by individual differences in generalization rules

Alex.H.K. Wong [a,b,*], Jessica C. Lee [c], Paula Engelke [b], Andre Pittig [d]

[a] Department of Psychology, Educational Sciences, and Child Studies, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
[b] Department of Psychology (Biological Psychology, Clinical Psychology, and Psychotherapy), University of Würzburg, Marcusstraße 9-11, 97070 Würzburg, Germany
[c] School of Psychology, University of New South Wales Sydney, Kensington NSW 2052 Sydney, Australia
[d] Translational Psychotherapy, Department of Psychology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraße 49b, 91052, Erlangen, Germany

A B S T R A C T

Exposure-based treatment involves repeated presentation of feared stimuli or situations in the absence of perceived threat (i.e., extinction learning). However, the stimulus or situation of fear acquisition (CS+) is highly unlikely to be replicated and presented during treatment. Thereby, stimuli that resemble the CS+ (generalization stimuli; GSs) are typically presented. Preliminary evidence suggests that depending on how one generalizes fear (i.e., different generalization rules), presenting the same GS in extinction leads to differential effectiveness of extinction learning. The current study aimed to extend this finding to safety behaviors. After differential fear and avoidance conditioning, participants exhibited discrete generalization gradients that were consistent with their reported generalization rules (Similarity vs Linear). The Linear group showed stronger safety behaviors to a selected GS compared to the Similarity group, presumably due to higher threat expectancy. After extinction learning to this GS, the Linear group exhibited stronger reduction in safety behaviors generalization compared to the Similarity group. The results show that identifying distinct generalization rules allows one to predict expectancy violation to the extinction stimulus, in addition to corroborating the idea that strongly violating threat expectancy leads to better extinction learning and its generalization. With regard to clinical implications, identifying one's generalization rule (e.g., threat beliefs) help designing exposure sessions that evoke strong expectancy violation, enhancing the reduction in the generalization of maladaptive safety behaviors.

1. Introduction

Exposure-based therapy has been widely considered as one of the most effective treatments for anxiety-related disorders (Carpenter et al., 2018; Hofmann & Smits, 2008). A part of this treatment involves repeated confrontation to feared stimuli or situations in the absence of an anticipated threat, thus disconfirming maladaptive threat beliefs to these fear-evoking stimuli or situations. This learning about the absence of threat can be modelled in the laboratory via fear extinction training, which has been widely regarded as a valid laboratory analogue to exposure-based therapy (Carpenter, Pinaire, & Hofmann, 2019; Scheveneels, Boddez, Vervliet, & Hermans, 2016). Fear extinction involves repeated presentation of a conditioned stimulus (CS+) alone that previously predicted an aversive unconditioned stimulus (US). As a result of repeated exposure to an unreinforced CS+, conditioned fear decreases across extinction trials (e.g., Alvarez, Johnson, & Grillon, 2007; Hermans et al., 2005; Milad, Orr, Pitman, & Rauch, 2005; Orr et al., 2000).

There are, however, factors that affect the effectiveness of exposure-based therapy. Expectancy violation is thought to be one central mechanism that enables extinction learning. It refers to the mismatch between an expected outcome and the actual outcome (i.e., expectation of an aversive event and its actual absence). According to the Inhibitory model, the stronger the expectancy violation (i.e., the greater the mismatch between the expected and actual outcomes), the stronger extinction learning is and hence translates to better learning outcome (Craske, Treanor, Conway, Zbozinek, & Vervliet, 2014; Craske, Hermans, & Vervliet, 2018). One reason for suboptimal expectancy violation in exposure-based treatments may be that the exact stimulus or circumstances at the time of fear acquisition (i.e., the original CS+) are highly unlikely to be reproduced. Thereby, stimuli that are different from the CS+, but nonetheless resemble it, are presented in treatment. In the laboratory, this is equivalent to presenting generalization stimuli (GSs) that perceptually or conceptually resemble the CS+ in extinction. However, empirical studies have shown that extinction learning to a GS does not effectively generalize to the original CS+ (e.g., Barry, Griffith, Vervliet, & Hermans, 2016; Vervliet & Geens, 2014; Vervliet, Vansteenwegen, Baeyens, Hermans, & Eelen, 2005; Vervliet, Vansteenwegen, & Eelen, 2004; Vervoort, Vervliet, Bennett, & Baeyens, 2014; Wong & Lovibond, 2020) or to other GSs that resemble the CS+ (e.g., Vervliet et al., 2004; Vervoort et al., 2014; Wong & Lovibond, 2020).

* Corresponding author. Erasmus School of Social and Behavioural Sciences, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands.
E-mail address: h.k.wong@essb.eur.nl (Alex.H.K. Wong).

https://doi.org/10.1016/j.brat.2022.104233
Received 29 July 2022; Received in revised form 7 October 2022; Accepted 17 November 2022
Available online 18 November 2022

This ineffective generalization of extinction learning to a GS is presumably due to the GS not evoking as much expectancy violation as a CS+ (i.e., due to generalization decrement). That is, expectancy violation is not maximized when presenting a GS in extinction, thus limited extinction learning generalizes to other stimuli. Thereby, poor responses to exposure-based treatment are presumably partly due to limited expectancy violation to the GS in extinction. It has also been proposed that weak generalization of GS extinction is a potential new pathway for the 'return of fear' (Wong & Lovibond, 2020); limited extinction learning to a GS failed to generalize to other novel stimuli, thus leading to an apparent relapse after successful treatment.

Recent studies have examined whether presenting a GS in extinction that evokes at least as much threat expectancy as the CS+ promotes generalization of GS extinction (Struyf, Hermans, & Vervliet, 2018; Wong, Glück, Boschet, & Engelke, 2020). This procedure induces strong expectancy violation, leading to robust extinction learning, and hence potentially translates to better treatment outcome. One way to search for GSs that evoke strong expectancy violation is to identify individual generalization rules. Recent laboratory studies (e.g., Ahmed & Lovibond, 2019; Lee, Hayes, & Lovibond, 2018; Wong & Lovibond, 2017) have assessed the different generalization gradients formed by the adaptation of distinct rules (rule-based generalization). Participants were first trained in a differential conditioning procedure with a CS+ at the centre of a stimulus dimension (e.g., aqua color circle) and a safety stimulus (CS-) off-centre (e.g., green circle) along the dimension (e.g., green-blue continuum). In a following generalization test, all stimuli along the dimension were presented individually, and participants exhibited distinct generalization gradients despite receiving highly similar training conditions (e.g., Ahmed & Lovibond, 2019; Lee et al., 2018; Wong & Lovibond, 2017). These distinct generalization gradients were due to participants inferring different self-reported rules, which guided fear generalization. For instance, participants who responded based on a linear rule (e.g., the bluer the stimulus, the more likely the US) exhibited a linear gradient, with the strongest conditioned fear to stimuli located at the opposite end of CS- along the stimulus dimension (see Fig. 1). In contrast, participants who responded based on a similarity rule (e.g., the more aqua color a stimulus is, the more likely the US) showed a peaked gradient with responding peaked at the CS+ in the middle of the stimulus dimension. In other words, these self-reported rules predict how fear is generalized, thereby forming distinct generalization gradients. The stimulus at the end of the stimulus dimension in the opposite direction of the CS- (e.g., S9 in Fig. 1) evokes strong generalized fear in the linear rule group, whereas it evokes limited generalized fear in the similarity rule group. We have recently found that presenting this stimulus in extinction resulted in different degrees of generalization of extinction in conditioned fear (Wong et al., 2020): the linear rule group showed greater generalization of extinction compared to the similarity rule group (i.e., flattening the post-extinction gradient to a greater extent), due to greater expectancy violation during extinction.

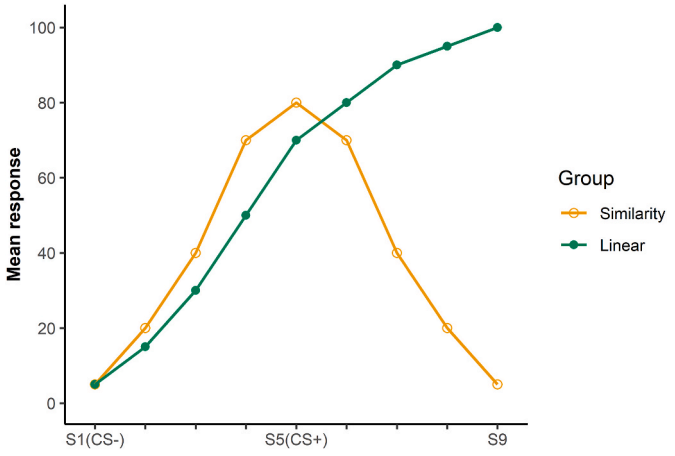

Fig. 1. Example generalization gradients. After differential conditioning training, two different generalization rules are typically inferred: Similarity and Linear. These generalization rules presumably drive generalization, leading to distinct generalization gradients. Gradients are plotted for 9 equally spaced stimuli along a stimulus dimension (S1 –S9). S1 serves as the CS- whereas S5 serves as the CS+.

