Test-Retest Reliability of Human Threat Conditioning and Generalization

Samuel E. Cooper[1], Joseph E. Dunsmoor[1,2], Kathleen A. Koval[3], Emma R. Pino[3], & Shari A. Steinman[3]

[1]Department of Psychiatry and Behavioral Sciences, University of Texas at Austin

[2]Institute for Neuroscience, University of Texas at Austin

[3]Department of Psychology, West Virginia University

Contacts: samuel.cooper@austin.utexas.edu, shari.steinman@mail.wvu.edu

Abstract

Given the increasing use of threat conditioning and generalization for clinical-translational research efforts, establishing test-retest reliability of these paradigms is necessary. Specifically, it is an empirical question whether the same participant evinces a similar generalization gradient of conditioned responses across two sessions with the identical contingencies and stimuli. Here, 51 human volunteers participated in an identical auditory threat acquisition and generalization protocol at two sessions separated by 9 days. Skin conductance responses (SCR) and trial-by-trial shock risk ratings served as primary dependent measures. We used linear mixed effects modeling to test differential threat responses and generalization gradients, and Generalizability (G) theory coefficients to formally assess test-retest reliability of intra-individual stability and change across time. Results showed largely invariant differential conditioning and generalization gradients across time. G coefficients indicated fair reliability for the majority of acquisition and generalization measures. Our findings support reliability of the threat conditioning and generalization paradigm and highlight their utility for assessments of behavioral interventions in mental health research.

For well over a century, Pavlovian conditioning paradigms have served as one of the most popular, reliable, and validated experimental tools for investigating learning and memory processes across species (Vervliet & Boddez, 2020). Pavlovian conditioning paradigms are increasingly popular for mental health research applications, as conditioning-based models provide a theoretical foundation for the etiology and treatment of a number of psychopathologies, such as anxiety disorders, obsessive-compulsive disorder, and posttraumatic stress disorder (Cooper & Dunsmoor, 2021; Dunsmoor et al., 2022; Pittig et al., 2018). Conditioning paradigms also provide objective measures for assessing the efficacy and potential mechanisms of therapeutic interventions, such as exposure therapy (Ball et al., 2017; Forcadell et al., 2017; Raeder et al., 2020).

In the standard human threat (fear) conditioning design, participants learn that a conditioned stimulus (CS; e.g., a picture or a tone) predicts an aversive unconditioned stimulus (US; e.g., a shock or a loud noise). Through the acquisition of the CS-US association, the CS alone can elicit increases in autonomic arousal (e.g., skin conductance response), subjective expectancy of the US, and changes in affective judgments of valence and arousal toward the CS. These conditioned responses (CR) tend to generalize to other stimuli that are perceptually and/or conceptually related to the CS, but have not been directly paired with the US (Dymond et al., 2015). The threat (fear) generalization paradigm is increasingly popular for clinical translational research efforts, as the overgeneralization of defensive responses toward stimuli that resemble known threats is a possible transdiagnostic marker that cuts across anxiety-related disorder categories (Cooper, Dis, et al., 2022; Dunsmoor & Paz, 2015; Lissek, 2012). A key assumption underlying these paradigms is that conditioning-related laboratory indices reflect traits that are stable within individuals across time. Consequently, Pavlovian conditioning paradigms should

provide test-retest reliability. Given the steady use of Pavlovian conditioning and generalization paradigms in basic and translational sciences and continued work to align these paradigms with clinical practices (e.g., Adolph et al., 2022), efforts to confirm their test-retest reliability are needed.

Whether consistent patterns of responses to learned and generalized threats can be reproduced within the same individual across time is not a simple matter and requires multiple investigations with different parameters to approach a consensus. As conditioning is a learning paradigm, there could be substantial differences at a follow-up test merely because participants learned the CS-US association at the initial test. For instance, human conditioning protocols commonly incorporate a discriminative design that includes an acquisition phase that uses a CS that predicts the US (i.e., CS+) and a CS that is never paired with the US (i.e., CS-). Therefore, when participants complete an acquisition phase a second time, they will presumably find it easier to discriminate between the CS+ and CS- due to their prior learning of the CS-US association and also due to familiarity with task procedures (commonly known as "practice effects" in other areas of psychology, e.g., Bird et al., 2003). This issue of prior learning is perhaps even more consequential for conditioned generalization paradigms, as the generalization test typically involves a number of ambiguous generalization stimuli (GS) that are never paired with the US, but might nonetheless elicit CRs in proportion to their similarity to the CS+. Thus, upon a *follow-up* generalization test, participants might remember from their initial experience that no GS was paired with the US. This memory of the previous session could systematically lower within-subject stability of generalization across time due to near non-responding on follow-up tests.

Individual differences in arousal measures, including skin conductance responses (SCR) and fear-potentiated startle (FPS) are well documented in human conditioning literature, with considerable variability across subjects that could potentially translate to across time variability (Lonsdorf & Merz, 2017). Some arousal variability might be explained by individual variability in psychological traits that broadly affect conditioning indices, such as intolerance of uncertainty (e.g., Hunt et al., 2019; Mertens & Morriss, 2021) or trait-anxiety (e.g., Barrett & Armony, 2009). However, intraindividual changes in psychophysiological arousal could fluctuate across sessions for a number of other reasons. For example, the first test session could generate relatively higher arousal because the participant is nervous to participate in a study with electrical shocks; but by the next session arousal has decreased because they are acquainted with the procedure and aware that the shock isn't as painful as they feared. Another potential influence on test-retest reliability of conditioning paradigms is that arousal during a given experimental session is likely impacted by state variables (e.g., emotional state, sleep) with no guarantee that arousal levels will be consistent in the same individual across testing sessions.

Empirical research quantifying the stability of individual differences in CRs in humans across time are limited, but so far provide evidence for test-retest reliability within the same individuals. For instance, Zeidan et al. (2012) found strong test-retest reliability for acquisition and extinction-recall SCRs across three identical test sessions separate by up to 3 months each. Torrents-Rodas et al. (2014) found strong test-retest reliability for acquisition and threat generalization in the same individuals across two test sessions separated by 8 months, including across measures of SCR, FPS, and shock expectancy. These studies provide promising evidence of test-retest reliability across conditioning indices. However, work in this area is scarce and limited to relatively long intervals between tests (i.e., several months). Over long testing

intervals, subjects may forget specific stimulus attributes of the CS and GSs, as well as crucial elements of the experimental protocol (Jasnow et al., 2012; Riccio & Joynes, 2007).

Notably absent from the literature of reliability of conditioning paradigms are investigations of shorter interval test-retest reliability, particularly studies of temporal stability on the scale of weeks as opposed to months. These relatively shorter test-retest intervals are important, as many exposure therapy studies assess changes in symptoms and related psychological variables (e.g., treatment mediators) every week or biweekly (e.g., Kothgassner et al., 2019; Mataix-Cols et al., 2017). Further, some studies administer self-report measures designed explicitly to test components of conditioning models (e.g., expectation violation; (Elsner et al., 2022), but do not assess objective *in vivo* measures of conditioned responding (e.g., psychophysiology). If intervention scientists seek to directly test temporal dynamics of candidate conditioning-related mechanisms of change during a treatment study, the reliability of conditioning tasks on the scale of weeks must be established.

The goal of this report is to investigate the test-retest reliability of threat acquisition and generalization and contribute to an important literature on the psychometric properties of these commonly used tasks. For instance, a strong base of test-retest evidence, comprised of multiple studies, is necessary to support conditioning tasks as reliable probes into the neurobehavioral mechanisms of anxiety-related psychology. In the current study, we employed an auditory threat generalization paradigm that was retested after a 9-day period in a sample recruited for elevated intolerance of uncertainty. We predicted that SCR and expectancy generalizes in a graded fashion, such that responses gradually diminish in magnitude as auditory stimuli decreasingly resemble the CS+ along a frequency dimension (Dunsmoor, Kroes, et al., 2017; Dunsmoor, Otto, et al., 2017). Based on prior work involving test-retest of conditioning indices, we predicted that

these generalization gradients would show overall adequate test-retest reliability across a 9-day period.

Method

Participants

Participants were recruited through West Virginia University's psychology department participant pool and through flyers posted in the psychology department. Interested individuals completed the Intolerance of Uncertainty Scale (IU; 27-item) online, and those with elevated IU (IU ≥ 72.22; one SD about the mean in a previous student sample; Buhr & Dugas, 2002) were invited to participate. A total of 72 participants provided consent. Of these 72, we excluded 18 participants who only completed the threat generalization task at the first session, and three participants were excluded due to unusable SCR data (technical issues or all zero values, which would result in artificially perfect test-retest reliability and were therefore inappropriate for our analyses), leaving N = 51 for the analyses described in the current effort. Participants all completed a hearing test at the conclusion of the study to ensure the ability to perceptually discriminate between the tone frequencies used in the experiment. No participants were excluded based on the results of a hearing test. See Table 1 for sample characteristics.