Although preliminary evidence suggests that individual generalization rules determine the effectiveness of generalization of GS extinction, it is unknown whether this finding extends to safety behaviors. Safety behaviors refer to behaviors that prevent or minimize the onset of threat when confronting a warning signal. In the laboratory, safety behaviors are modelled by behavioral responses performed during CS+ presentations that prevent US onset i.e., US-avoidance (Krypotos, Vervliet, & Engelhard, 2018; Pittig, Wong, Glück, & Boschet, 2020). Safety behaviors are usually adaptive when they prevent realistic harm; however, they become pathological when they persist in the absence of threat in addition to inflicting impairments and results in cost to the individual (i.e., costly safety behaviors). Empirical studies (Lovibond, Mitchell, Minard, Brady, & Menzies, 2009; Pittig, 2019; Rattel, Miedl, Blechert, & Wilhelm, 2017; Volders, Meulders, de Peuter, Vervliet, & Vlaeyen, 2012) showed that if US-avoidance was constantly engaged to the CS+ during extinction trials, participants would attribute the absence of a US to their US-avoidance, thus protecting them from extinction learning (i.e., protection from extinction). Therefore it is of clinical interest to reduce safety behaviors as it likely reduces protection from extinction and the inflicted impairments, thus enhancing the effectiveness of exposure-based treatments (see also Helbig-Lang & Petermann, 2010; Wells et al., 1995). This leads to the speculation that different generalization rules might lead to different degrees of extinction learning to the GS in extinction, thus leading to differential reduction of safety behaviors and its generalization. Therefore individuals having weak threat expectancy to the extinction stimulus may show limited generalization of reduction in safety behaviors. This limited generalization of safety behaviors reduction is likely to lead to protection from extinction to a broader range of stimuli or situations, impeding responses to treatment.

The current study therefore sought to examine how presenting a GS to participants with different generalization rules (assessed via rule-based generalization) differentially affects generalization of GS extinction in safety behaviors. We employed a differential fear and avoidance conditioning procedure, with the stimulus in the middle of a green-blue stimulus dimension serving as the CS+ (aqua-color) and the stimulus at the green end serving as the CS- (parallel to S1 in Fig. 1). After acquiring stronger conditioned fear and costly safety behaviors to the CS+ compared to the CS- (i.e., differential responding to the CSs), all nine stimuli along the dimension were then presented to assess the generalization of costly safety behaviors. In the next phase, the GS at the other end of the dimension (blue color; as in S9 in Fig. 1) was presented in response prevention extinction, i.e. an unreinforced stimulus was presented without the option for costly safety behaviors. In the final test phase, generalization of GS extinction was assessed by presenting the full range of GSs along the stimulus dimension again, and costly safety behaviors to each of these GSs were measured. After the experiment, we assessed whether participants adopted any rules to guide their responses during the generalization test prior to GS extinction. We expected that most participants would report entertaining either a similarity- or a linear-based rule. It was expected that participants who responded with a linear rule would engage in stronger costly safety behaviors to the GS at the extreme blue end of stimulus dimension (S9). Thus, presenting this GS in response prevention extinction would lead to strong expectancy violation, and lead to robust generalization of GS extinction in costly safety behaviors. In contrast, the same GS would evoke limited expectancy violation in participants who responded with a similarity rule, therefore resulting in little if any generalization of GS extinction effect. The current study also explored whether risk factors, such as trait

anxiety or intolerance of uncertainty, would have any impact on the generalization of costly safety behaviors before and after GS extinction (cf. Flores, López, Vervliet, & Cobos, 2018; Kaldewaij et al., 2021; Pittig & Scherbaum, 2020).

2. Method

Pre-registration and data are available at OSF (https://osf.io/xfj26).

2.1. Participants

Undergraduate students from the University of Würzburg or the general community were recruited and were compensated with partial course credit or 10€ for 1 hour of participation. Of note, participants also received extra monetary reward depending on their avoidance ratings throughout the experiment. We followed a recruitment strategy similar to our previous study (Wong et al., 2020), in which we stopped recruitment when the two expected rule groups had at least 35 participants after exclusions. A simulation-based power analysis (Kumle, Võ, & Draschkow, 2021) revealed that 35 participants in each rule group provided 92% power to detect a group difference of the smallest effect size of interest (see Kumle et al., 2021) in post-extinction US-avoidance gradients (see the pre-registered report for more details). This recruitment strategy led to a total sample of 84 participants. This study was approved by the Ethics Committee of the Institute of Psychology at the University of Würzburg in accordance to the Declaration of Helsinki.

2.2. Apparatus and materials

We employed a subset of visual stimuli adapted from Lovibond, Lee, and Hayes (2020). This subset of stimulus dimension consisted of nine rectangles (each measuring 354 × 177 pixels) varying in hue along a green-blue dimension (Fig. 2A). The hue values ranged from 0.42 to 0.58 and were equally spaced between stimuli. The saturation and brightness were kept constant across all stimuli at 1 and 0.75, respectively. S5, the aqua color rectangle served as CS+, whereas S1, the green color rectangle, served as CS-. All stimuli were individually presented on a white screen.

A computer equipped with Presentation software (Neurobehavioral Systems Inc., Berkeley, CA, Version 20.1) presented the instructions, all visual stimuli, and recorded the US expectancy ratings and US-avoidance responses. Skin conductance were measured via two Ag/AgCl electrodes at a sampling rate of 1000 Hz, which was recorded by another computer equipped with BrainVision Recorder (Brain Products GmbH, Gilching, Germany). The US was an electric stimulation with a duration of 625 ms, generated by a DS7A Digitmer stimulator and delivered through a bar electrode attached to participants' wrist.

2.3. Procedure

After providing informed consent, we attached skin conductance electrodes filled with isotonic gel to the hypothenar muscles on the palm of participants' non-dominant hand. Next, participants filled in the German version of Intolerance of Uncertainty scale (UI-18; Freeston, Rhéaume, Letarte, Dugas, & Ladouceur, 1994; Gerlach, Andor, & Patzelt, 2008) and the short German version of Depression, Anxiety, and Stress Scale (DASS-21; Lovibond & Lovibond, 1995; Nilges & Essau, 2015). The UI-18 measures cognitive, emotional, and behavioral responses to uncertainty (see Carleton, Norton, & Asmundson, 2007), whereas the DASS-21 measures and discriminates three different constructs: depression, anxiety, and stress. Afterwards, US electrodes were attached to participants' wrist on the non-dominant hand.

Participants were then led through a US workup procedure, in which an US intensity of 0.2 mA was gradually increased until it reached a level that was perceived as 'definitely unpleasant but not painful'. Next, we carried out a reward-matching procedure, which aimed to identify a reward that was neither too low nor too high to prevent ceiling or floor effects on US-avoidance (see Schlund et al., 2016; Wong & Pittig, 2022; Wong and Pittig, 2022b). This procedure entailed a series of questions "Are you willing to tolerate the selected level of electric stimulation if you are given €_?", with the amount of reward ranging from 5 to 31 cents in odd numbers (i.e., 5 cents, 7 cents, 9 cents, …, 31 cents) presented in a randomized order. A total of 14 questions were presented. Participants were prompted to answer either 'Yes' or 'No' to each of these questions. The amount of reward exactly between the highest amount that received a 'No' and the lowest amount that received a 'Yes' (i.e., averaged between these two values) was selected as the incentive for US-avoidance disengagement. For instance, if an individual participant was unwilling to tolerate an US up to 19 cents (i.e., answering 'No' up to 19 cents), but was willing to tolerate it when given 21 cents or more (i.e., answering 'Yes' from 21 cents onwards), the amount in between (20 cents) would be chosen as the maximum amount of competing reward per trial.

This experiment consisted of five consecutive phases: *Pavlovian fear acquisition, Costly US-avoidance acquisition, Pre-extinction generalization test, GS extinction, and Post-extinction generalization test* (see Table 1).

Pavlovian fear acquisition. Participants were instructed that some geometric shapes would appear on the screen, which might or might not be followed by a US (all exact instructions were reported in the Supplementary Materials). They were instructed to learn the relationship between the shapes and the US (cf. Mertens, Boddez, Krypotos, & Engelhard, 2021). Twelve trials of CS+ (S5) and twelve trials of CS- (S1) were presented alongside a US expectancy scale. The US expectancy scale ranged from 0% to 100% in which 0% indicates certain no US and 100% indicates certain US. Participants were asked to indicate their US

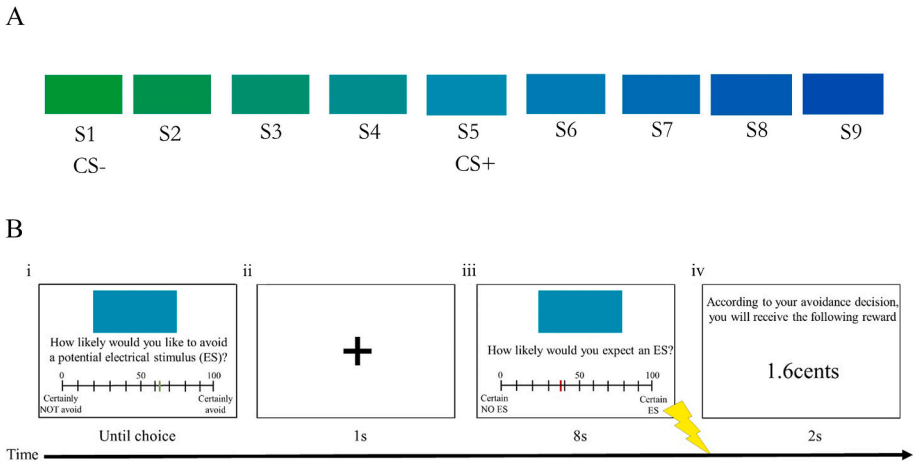

Fig. 2. (A) Green-blue stimulus dimension: S1 served as the CS- whereas S5 served as the CS+. The stimulus labels (S1–S9) were not presented in the experiment. (B) Trial structure when US-avoidance was available. (i) Participants were prompted to indicate US-avoidance. (ii) A black fixation cross was presented in the centre of the screen for 1 s (iii) The stimulus was presented again along with an US expectancy scale for 8 s. Participants had to indicate their US expectancy ratings. In Costly US-avoidance acquisition, an electrical US would be administered immediately after CS+ offset depending on the US-avoidance made. In Pre-extinction generalization test and Post-extinction generalization test, no US was administered regardless of US-avoidance made. (iv) The reward feedback appeared on screen for 2 s. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
S1 to S9 indicate the different stimuli across the green-blue stimulus dimension; + indicates US presentation; - indicates US omission; * indicates the availability of US-avoidance; + in parentheses indicates that US presentation depends on US-avoidance; € in parentheses indicates the amount of competing reward depends on US-avoidance; number in parentheses indicates the number of trials per trial type.