Table 1. Sample characteristics (N=51)

Age	
Mean (SD)	20.0 (2.88)
Gender	
Female	39 (76.5%)
Male	12 (23.5%)
Race	
Black/African origin	1 (2.0%)
East Asian	3 (5.9%)
Other or Unknown	6 (11.8%)
White/European origin	40 (78.4%)
American Indian/Alaska Native	1 (2.0%)
Ethnicity	
Hispanic or Latino	8 (15.7%)
Not Hispanic or Latino	43 (84.3%)
Education	
Associate's Degree	1 (2.0%)
Bachelor's Degree	3 (5.9%)
High School Graduate	5 (9.8%)
Some College	42 (82.4%)

Threat Generalization Task

The experimental task contained two phases, acquisition (discriminative threat conditioning) and generalization based on Dunsmoor, Kroes, et al. (2017, see Figure 1). Stimuli consisted of pure tone sine waves presented at a moderate volume (< 60 decibels) through two dedicated Dell Computers external speakers for 2.5 s each and separated by a 7–8 s inter-trial interval. Stimulus presentation was controlled using E-Prime 2.0 (Psychology Software Tools, Sharpburg, PA). CSs were a 1000 Hz and 550 Hz tone that signaled the presence (CS+) or absence (CS-) of the US, respectively. The acquisition phase included 12 presentations each of unpaired CS+ and CS-, and an additional 8 CS+ trials paired with the US (8 of 20 CS+ trials; 40% reinforcement rate). We excluded all CS+ trials paired with the US from analysis to mitigate potential confounds introduced by the US given the relatively short duration CS.

After fear-conditioning, participants received 6 novel GS tones of ranging between the CS− and CS+ (650, 800, and 900) and extending beyond the CS+ (1100, 1200, and 1350) During the generalization test, each tone (including unpaired CS+ and CS-) were presented 7 times each, for a total of 42 trials. We also included an additional 5 CS+ trials paired with the US during generalization to prevent extinction and habituation over the course of the lengthy generalization test (steady-state generalization testing; see also Blough, 1975; Dunsmoor et al., 2009; Lissek et al., 2008)

In all phases, we collected SCRs and trial-by-trial shock expectancy. These ratings consisted of a three alternative-forced-choice scale corresponding to ‘no risk,’ ‘moderate risk,’ and ‘high risk’ for receiving the US, based on prior generalization studies (e.g., Lissek et al, 2008). We informed participants that their button presses did not affect the outcome on a trial to mitigate the potential for participants to attribute the outcome to their choice or reaction times

(i.e. to prevent an illusory correlation). We instructed participants to try to learn the association between the tones and the shock, but no explicit information was given regarding the CS-US contingencies. Presentation was pseudo-randomized so that no more than 3 presentations of the same tone occurred in a row. After generalization testing, participants underwent a hearing test, which validated that all participants had normal hearing and the capacity to discriminate between each tone frequency used in the experiment.

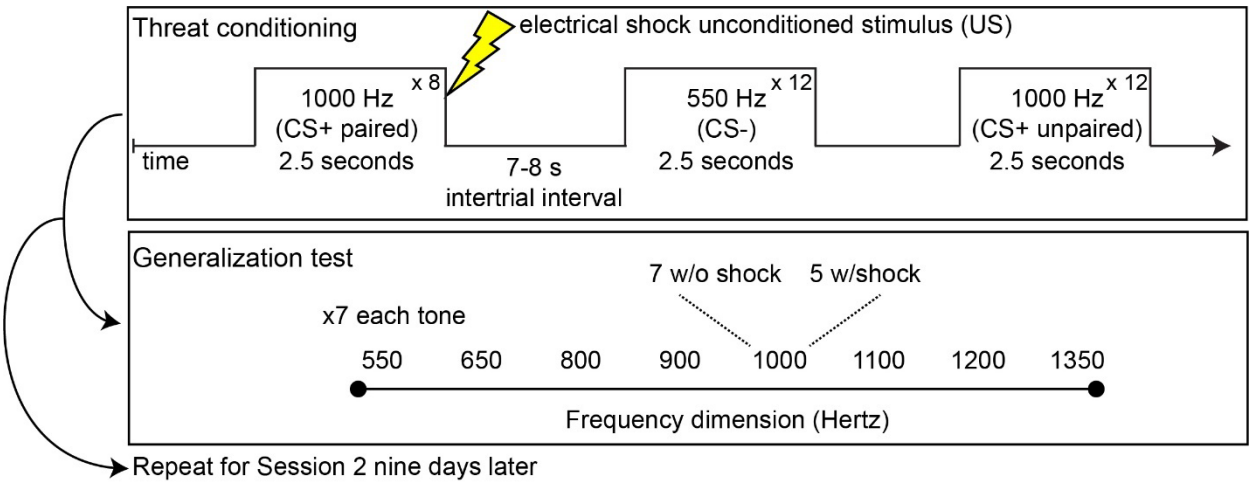

Figure 1. **Threat conditioning design.** Discriminative threat conditioning included pure tone conditioned stimuli paired (CS+, 1000 Hz) or unpaired (CS−, 550 Hz) with an aversive US. Generalization stimuli (GSs) were novel tones spanning a frequency continuum between the CS- and CS+ and beyond the CS+. CS- = conditioned safety cue; CS+ = conditioned threat cue; US = unconditioned stimulus.

Psychophysiology Collection and Shock Delivery

SCRs were acquired from the hypothenar eminence of the left palmar surface using disposable pre-gelled snap electrodes connected to the MP-100 BIOPAC System (BIOPAC Systems). Analysis of SCRs used previously described procedures (Dunsmoor et al., 2015; Dunsmoor, Kroes, et al., 2017). In brief, an SCR was considered related to CS presentation if the trough-to-peak deflection occurred 0.5–3 seconds following CS onset, lasted between 0.5 and 5.0 s, and was greater than 0.02 microsiemens (μS). Responses that did not fit these criteria were scored as zero. SCR values were obtained using a custom MATLAB (The Mathworks Inc., Natick, MA) script that extracts SCRs for each trial using the above criteria (Green et al., 2014) and subsequently inspected by an independent blinded rater. CS+ trials paired with the US were excluded from all analyses. Raw SCR scores were square root transformed prior to statistical analysis to normalize the distribution (Lykken & Venables, 1971).

Two electrodes were attached to the participants' right wrist to deliver shocks, which functioned as the US in this study. Shocks were generated by the BIOPAC STIMISOC adapter and lasted 200 ms. Each participant completed a shock work-up to determine a shock level that was highly annoying but not painful. In this procedure, shocks were calibrated using an ascending staircase procedure starting with a low voltage setting near a perceptible threshold and continuing until the participant endorsed the shock that was at a four or five on a 10-point intensity scale.

Procedure

The data in this paper are from a larger 4-session study, assessing the effect of cognitive bias modification for interpretations (CBM-I) compared to a control condition (sham CBM-I, designed not to affect interpretations) on intolerance of uncertainty. The primary aims and

outcomes of this intervention are described elsewhere (Koval et al., 2022). Preliminary analyses determined the intervention did not influence any conditioning task variables. This manuscript describes data from the threat generalization task completed at Session 1 ("initial session") and Session 4 ("follow-up session"), which were approximately nine days apart. Trained researchers attached SCR and shock electrodes and were guided through the shock workup procedure. Participants then received task instructions and completed the threat generalization task. After the task, participants completed brief post-task questionnaires and a hearing test.

Analytic Plan

All code for the current analyses can be found on this project's OSF repository, https://osf.io/zqfkj/?view_only=b8fcfa394f774438aed27a9117ebaec4.

Linear Mixed Models

We used linear mixed models (i.e., multilevel models) to model and test generalization gradients (see Vanbrabant et al., 2015 for applicability of these models to generalization data), All models were fit with the *lme4* library for R (Barr et al., 2013; R Core Team, 2018). All models contained a random-intercept of participant and fixed effects of stimulus, session, and the stimulus x session interaction. The addition of a session and stimulus random-effect was tested for improved fit using Likelihood ratio tests (LRTs) comparing models with and without the term, per standard mixed-effects regression recommendations (e.g., Barr et al. 2013; Gelman & Hill, 2006). We report standardized betas, 95% CIs, and Wald *t*-tests using Satterwhite approximated degrees of freedom for all terms from primary models. We also used linear mixed models for manipulation checks of differential conditioning during acquisition and to determine if participants continued to differentiate between the CS+ and CS- during the generalization phase; these models contained a fixed effect of stimulus with only CS+ and CS- trials included.