Pavlovian fear acquisition	Costly US-avoidance acquisition	Pre-extinction generalization test	GS extinction	Post-extinction generalization test
S5+ (9) S5- (3) S1- (12)	S5* (+, €) (8) S1*- (€) (8)	[S1 – S9]*- (€) (1)	S9- (6)	[S1 – S9]*- (€) (1)

expectancy during CS presentations. The CSs and US expectancy scales were presented on screen for 8 s. Of note, the CS+ trials were reinforced by a US at a 75% rate (i.e., 9 out of 12 trials) whereas the CS- trials were never reinforced. The presentation order was pseudo-randomized such that the same trial type was not presented more than twice in a row, and the first and last trials of CS+ were always reinforced. The inter-trial intervals (ITIs) varied between 15 and 18 s and were applied to all the following phases.

Costly US-avoidance acquisition. Participants were informed that they can avoid the US by indicating their avoidance ratings at the US-avoidance scale presented alongside the CSs. Avoidance ratings were negatively proportional to the chance of US presentation. For instance, an avoidance rating of 60% would result in a 60% chance of US omission for that particular trial. Participants were, however, not informed that US-avoidance ratings determined US omission on CS+ trials only, whereas CS- trials were not reinforced regardless of US-avoidance. Furthermore, each trial included a competing reward that was determined individually by the initial reward-matching procedure. The amount of this reward was, however, inversely proportional to the US-avoidance made. For instance, an avoidance rating of 60% would lead to a gain of 40% of the maximum reward. Participants were instructed that all rewards gained throughout the experiment would be paid off after the conditioning task. This phase consisted of 8 CS+ and 8 CS- trials. On every trial, a CS and the US-avoidance scale were presented on screen until response. After US-avoidance ratings had been indicated, a fixation cross appeared for 1 s. After that, the same CS was presented alongside the US expectancy scale for 8 s, and participants were prompted to indicate their US expectancies. Immediately after CS offset, a US would either be presented or omitted depending on the US-avoidance made and CS type. Specifically, US presentation was determined probabilistically on each CS+ trial. Reward feedback was then presented for 2 s (see Fig. 2B).

Pre-extinction generalization test. This phase continued seamlessly from the previous phase and was carried out under extinction. All nine stimuli (S1 – S9) were presented on screen once in a randomized order. The trial structure was identical to the previous phase, with the exception that none of the stimuli were reinforced by a US regardless of US-avoidance made, that is, none of the CSs and GSs were reinforced. Of note, the dimensional measure of avoidance was able to minimize confounding effect of extinction learning to the test stimuli in this phase and in *Post-extinction generalization test* (see Wong & Pittig, 2022a; Wong and Pittig, 2022b).

GS extinction. Before this phase started, participants were asked to take a 30 s break. The word "Pause" appeared on screen throughout the break. At the end of the break, participants were informed that avoidance was unavailable for the following phase, and they were asked to continue indicating their US expectancies only. S9 was presented for 6 trials in this phase, the same as in previous studies examining the effect of expectancy violation in generalization of GS extinction (Struyf et al., 2018; Wong et al., 2020). Each stimulus was presented on screen alongside the US expectancy scale for 8 s.

Post-extinction generalization test. Participants were informed that US-avoidance was available again, and they could choose to potentially prevent the US that might followed the stimuli. Participants were also reminded their US-avoidance rating determined the amount of reward on each trial. All nine stimuli along the stimulus dimension (S1 to S9) were presented once each in a randomized order. The trial structure was identical to *Pre-extinction generalization test*. None of these stimuli were reinforced by a US.

After the conditioning task, participants were asked to fill in a questionnaire assessing potential rules they might have employed during the task. The experimenter wrote down the US-avoidance ratings that the individual participant had made to S1 and S9 during *Pre-extinction generalization test*. Participants were explicitly asked whether they came up with a rule or strategy *before the break* (before *GS extinction* took place). In prior studies (Wong et al., 2020; Wong & Lovibond, 2017), we found that participants entertained different rules once extinction learning took place. Thus, we specifically asked for the rules inferred before the break to probe the exact generalization rule employed before *GS extinction*. To indicate generalization rules used during the conditioning task, participants were given a force choice question with five alternative options: Blue linear (US administration is more likely the *bluer* the rectangle), Green linear (US administration is more likely the *greener* the rectangle), Similar (US administration is more likely the *more* aqua *color* the rectangle), No rule (US administration is *not related to the color* of the rectangles), and Other (None of the above). Participants would be categorized into rule groups according to their responses to the five-alternative force-choice question in the post-experimental questionnaire.

2.4. Scoring and analysis

Although skin conductance was recorded online throughout the entire experiment, only skin conductance measured during the 8 s stimulus presentation (i.e., when participants were prompted to indicate their US expectancies) were included for analysis. We applied a 50 Hz notch filter and a 1 Hz high-pass filter to the SCR data. The SCRs were calculated by identifying the difference between response peak and the corresponding trough (i.e., minimum response) in the interval of 1 s after stimulus onset to stimulus offset (i.e., SCRs peaks were measured from 1s to 8s after CS onset). We then square root transformed the SCRs to reduce skewness (Boucsein et al., 2012).

We conducted linear mixed models for all analyses within a traditional frequentist framework. Analyses that examined the differences (i.e., differences between groups, differences between phase) in generalization gradients were followed up by a recent novel technique, augmented Gaussian modelling (Lee, Mills, Hayes, & Livesey, 2021). This analysis fits an augmented (i.e., asymmetric) Gaussian function to individual gradients in a hierarchical model and estimates parameters corresponding to the height, mean, and width on the left- and right-hand side of the function. The height parameter represents the height of the gradient peak, the mean parameter represents the peak location on the stimulus dimension, and the two width parameters represent the width of the gradient on either side of the gradient peak (width- for the side containing the CS-, and width+ for the other side). This analysis quantifies multiple descriptive features of generalization gradients and provides a way to test exactly how generalization changes between groups or phases. For example, broadening of gradients could be captured via group differences in one, or both, width parameters.

The analysis was conducted in a Bayesian framework, producing posterior distributions for each of the four parameters (mean, height, and two width parameters). Contrasts were tested by simulating

differences between phases or subgroups, followed by calculating the proportion (p(direction)) of the posterior samples that were above or below 0 (whichever was most probable). Specifically, a p(direction) larger than 0.975 is considered as equivalent to a significant difference (Makowski, Ben-Shachar, & Lüdecke, 2019). We only reported estimated parameters that reached this significant threshold. The analyses were separated into three parts: manipulation check, main hypotheses, and exploratory analyses. All analyses were pre-registered (see https://osf.io/xfj26).

2.5. Manipulation check

We first analyzed whether participants acquired stronger conditioned fear and costly US-avoidance to the CS+ compared to the CS- in *Pavlovian fear acquisition* and *Costly US-avoidance acquisition*, respectively. To this end, US expectancy, SCRs, or US-avoidance served as dependent variable, whereas CS type (CS+ vs CS-) and Trial (a linear trend repeated measures across trials) served as fixed effects. Participants served as a random effect; this was applied to all the linear mixed models.

2.6. Main hypotheses

First, to examine the different shapes in generalization gradients in both *Pre-extinction generalization test* and *Post-extinction generalization test*, we used two orthogonal polynomial trend repeated measures contrasts across test stimuli as fixed effects. Specifically, a linear trend repeated measures contrast across test stimuli (represented by "Stimulus"), and a quadratic trend repeated measures contrast across test stimuli (represented by "Stimulus^2"). To evaluate any group differences in the generalization gradients, Group (Linear vs Similarity) served as another fixed effect. In sum, US-avoidance served as dependent variable, whereas Stimulus, Stimulus^2, and Group served as fixed effects. The interactions between the polynomial trend repeated measures and Group were of primary interest to evaluate any group differences in the shapes (linearity or curvature) of the generalization gradients. It is expected that the Linear group would show a greater linear trend across stimuli compared to the Similarity group (as indicated by a Group*Stimulus interaction), whereas the Similarity group would show a stronger quadratic trend across stimuli compared to the Linear group (as indicated by a Group*Stimulus^2 interaction). In addition, we included an extra contrast that evaluated the group differences in differential US-avoidance to the CS+ and the CS- in *Post-extinction generalization test*, in which the Linear group was expected to show a weaker differential responding to the CSs compared to the Similarity group (as indicated by a Group*CS type interaction). Thus, CS type and Group served as fixed effects in this contrast. This additional contrast was non-orthogonal to the two polynomial trend repeated measures contrasts (i.e., the contrasts were correlated), therefore Bonferroni-corrections were applied to the p-values for the contrasts in *Post-extinction generalization test*.

Second, group differences in the rate of extinction were analyzed in separate linear mixed models. Rate of extinction was analyzed via Trial (a linear trend repeated measures across trials) i.e., the slope of change in responding to the GS during extinction. To this end, US expectancy or SCRs served as dependent variable, whereas Trial and Group served as fixed effects. It was expected that the Linear group would show stronger responding to the extinction stimulus on early trials, followed by faster decrease in responding compared to the Similarity group (as indicated by a Group*Trial interaction).

Third, a separate cross-phase linear mixed model specifically examined the change in generalization gradients prior to and after GS extinction between groups. In this model, US-avoidance served as dependent variable, whereas Phase (*Pre-extinction generalization test* vs *Post-extinction generalization test*), Stimulus, Stimulus^2, and Group served as fixed effects. Furthermore, we included a non-orthogonal contrast to examine whether the groups differed in differential US-avoidance to the CSs prior- compared to post-extinction (i.e., whether the groups differed in how effectively GS extinction generalized to the CS+). To this end, US-avoidance served as dependent variable, whereas Phase, CS type, and Group served as fixed effects. All contrasts in this model were Bonferroni-corrected. It was expected that the Linear group would show a stronger decrease in linear trend across stimuli after GS extinction compared to the Similarity group (as indicated by a Group*Stimulus*Phase interaction; see Wong et al., 2020), in addition to a greater decrease in differential responding to the CSs (as indicated by a Group*CS type*Phase interaction).

Fourth, the group differences in generalization gradients in *Pre-extinction generalization test* and *Post-extinction generalization test* were further evaluated by two separate analyses (i.e., group differences in *Pre-extinction generalization test* only and in *Post-extinction generalization test* only), and estimating the four parameters (mean, height and two width parameters) of an augmented Gaussian for each analysis.

2.7. Exploratory analyses

To explore the effect of trait anxiety or intolerance of uncertainty on the gradients in *Pre-extinction generalization test* and *Post-extinction generalization test*, we separately added these two constructs as a continuous fixed effect to the aforementioned models. That is, US-avoidance served as dependent variable, whereas Stimulus, Stimulus^2, Group, and trait anxiety/intolerance of uncertainty served as fixed effects.

Of note, although US expectancy ratings and SCRs were recorded after US-avoidance had been made in the test phases (*Pre-extinction generalization test & Post-extinction generalization test*), these analyses were moved to Supplementary Materials. This was because these measures were measured after US-avoidance had been made and thus were influenced by the level of US-avoidance. Thereby, these measures did not fully reflect one's generalization rules if US-avoidance would have been unavailable (cf. Lee, Mills, et al., 2021; Wong & Lovibond, 2017). The degree of significance was reported with Satterthwaite approximation for degrees of freedom (Satterthwaite, 1941). All analyses were carried out using R (R Core team, 2022), with *lmer* package for frequentist models (Bates, Mächler, Bolker, & Walker, 2015), and *rstan* (Stan Development Team, 2018) and *BayesTestR* (Makowski, Ben-Shachar, & Lüdecke, 2019) for the augmented Gaussian modelling.

3. Results

Exclusion criteria were pre-registered at https://osf.io/xfj26. Overall, data of three participants were incomplete due to technical issues, and thus were excluded from the final sample. This led to a sample of 81 participants.

Questionnaire coding: Participants were categorized into rule groups according to their responses to the five-alternative force-choice question in the post-experimental questionnaire. As expected, most participants identified either a Similarity rule (n = 38) or a Linear rule (n = 35). A small number of participants failed to identify any clear rules (No rule; n = 8). Participants identifying a Similarity rule reported stronger US-

Table 2
Demographic and questionnaire data. Means (standard deviations).

	Similarity group (n = 38)	Linear group (n = 35)	F or χ^2	p
Age	24.58 (3.97)	23.14 (3.80)	2.49	.119
Sex - Female	29 (76.3%)	27 (77.1%)	18.09	.319
US intensity (mA)	1.31 (0.61)	1.42 (2.93)	0.036	.850
DASS 21-Anxiety	2.79 (3.17)	5.14 (4.74)	6.31	**.014**
DASS 21-Depression	4.79 (5.33)	6.46 (5.78)	1.64	.204
DASS 21-Stress	7.11 (5.42)	10.69 (8.03)	5.06	**.028**
UI-18	40.74 (10.75)	43.86 (11.72)	1.41	.239

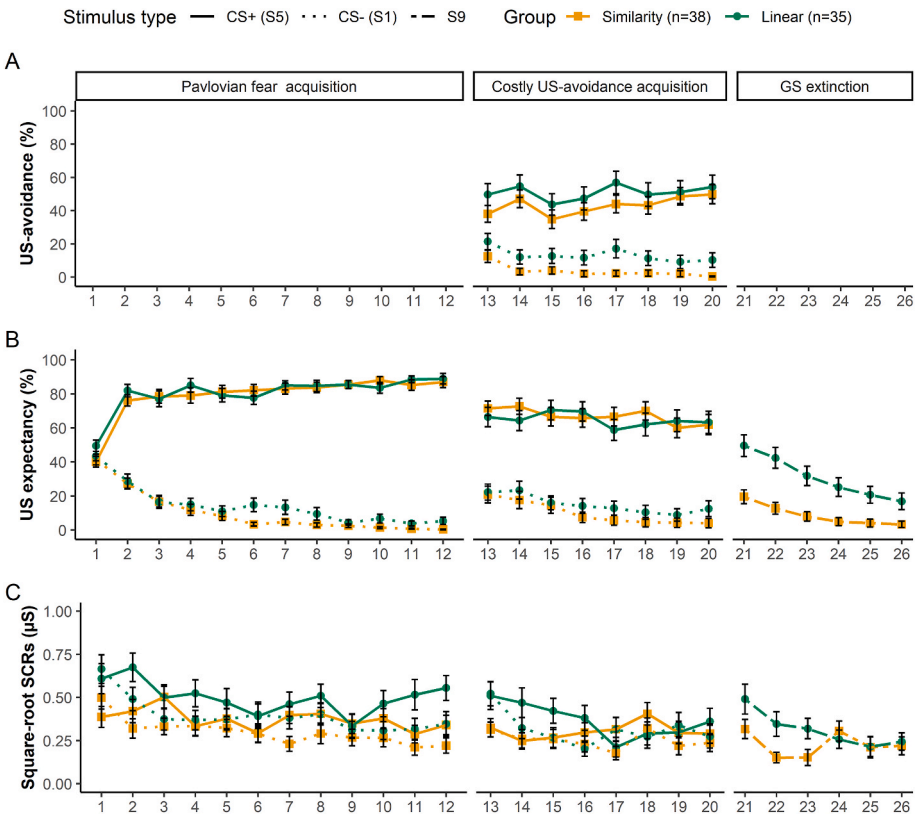

Fig. 3. US-avoidance (top panel), US expectancy ratings (middle panel), and square-root SCRs (bottom panel) in *Pavlovian fear acquisition*, *Costly US-avoidance acquisition*, and *GS extinction*. Error bars indicate the standard error of the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

avoidance when the stimulus was more perceptually similar to the CS+ whereas participants identifying a Linear rule reported stronger US-avoidance when the stimulus was bluer. As pre-registered, we excluded participants in the No rule group from statistical analyses due to the small sample size (and thus the lack of statistical power). The final sample thus comprised 73 participants (data available at https://osf.io/b2gzv/). Groups did not differ in age, sex, US intensity, depression, or intolerance of uncertainty. However, the Linear group scored higher in trait anxiety and stress compared to the Similarity group. (see Table 2).

3.1. Manipulation check

Pavlovian fear acquisition. All acquisition data were analyzed in linear mixed models with CS type and Trial as fixed factors. Fig. 3B shows the mean US expectancy ratings across acquisition trials in each rule group. Main effects were not reported if the higher-order interactions were significant. Averaged across Group, participants developed higher US expectancy ratings to the CS+ across trials, whereas an opposite pattern was observed to the CS-. This pattern was confirmed by a significant interaction between CS type and Trial, $b_{CS\ type*Trial} = 733.64$, SE = 38.11, $p < .001$. No interactions involving Group reached significance (smallest $p = .131$), suggesting no evidence for any group differences in the acquisition of differential expectancy ratings to the CSs.

Fig. 3C shows the square root SCRs across acquisition trials in each rule group. Averaged across Group, stronger SCRs to the CS+ compared to the CS- were developed across acquisition trials, supported by a significant interaction between CS type and Trial, $b_{CS\ type*Trial} = 1.36$, SE = 0.63, $p = .031$. There was no evidence for any group differences in the acquisition of SCRs to the CSs (smallest $p = .554$).

Costly US-avoidance acquisition. Fig. 3A shows the mean US-avoidance across trials in each rule group in *Costly US-avoidance acquisition*. Averaged across Group, participants showed stronger US-avoidance to the CS+ compared to the CS-, in which this pattern became more pronounced across trials. This pattern was supported by a significant interaction between CS type and Trial, $b_{CS\ type*Trial} = -164.28$, SE = 47.88, $p < .001$. There was no evidence that the groups differed in the differential acquisition of costly US-avoidance to the CSs (smallest $p = .521$).

In sum, participants successfully acquired differential conditioned fear and costly US-avoidance to the CSs without any group differences.

3.2. Main hypotheses

Pre-extinction generalization test. Group differences in both generalization test phases were analyzed in linear mixed models with polynomial repeated measures contrasts (Stimulus & Stimulus^2) and Group. Fig. 4A shows the generalization gradient of US-avoidance in each rule group.[1] The Similarity group exhibited a bell-shaped gradient with the strongest US-avoidance to the CS+, whereas the Linear group revealed an S-shaped gradient with strong US-avoidance to stimuli right of the CS+. The Linear group showed a stronger linear trend across stimuli than the Similarity group, $b_{Stimulus*Group} = 318.58$, SE = 48.95, $p < .001$, suggesting that there was more linearity in the gradient in the Linear group than the Similarity group. The Similarity group, on the other hand, exhibited a stronger quadratic trend across stimuli than the Linear group, $b^2_{Stimulus*Group} = -142.81$, SE = 48.95, $p = .004$, suggesting that there was greater curvature in the gradient in the Similarity group than the Linear group.

GS extinction. We examined whether there were any group

[1] Gradient for each individual participant can be seen in the Supplementary Materials.

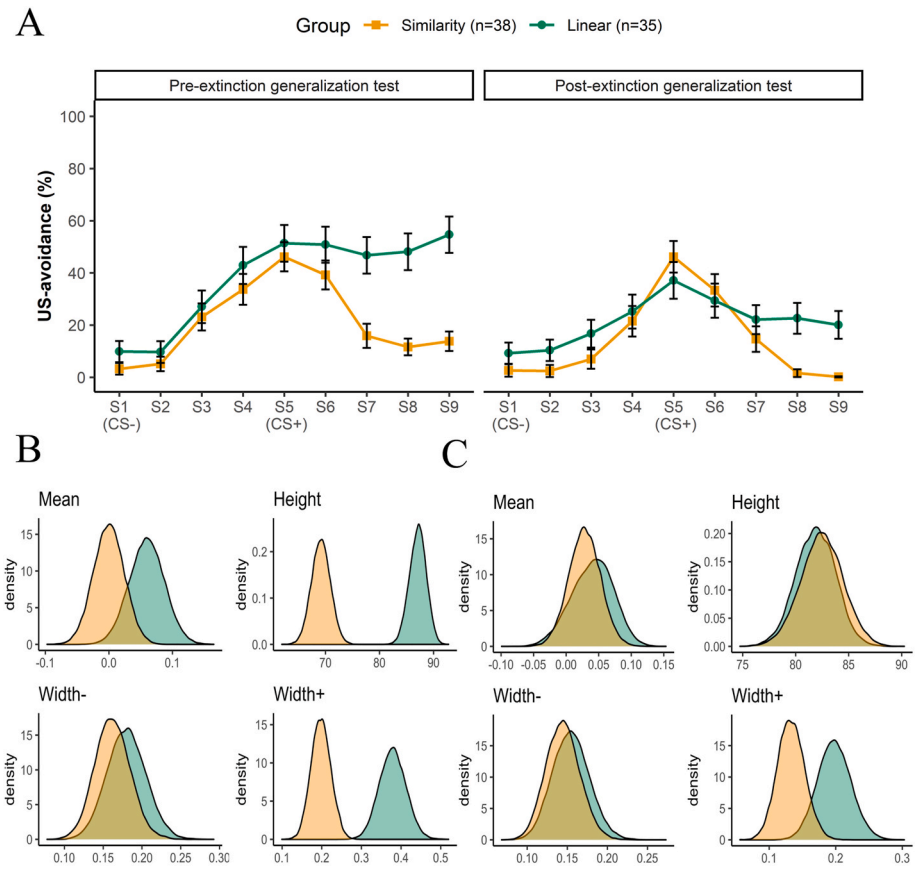

Fig. 4. Top panel: (A) US-avoidance made during *Pre-extinction generalization test* (left panel) and during *Post-extinction generalization test* (right panel). Error bar indicates the standard error of the mean. Bottom panel: Posterior distributions for the four estimated parameters in (B) *Pre-extinction generalization test* and in (C) *Post-extinction generalization test*. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

differences in the rate of extinction to the extinction stimulus (S9). This was examined by linear mixed models with Group and Trial as fixed factors. Fig. 3B shows the mean US expectancy ratings across extinction trials in each rule group. As expected, the Linear group exhibited higher US expectancy ratings to the extinction stimulus on early trials compared to the Similarity group, and a more rapid extinction. This pattern was supported by a significant interaction between Group and Trial, $b_{Group*Trial} = -127.96$, SE = 30.97, $p < .001$. Follow-up analyses revealed that both groups showed a significant decrease in US expectancies to the extinction stimulus across trials (smallest $p < .001$).

Fig. 3C shows the mean square-root SCRs across extinction trials in each rule group. Similar to the expectancy data, the Linear group showed stronger SCRs to the extinction stimulus on early extinction trials than the Similarity group. This pattern was supported by a significant interaction between Group and Trial, $b_{Group*Trial} = -1.57$, SE = 0.56, $p = .006$. Follow-up analyses further revealed that only the Linear group exhibited a significant decrease in SCRs to the extinction stimulus across trials, $b_{Trial} = -1.72$, SE = 0.41, $p < .001$, but not the Similarity group, $b_{Trial} = -0.16$, SE = 0.39, $p = .687$.

Post-extinction generalization test. Fig. 4A shows the post-extinction generalization gradient of US-avoidance in each group. Both groups exhibited a peaked gradient with the peak responding to CS+. The Linear group showed a weaker quadratic trend across stimuli compared to the Similarity group, $b^2_{Stimulus} = 161.08$, SE = 46.77, $p = .002$, suggesting that the Linear group had a less curve gradient than the Similarity group. Follow-up analyses revealed that both groups exhibited a significant peaked gradient (smallest $p < .001$). However, there was no evidence for any group differences in linearity, $b_{Stimulus*Group} = 88.51$, SE = 46.77, $p = .179$.

In addition, averaged across Group, US-avoidance was stronger to the CS+ compared to the CS-, $b_{CStype} = 36.04$, SE = 4.47, $p < .001$, suggesting the persistence of differential US-avoidance to the CSs post-extinction. There was, however, no evidence for any group differences in differential US-avoidance to the CSs, $b_{CS\ type*Group} = 15.67$, SE = 8.75, $p = .233$.

Cross-phase analyses: The group differences in generalization of GS extinction. For this analysis, a linear mixed model was employed with three non-orthogonal contrasts. The first two contrasts compared the change in group differences in generalization gradients before and after GS extinction, whereas the third contrast compared the change in group differences in differential US-avoidance to the CSs before and after *GS extinction*. With regard to the first two contrasts, the US-avoidance gradients decreased in linearity after *GS extinction* as compared to before GS extinction, presumably due to a decrease in US-avoidance to stimuli right of the CS+. This pattern was more pronounced in the Linear group compared to the Similarity group, supported by a significant 3-way interaction involving Group, Phase, and Stimulus, $b_{Group*Phase*Stimulus} = -325.36$, SE = 98.70, $p = .003$. Follow-up analyses further confirmed that the Linear group showed a significant decrease in linearity after GS extinction, $b_{Phase*Stimulus} = -384.76$, SE = 71.21, $p < .001$, whereas there was no evidence of changes in linearity after GS extinction in the Similarity group, $b_{Phase*Stimulus} = -59.40$, SE = 68.34, $p = .999$. For the second contrast, no interactions involving Stimulus^2 reached significance (all Bonferroni-adjusted $p = .999$).

For the third contrast, there was no evidence for any changes in differential US-avoidance to the CSs after *GS extinction*, $b_{Phase*CS\ type} = -6.22$, SE = 5.58, $p = .798$, nor there was evidence for any group differences in differential US-avoidance to the CSs before and after *GS extinction*, $b_{Group*Phase*CS\ type} = -14.23$, SE = 11.02, $p = .594$.

To further explore the rule group differences pre- and post-

extinction, we modelled the gradients as augmented Gaussians (Lee, Mills, et al., 2021) and estimated the height, mean, width-, and width + parameters for the Linear and Similarity subgroups, for each generalization phase. Fig. 4B and C shows the posterior distributions for the *group differences* (Linear vs. Similarity) in each parameter pre-extinction (Fig. 4B) and post-extinction (Fig. 4C). Examining the empirical gradients (Fig. 4A), the Linear gradient appears more linear and higher than the Similarity gradient pre-extinction, but becomes more similar to the Similarity gradient post-extinction. In line with these observations, prior to extinction (Fig. 4B), the posterior densities suggest a clear group difference in the height and width+ parameters, while post-extinction (Fig. 4C), the group difference disappears for the height parameter and appears reduced for the width+ parameter. The visual differences in the posteriors were supported by the computed p(direction) statistics for the group difference, which were greater than 0.99 for both height and width+ pre-extinction. However post-extinction, there was a group difference for the width+ parameter only, with p(direction) greater than 0.975 (see Supplemental materials). These results suggest that extinction of a GS reduced US-avoidance to stimuli similar to the extinguished GS, and changed the shape of the gradient from linear to more peaked. However, a group difference in the width+ parameter remained, suggesting that even post-extinction, the Linear group still exhibited stronger US-avoidance generalization to stimuli right of the CS+ (see Supplementary Materials for the complete analysis for the augmented Gaussian modelling).

3.3. Explorative analyses

Pre-extinction generalization test. These exploratory analyses were examined in linear mixed models with polynomial repeated measures contrasts, Group, and risk factors (trait anxiety or intolerance of uncertainty) as fixed factors, whereas US-avoidance served as dependent variable. Fig. 5A and B show the effect of trait anxiety and IU in *Pre-extinction generalization test*, respectively. We observed a three-way interaction involving Stimulus^2, Group, and Anxiety, $b^2_{\text{Group*Anxiety}} = 29.59$, SE = 13.13, $p = .025$, suggesting as anxiety increases, the group differences in quadratic trend across stimuli increases. Follow-up analyses suggested that this pattern was isolated to the Similarity group, $b^2_{\text{Anxiety}} = -21.39$, SE = 10.76, $p = .047$. In other words, an increase in trait anxiety was associated with a broader gradient in the Similarity group. In contrast, no interactions involving IU reached significance (smallest $p = .146$).

Post-extinction generalization test. Fig. 5A and B show the effect of trait anxiety and IU on generalization in *Post-extinction generalization test*, respectively. With regard to trait anxiety, there was a group difference in how trait anxiety was associated with Stimulus^2, $b^2_{\text{Group*Anxiety}} = -43.43$, SE = 12.44, $p = .002$. Follow-up analyses revealed that an increase in trait anxiety was associated with a weaker quadratic trend across stimuli in the Similarity group, $b^2_{\text{Stimulus *Anxiety}} = -39.70$, SE = 10.20, $p < .001$, but not in the Linear group, $b^2_{\text{Stimulus *Anxiety}} = 3.73$, SE = 7.13, $p = .999$. This suggested that higher trait anxiety was associated with a flatter gradient, but only in the Similarity group. There was no evidence that trait anxiety was associated with differential US-avoidance to the CSs in the Post-extinction generalization test (smallest Bonferroni-adjusted $p = .446$).

An increase in IU was associated with a stronger quadratic trend across stimuli averaged across Group, supported by a significant interaction between Stimulus^2 and IU, $b^2_{\text{Stimulus *IU}} = -5.51$, SE = 2.11, $p = .027$. This suggested that individuals high in IU were associated with less flattening in their gradients post-extinction (i.e., a reduction in US-avoidance generalization after GS extinction), suggesting that IU is associated with poorer generalization of GS extinction. No other interactions involving IU reached significance (smallest Bonferroni-adjusted $p = .409$).

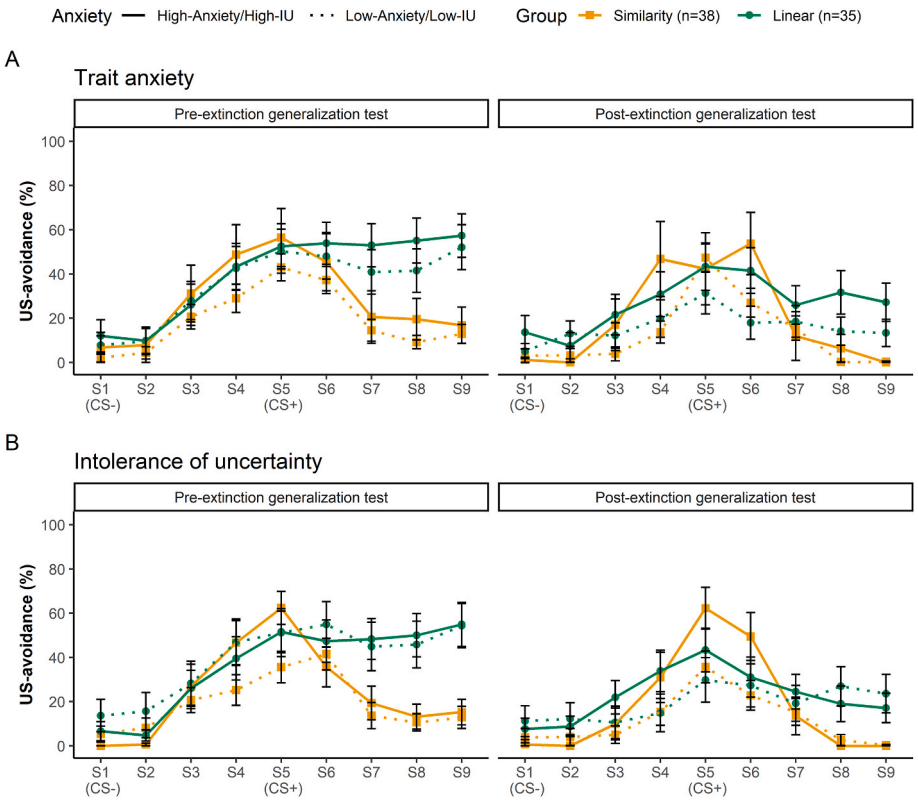

Fig. 5. The effect of trait anxiety (A) or intolerance of uncertainty (B) on US-avoidance generalization pre- and post-extinction. Trait anxiety/Intolerance of uncertainty was divided into high and low values (via median split) for descriptive purpose. Error bars indicate standard error of the mean. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Cross-phase analyses: The effect of risk factors on generalization of GS extinction. There was no evidence that trait anxiety was associated with changes in US-avoidance gradients nor differential US-avoidance to the CSs before and after GS extinction (smallest Bonferroni-adjusted p = .999).

IU, on the other hand, had a descriptively more pronounced association with elevated generalized US-avoidance after *GS extinction* compared to before GS extinction. This pattern, however, did not reach significance, $b_{Phase * IU} = 0.28$, SE = 0.14, $p = .116$. No other interactions involving IU reached significance (smallest Bonferroni-adjusted p = .128).

4. Discussion

Empirical studies have shown that extinction learning to a GS does not always effectively generalize to other fear-related stimuli (e.g., Vervliet et al., 2004; Vervoort et al., 2014; Wong & Lovibond, 2020). The current study aimed to examine the impact of generalization rules on the levels of generalization of GS extinction; we examined whether this effect extended to costly US-avoidance. The rule differences in generalization of GS extinction were presumably due to differences in expectancy violation.

We found that the US-avoidance generalization gradients prior to GS extinction were highly consistent with participants' self-reported rules. In particular, participants reporting a similarity rule exhibited peak US-avoidance to the CS+, with decreasing US-avoidance to stimuli that were further away from the CS+ along the stimulus dimension. Participants reporting a Linear rule, on the other hand, showed strong generalized US-avoidance to bluer stimuli. These patterns aligned with our previous work that self-generated rules were highly consistent with the shape of generalization gradients (Ahmed & Lovibond, 2019; Lee et al., 2018; Wong & Lovibond, 2017, 2018), in addition to extending the findings to generalized US-avoidance (see also Wong and Pittig, 2022b). The current study also aligns with studies that have examined the role of higher-order cognitive processes in the generalization of US-avoidance. For example, participants were more likely to engage in US-avoidance to a novel word that was a synonym of the CS+ (Boyle, Roche, Dymond, & Hermans, 2016), or to novel stimuli that shared the same artificial category with the CS+ (Dymond et al., 2011, 2014). Combined, these findings offer strong support that higher-order cognitive processes play an important role in the generalization of US-avoidance.

As predicted, the strength of extinction learning to a GS was determined by participants' self-reported rules. The Linear group showed stronger conditioned fear on early extinction trials to the extinction stimulus compared to the Similarity group, as indexed by both US expectancy ratings and SCRs. This pattern aligned with our prediction that the extinction stimulus more strongly violated outcome expectancy, thus leading to stronger extinction learning in the Linear group. In contrast, little to no expectancy violation took place in the Similarity group, thus leading to limited extinction learning.

A key finding was that after GS extinction, the linear-based gradient in the Linear group transformed to a flatter, peaked gradient, featured by a reduction in US-avoidance to the GSs right of the CS+ (i.e., bluer stimuli). The augmented Gaussian modelling suggested that the height and breadth of US-avoidance generalization to stimuli right of the CS+ were both reduced after *GS extinction* in the Linear group. This suggested that extinction learning to the extinction stimulus generalized effectively to other GSs, at least to GSs that resemble both the extinction stimulus and the CS+. On the other hand, the shape of US-avoidance gradient in the Similarity group remained largely unchanged before and after *GS extinction*, due to limited expectancy violation during *GS extinction*. Overall, these findings are consistent with the notion of the Inhibitory learning model (Craske et al., 2014, 2018), which puts forward the idea that the stronger the expectancy violation, the stronger extinction learning is. The Linear group experienced strong expectancy violation during *GS extinction*, thus showed stronger reduction in generalization of US-avoidance to the GSs right of the CS+ compared to the Similarity group in the *Post-extinction generalization test*.

Despite the enhanced expectancy violation in the Linear group leading to greater generalization of US-avoidance reduction, extinction learning to this GS did not lead to a significant reduction in US-avoidance to the CS+. In addition, the Linear group still showed stronger generalization to stimuli right of the CS+ in the *Post-extinction generalization test*. One explanation is residual threat expectancies to the GS due to the limited amount of extinction trials, as observed in the apparent higher US expectancies on the last extinction trial in the Linear group compared to the Similarity group. Thus, this residual threat expectancy to the GS might have reduced the magnitude of generalization of GS extinction to the CS+. Another explanation is that a mere strong expectancy violation is not sufficient to generalize GS extinction to the CS+. Perhaps multiple interventions are required to further enhance the generalization of US-avoidance reduction, for instance, presenting multiple GSs in extinction (e.g., Shiban, Schelhorn, Pauli, & Muhlberger, 2015; Waters, Kershaw, & Lipp, 2018; Zbozinek & Craske, 2018), inducing positive affect during extinction (e.g., Zbozinek, Holmes, & Craske, 2015), or enhancing the generalization of safety learning to the CS- (Laing, Felmingham, Davey, & Harrison, 2022).

Exploratory analyses regarding the role of risk factors in US-avoidance generalization and its reduction were less clearcut. Trait anxiety was associated with a broader US-avoidance generalization before and after *GS extinction*, but only in the Similarity group. Given that limited extinction learning took place in the Similarity group during *GS extinction*, we interpreted that trait anxiety was associated with a broader generalization of US-avoidance but not with impaired generalization of GS extinction. The association between trait anxiety and enhanced US-avoidance generalization aligned with empirical findings of excessive US-avoidance in trait anxiety (e.g., Gorka, LaBar, & Hariri, 2016; Lau et al., 2012; Pittig & Scherbaum, 2020). With regard to IU, there was no evidence that it was associated with a broader US-avoidance generalization, in contrast to preliminary evidence that IU was associated with enhanced fear generalization (e.g., Bauer et al., 2020; Hunt, Cooper, Hartnell, & Lissek, 2019). After *GS extinction*, IU was associated with broader US-avoidance generalization gradients across groups, suggesting that it was related with impaired generalization of GS extinction. Overall, this preliminary, exploratory analysis suggested that trait anxiety was associated with broader US-avoidance generalization, whereas IU was associated with impaired generalization of GS extinction. However, one must be cautious with this interpretation, given that our power analyses were calculated based on observing the rule group differences in post-extinction gradients. It was possible that the current study was underpowered to examine the effect of risk factors on the acquisition and extinction of US-avoidance generalization (see Morriss, Zuj, & Mertens, 2021).

The current study categorized participants into a single rule group, assuming that each individual participant entertained one single rule. However, a recent study suggests that participants might entertain multiple rules simultaneously, and that modelling generalization as a mixture of rules was more predictive than a single rule (see Lee, Mills, et al., 2021). However, it is worth noting that Lee, Lovibond, Hayes, and Lewandowsky (2021) found that although most participants held a moderate-high degree of belief in both similarity and relational (i.e., linear) rules, these beliefs were negatively correlated, suggesting that most participants had a higher degree of belief in one of the two rules. Further, previous studies dividing participants into discrete rule subgroups have consistently showed differential generalization gradients consistent with the reported rules (Ahmed & Lovibond, 2019; Lee et al., 2018; Lovibond et al., 2020; Wong & Lovibond, 2017). Thus, separating participants into linear and similarity subgroups is a simple but valid method of analysing individual differences in generalization rules.

With regard to clinical implications, the present findings suggest selecting stimuli that maximize expectancy violation for exposure-based

treatments. This can be achieved by identifying individual threat beliefs. This is crucial as the exact stimuli of acquisition are highly unlikely to be reproduced in treatments. Our findings also suggest that presenting stimuli that maximize expectancy violation promotes the generalization of safety behaviors reduction to resembling stimuli. This is of clinical importance as it can aid in reducing "protection from extinction" to a wide range of fear-related stimuli, facilitating the generalization of extinction learning. In addition, enhancing generalization of GS extinction in safety behaviors can help alleviate impairments caused by costly safety behaviors that persists in the absence of realistic threat. It is of importance to note that despite generalization of GS extinction was enhanced in the Linear group, it did not effectively reduce US-avoidance to the CS+. It remained, however, unclear whether this pattern was due to GS extinction failing to reduce US-avoidance to the stimulus of fear acquisition (cf. Wong et al., 2020), or due to the limited amount of extinction trials.

The current study had some limitations. First, the Linear group showed significantly higher anxiety and stress levels than the Similarity group. However, this group difference was unlikely to have confounded the current findings given that the Linear group still showed less US-avoidance averaged across stimuli after *GS extinction* compared to the Similarity group. Second, we assessed the rules after the experimental task. This meant that we were not able to measure any changes in rules after *GS extinction* (e.g., changing from a linear rule to a similarity rule after GS extinction). Although the post-experimental questionnaire emphasized that participants should respond with the rule adopted before *GS extinction*, it is still possible that participants' responses could have been affected by the *GS extinction* phase. Third, we have limited power for detecting how risk factors affect the generalization of US-avoidance and the generalization of GS extinction, given our power analysis was calculated based on the rule group differences after *GS extinction*.

In conclusion, the current work replicated findings on self-inferred rules determining the shape of US-avoidance generalization gradient (Wong and Pittig, 2022b), resembling how threat beliefs determine the extent of US-avoidance engagement to a broad range of stimuli that resemble the feared stimulus. Different generalization rules determine threat expectancy to the same GS, thus presenting this stimulus in extinction evokes different levels of expectancy violation in different rule groups, leading to different levels of generalization of GS extinction (Wong et al., 2020). The current study extends this finding to US-avoidance. Participants who had their threat expectancy strongly violated showed a great generalization of US-avoidance reduction to similar stimuli, whereas those who had experienced minimal expectancy violation showed very limited generalization of US-avoidance reduction. It is worth noting the strengths of the current study included the clear rule group differences in US-avoidance generalization gradients, leading to the identification of different levels of expectancy violation for a range of stimuli. Additionally, the augmented Gaussian modelling provided more details in where the group differences lie in. The current work emphasizes the identification of different threat beliefs and selection of stimuli that maximize expectancy violation to reduce US-avoidance, minimizing protection from extinction to a broad range of stimuli.

Authors' declaration

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We further confirm that any aspect of the work covered in this manuscript that has involved either experimental animals or human patients has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from h.k.wong@essb.eur.nl.

This declaration is confirmed by all authors as listed below:

Alex H. K. Wong Jessica C. Lee Paula Engelke Andre Pittig

CRediT authorship contribution statement

Alex.H.K. Wong: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Software, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Jessica C. Lee:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Paula Engelke:** Conceptualization, Methodology, Writing – review & editing. **Andre Pittig:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

None.

Data availability

Data for this study available at https://osf.io/b2gzv/

Acknowledgements

AHKW was supported in part by funds of the Bavarian State Ministry of Science and the Arts and the University of Würzburg to the Graduate School of Life Sciences (GSLS), University of Würzburg. JCL was supported by a Discovery Early Career Researcher Award from the Australian Research Council (DE210100292). PE was supported by a scholarship by the Studienstiftung des Deutschen Volkes (German National Academic Foundation). The authors would like to thank Anica Pilz, Negin Rahmani, Sevval Eryüksel, Maria Yepez, and Elena Duscher for their help in data collection and processing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2022.104233.

References

Ahmed, O., & Lovibond, P. F. (2019). Rule-based processes in generalisation and peak shift in human fear conditioning. *Quarterly Journal of Experimental Psychology, 72*(2), 118–131. https://doi.org/10.1177/1747021818766461

Alvarez, R. P., Johnson, L., & Grillon, C. (2007). Contextual-specificity of short-delay extinction in humans: Renewal of fear-potentiated startle in a virtual environment. *Learning & Memory, 14*(4), 247–253. https://doi.org/10.1101/lm.493707

Barry, T. J., Griffith, J. W., Vervliet, B., & Hermans, D. (2016). The role of stimulus specificity and attention in the generalization of extinction. *Journal of Experimental Psychopathology, 7*(1), 143–152. https://doi.org/10.5127/jep.048615

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.

Bauer, E. A., MacNamara, A., Sandre, A., Lonsdorf, T. B., Weinberg, A., Morriss, J., et al. (2020). Intolerance of uncertainty and threat generalization: A replication and extension. *Psychophysiology, 57*(5), Article e13546. https://doi.org/10.1111/psyp.13546

Boucsein, W., Fowles, D. C., Grimnes, S., Ben-Shakhar, G., Roth, G., Dawson, W. T., & 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.

Boyle, S., Roche, B., Dymond, S., & Hermans, D. (2016). Generalisation of fear and avoidance along a semantic continuum. *Cognition & Emotion, 30*(2), 340–352. https://doi.org/10.1080/02699931.2014.1000831

Carleton, R. N., Norton, M. A. P. J., & Asmundson, G. J. G. (2007). Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders, 21*(1), 105–117. https://doi.org/10.1016/j.janxdis.2006.03.014.

Carpenter, J. K., Andrews, L. A., Witcraft, S. M., Powers, M. B., Smits, J. A., & Hofmann, S. G. (2018). Cognitive behavioural therapy for anxiety and related disorders: A meta- analysis of randomized placebo-controlled trials. *Depression and Anxiety, 35*(6), 502–514. https://doi.org/10.1002/da.22728

Carpenter, J. K., Pinaire, M., & Hofmann, S. G. (2019). From extinction learning to anxiety treatment: Mind the gap. *Brain Sciences, 97*(7), 164. https://doi.org/10.3390/brainsci9070164

Craske, M. G., Hermans, D., & Vervliet, B. (2018). State-of-the-art and future directions for extinction as a translational model for fear and anxiety. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences* (Vol. 373), Article 20170025. https://doi.org/10.1098/rstb.2017.0025, 1742, pii.

Craske, M. G., Treanor, M., Conway, 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

Dymond, S., Schlund, M. W., Roche, B., & Whelan, R. (2014). The spread of fear: Symbolic generalization mediates graded threat-avoidance in specific phobia. *Quarterly Journal of Experimental Psychology, 67*(2), 247–259. https://doi.org/10.1080/17470218.2013.800124

Dymond, S., Schlund, M. W., Roche, B., Whelan, R., Richards, J., & Davies, C. (2011). Inferred threat and safety: Symbolic generalization of human avoidance learning. *Behaviour Research and Therapy, 49*(10), 614–621. https://doi.org/10.1016/j.brat.2011.06.007

Flores, A., López, F. J., Vervliet, B., & Cobos, P. L. (2018). Intolerance of uncertainty as a vulnerability factor for excessive and inflexible avoidance behavior. *Behaviour Research and Therapy, 104*, 34–43. https://doi.org/10.1016/j.brat.2018.02.008.

Freeston, M. H., Rhéaume, J., Letarte, H., Dugas, M. J., & Ladouceur, R. (1994). Why do people worry? *Personality and Individual Differences, 17*(6), 791–802. https://doi.org/10.1016/0191-8869(94)90048-5.

Gerlach, A. L., Andor, T., & Patzelt, J. (2008). Die Bedeutung von Unsicherheitsintoleranz für die Generalisierte Angststörung Modellüberlegungen und Entwicklung einer deutschen Version der Unsicherheitsintoleranz-Skala. *Zeitschrift für Klinische Psychologie und Psychotherapie, 37*, 190–199. https://doi.org/10.1026/1616-3443.37.3.190.

Gorka, A. X., LaBar, K. S., & Hariri, A. R. (2016). Variability in emotional responsiveness and coping style during active avoidance as a window onto psychological vulnerability to stress. *Physiology & Behavior, 158*, 90–99. https://doi.org/10.1016/j.physbeh.2016.02.036

Helbig-Lang, S., & Petermann, F. (2010). Tolerate or eliminate? A systematic review on the effects of safety behavior across anxiety-related disorders. *Clinical Psychology: Science and Practice, 17*(3), 218–233. https://doi.org/10.1111/j.1468-2850.2010.01213.x

Hermans, D., Dirikx, T., Vansteenwegenin, D., Baeyens, F., Van den Bergh, O., & Eelen, E. (2005). Reinstatement of fear responses in human aversive conditioning. *Behaviour Research and Therapy, 43*(4), 533–551. https://doi.org/10.1016/j.brat.2004.03.013

Hofmann, S. G., & Smits, J. A. (2008). Cognitive-behavioral therapy for adult anxiety disorders: A meta-analysis of randomized placebo-controlled trials. *Journal of Clinical Psychiatry, 69*(4), 621–632. https://doi.org/10.4088/jcp.v69n0415

Hunt, C., Cooper, S. E., Hartnell, M. P., & Lissek, S. (2019). Anxiety sensitivity and intolerance of uncertainty facilitate associations between generalized Pavlovian fear and maladaptive avoidance decisions. *Journal of Abnormal Psychology, 128*(4), 315–326. https://doi.org/10.1037/abn0000422

Kaldewaij, R., Hashemi, M. M., Zhang, W., Koch, S. B. J., Figner, B., Roelofs, K., & Klumpers, F. (2021). Individual differences in costly fearful avoidance and the relation to psychophysiology. *Behaviour Research and Therapy, 137*, Article 103788. https://doi.org/10.1016/j.brat.2020.103788.

Krypotos, A.-M., Vervliet, B., & Engelhard, I. M. (2018). The validity of human avoidance paradigms. *Behaviour Research and Therapy, 111*, 99–105. https://doi.org/10.1016/j.brat.2018.10.011

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods, 53*, 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

Laing, P. A. F., Felmingham, K. L., Davey, C. G., & Harrison, B. J. (2022). The neurobiology of Pavlovian safety learning: Towards an acquisition-expression framework. *Neuroscience & Behavioral Reviews, 142*, Article 104882. https://doi.org/10.1016/j.neubiorev.2022.104882

Lau, J. Y. F., Hilbert, K., Goodman, R., Gregory, A. M., Pine, D. S., Viding, E. M., et al. (2012). Investigating the genetic and environmental bases of biases in threat recognition and avoidance in children with anxiety problems. *Biology of Mood & Anxiety Disorders, 2*, 12. https://doi.org/10.1186/2045-5380-2-12

Lee, J. C., Hayes, B. K., & Lovibond, P. F. (2018). Peak shift and rules in human generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0000558

Lee, J. C., Lovibond, P. F., Hayes, B. K., & Lewandowsky, S. (2021). A mixture of experts in associative generalization. *Proceedings of the Annual Meeting of thee Cognitive Science Society, 43*(43).

Lee, J. C., Mills, L., Hayes, B. K., & Livesey, E. J. (2021). Modelling generalisation gradients as augmented Gaussian functions. *The Quarterly Journal of Experimental Psychology, 74*(1), 106–121. https://doi.org/10.1177/1747021820949470

Lovibond, P. F., Lee, J. C., & Hayes, B. K. (2020). Stimulus discriminability and induction as independent components of generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(6), 1106–1120.

Lovibond, P. F., Mitchell, C. J., Minard, E., Brady, A., & Menzies, R. G. (2009). Safety behaviours preserve threat beliefs: Protection from extinction of human fear conditioning by an avoidance response. *Behaviour Research and Therapy, 47*(8), 716–720. https://doi.org/10.1016/j.brat.2009.04.013

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd). Sydney Psychology Foundation.

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40), 1541. https://doi.org/10.21105/joss.01541.

Mertens, G., Boddez, Y., Krypotos, A.-M., & Engelhard, I. M. (2021). Human fear conditioning is moderated by stimulus contingency instructions. *Biological Psychology, 158*, Article 107994. https://doi.org/10.1016/j.biopsycho.2020.107994.

Milad, M. R., Orr, S. P., Pitman, R. K., & Rauch, S. L. (2005). Context modulation of memory for fear extinction in humans. *Psychophysiology, 42*, 456–464. https://doi.org/10.1111/j.1469-8986.2005.00302.x

Morriss, J., Zuj, D. V., & Mertens, G. (2021). The role of intolerance of uncertainty in classical threat conditioning: Recent developments and directions for future research. *International Journal of Psychophysiology, 166*, 116–126. https://doi.org/10.1016/j.ijpsycho.2021.05.011

Nilges, P., & Essau, C. (2015). Die depressions-angst-stress-skalen: Der DASS–einscreeningverfahren nicht nur fur schmerzpatienten [depression, anxiety and stress scales: DASS – a screening procedure not only for pain patients]. *Der Schmerz, 29*(6), 649–657. https://doi.org/10.1007/s00482-015-0019-z.

Orr, S. P., Metzger, L. J., Lasko, N. B., Macklin, M. L., Peri, T., & Pitman, R. K. (2000). De novo conditioning in trauma-exposed individuals with and without posttraumatic stress disorder. *Journal of Abnormal Psychology, 109*(2), 290–298. https://doi.org/10.1037/0021-843X.109.2.290

Pittig, A. (2019). Incentive-based extinction of safety behaviors: Positive outcomes competing with aversive outcomes trigger fear-opposite action to prevent protection from fear extinction. *Behaviour Research and Therapy, 121*, Article 103463. https://doi.org/10.1016/j.brat.2019.103463

Pittig, A., & Scherbaum, S. (2020). Costly avoidance in anxious individuals: Elevated threat avoidance in anxious individuals under high, but not low competing rewards. *Journal of Behavior Therapy and Experimental Psychiatry, 66*, Article 101524. https://doi.org/10.1016/j.jbtep.2019.101524

Pittig, A., Wong, A. H. K., Glück, V. M., & Boschet, J. M. (2020). Avoidance and its bi-directional relationship with conditioned fear: Mechanisms, moderators, and clinical implications. *Behaviour Research and Therapy, 126*, Article 103550. https://doi.org/10.1016/j.brat.2020.103550

Rattel, J. A., Miedl, S. F., Blechert, J., & Wilhelm, F. H. (2017). Higher threat avoidance costs reduce avoidance behaviour which in turn promotes fear extinction in humans. *Behaviour Research and Therapy, 96*, 37–46. https://doi.org/10.1016/j.brat.2016.12.010

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6*, 309–316. https://doi.org/10.1007/BF022 88586

Scheveneels, S., Boddez, Y., Vervliet, B., & Hermans, D. (2016). The validity of laboratory- based treatment research: Bridging the gap between fear extinction and exposure treatment. *Behaviour Research and Therapy, 86*, 87–94. https://doi.org/10.1016/j.brat.2016.08.015

Shiban, Y., Schelhorn, I., Pauli, P., & Muhlberger, A. (2015). Effect of combined multiple contexts and multiple stimuli exposure in spider phobia: A randomized clinical trial in virtual reality. *Behaviour Research and Therapy, 71*, 45–53. https://doi.org/10.1016/j.brat.2015.05.014

Stan Development Team. (2018). *RStan: The R interface to Stan*. http://mc-stan.org/.

Struyf, D., Hermans, D., & Vervliet, B. (2018). Maximizing the generalization of fear extinction: Exposures to a peak generalization stimulus. *Behaviour Research and Therapy, 111*, 1–8. https://doi.org/10.1016/j.brat.2018.09.005

Vervliet, B., & Geens, M. (2014). Fear generalization in humans: Impact of feature learning on conditioning and extinction. *Neurobiology of Learning and Memory, 113*, 143–148. https://doi.org/10.1016/j.nlm.2013.10.002

Vervliet, B., Vansteenwegen, D., Baeyens, F., Hermans, D., & Eelen, P. (2005). Return of fear in a human differential conditioning paradigm caused by a stimulus change after extinction. *Behaviour Research and Therapy, 43*(3), 357–371. https://doi.org/10.1016/j.brat.2004.02.005

Vervliet, B., Vansteenwegen, D., & Eelen, P. (2004). Generalization of extinguished skin conductance responding in human fear conditioning. *Learning & Memory, 11*, 555–558. http://www.learnmem.org/cgi/doi/10.1101/lm.77404.

Vervoort, E., Vervliet, B., Bennett, M., & Baeyens, F. (2014). Generalization of human fear acquisition and extinction within a novel arbitrary stimulus category. *PLoS One, 9*(5), Article e96569. https://doi.org/10.1371/journal.pone.0096569

Volders, S., Meulders, A., de Peuter, S., Vervliet, B., & Vlaeyen, J. W. S. (2012). Safety behavior can hamper the extinction of fear of movement-related pain: An experimental investigation in healthy participants. *Behaviour Research and Therapy, 50*(11), 735–746. https://doi.org/10.1016/j.brat.2012.06.004

Waters, A. M., Kershaw, R., & Lipp, O. V. (2018). Multiple fear-related stimuli enhance physiological arousal during extinction and reduce physiological arousal to novel stimuli and the threat conditioned stimulus. *Behaviour Research and Therapy, 106*, 28–36. https://doi.org/10.1016/j.brat.2018.04.005

Wells, A., Clark, D. M., Salkovskis, P., Ludgate, J., Hackmann, A., & Gelder, M. (1995). Social phobia: The role of in-situation safety behaviors in maintaining anxiety and negative beliefs. *Behavior Therapy, 26*(1), 153–161. https://doi.org/10.1016/S0005-7894(05)80088-7

Wong, A, H. K., & Pittig, A (2022a). A dimensional measure of safety behavior: A non-dichotomous assessment of costly avoidance in human fear conditioning. *Psychological Research, 86*, 312–330. https://doi.org/10.1007/s00426-021-01490-w

Wong, A. H. K., Glück, V. M., Boschet, J. M., & Engelke, P. (2020). Generalization of extinction with a generalization stimulus is determined by learnt threat beliefs. *Behaviour Research and Therapy, 135*, Article 103755. https://doi.org/10.1016/j.brat.2020.103755

Wong, A. H. K., & Lovibond, P. F. (2017). Rule-based generalization in single-cue and differential fear conditioning in humans. *Biological Psychology, 129*, 111–120. https://doi.org/10.1016/j.biopsycho.2017.08.056

Wong, A. H. K., & Lovibond, P. F. (2018). Excessive generalisation of conditioned fear in trait anxious individuals under ambiguity. *Behaviour Research and Therapy, 107*, 53–63. https://doi.org/10.1016/j.brat.2018.05.012

Wong, A. H. K., & Lovibond, P. F. (2020). Generalization of extinction of a generalization stimulus in fear learning. *Behaviour Research and Therapy, 125*. https://doi.org/10.1016/j.brat.2019.10353. Article 103535.

Wong, A. H. K., & Pittig, A. (2022b). Threat belief determines the degree of costly safety behavior: Assessing rule-based generalization of safety behavior with a dimensional measure of avoidance. *Behaviour Research and Therapy, 156*, Article 104158. https://doi.org/10.1016/j.brat.2022.104158

Zbozinek, T. D., & Craske, M. G. (2018). Pavlovian extinction of fear with the original conditional stimulus, a generalization stimulus, or multiple generalization stimuli. *Behaviour Research and Therapy, 107*, 64–75. https://doi.org/10.1016/j.brat.2018.05.009

Zbozinek, T. D., Holmes, E. A., & Craske, M. G. (2015). The effect of positive mood induction on reducing reinstatement fear: Relevance for long term outcomes of exposure therapy. *Behaviour Research and Therapy, 71*, 65–75. https://doi.org/10.1016/j.brat.2015.05.016