Psychometric Framework and Calculations

As our primary measures of test-retest reliability, we calculated coefficients based on generalizability (G) theory. Briefly, G theory is a psychometric approach that decomposes an observed score into multiple sources of variance to produce coefficients that describe different types of reliability (G coefficients), which expands on the classical test theory concept of reliability that recognized only single sources of non-error variance (Brennan, 2001; Cronbach et al., 1972; Shrout & Lane, 2012). G theory is a particularly noteworthy development for factorial experimental tasks that use psychophysiological measures, as these types of designs contain multiple sources of variances due to their signal-to-noise properties, multiple experimental parameters, and other attributes that together make classical test theory a poor fit to assess their reliability. In the current effort, for SCRs and risk ratings in both phases, we first used the *psych* and *lme4* libraries for R to obtain variance components via the "mlr" and "lmer" functions (Bates et al., 2015; Revelle, 2017) and used functions from the *gtheory* library (Moore, 2016) to extract components. With these components, we calculated two G coefficients. The first of these we term R_{IRS} and was proposed by Hinz et al. (2002) as a metric of "individual response stability", the proportion of within-person responding to experimental stimuli that is stable across time and is best suited to capturing the stability of patterns of stimulus generalization. Equation 1 was used to calculate R_{IRS}:

1

$$R_{IRS} = \frac{\sigma^2_{Participant\ x\ Stimulus}}{\left(\sigma^2_{Participant\ x\ Stimulus} + \sigma^2_{Residual}\right)}$$

In Equation 1, $\sigma^2_{Participant\ x\ Stimulus}$ refers to individual variability in response to the experimental stimuli, and $\sigma^2_{Residual}$ refers to error variance that is not accounted for by other

components (i.e., variance that cannot be explained by the tested factors). Notably, R_{IRS} was the G coefficient reported in the only prior study of test-retest of threat generalization, Torrents-Rodas et al. (2014)[1], which also collected SCR and ratings. Thus, we have the opportunity to directly compare this form of reliability between two different studies. Larger R_{IRS} coefficients indicate that a pattern of individual responding is consistent across time and can increase confidence that conditioning tasks are capturing a relatively stable associative learning process.

In addition to R_{IRS}, we report R_C (Equation 2), which was first proposed by Cranford et al. (2006) and further discussed by Shrout and Lane (2012) as a measure of the reliability of *change* in responses across individuals between time points, as opposed to stability of a particular response pattern:

2

$$R_C = \frac{\sigma^2_{Participant\ x\ Session}}{\left(\sigma^2_{Participant\ x\ Session} + [\sigma^2_{Residual}/m]\right)}$$

In Equation 2, $\sigma^2_{Participant\ x\ Session}$ refers to individual variability at each session (i.e., across time). The $\sigma^2_{Residual}$ term continues to refer to error variance, but in this case, it is divided by *m* number of sessions, which results in a fixed effect coefficient (i.e., the estimate is specific to number of sessions specified, which is 2 in the current study). We report R_C due to the continued interest in and practice of using conditioning tasks as biobehavioral measures of underlying pathological mechanisms that are targets of intervention research, particularly exposure therapy research (Craske et al., 2014; Raeder et al., 2020). Larger R_C coefficients

[1] Torrents-Rodas et al. (2014) refer to the R_{IRS} coefficient with the more general notation for a G coefficient, $E\rho^2$. We use the notation from Hinz et al. (2002) to align with G theory work by Cranford et al. (2006), Shrout and Lane (2012), and others and to facilitate additional investigations using these coefficients.

would provide support for a measure being useful to track changes in responding related to an intervention, as opposed to change as a result of random error.

Rc coefficients complement R_{IRS} coefficients by quantifying a person's non-stable variance (i.e., the variance that is unreliable according to R_{IRS}) and determining how much of said variance is related to change across timepoints. Accordingly, it is possible to have both adequate R_C and R_{IRS} coefficients from the same measure, but as one increases, the available variance to quantify for the other measure decreases. It is therefore not possible to have very high R_C and R_{IRS} simultaneously, nor would be it desirable for a treatment measure because it would suggest that measure is no amenable to intervention-related change.

For both types of coefficients, we constructed 95% confidence intervals using the method provided in Table 7 in McGraw and Wong (1996). CIs that do not contain zero within its interval indicate that the coefficient is significantly different from zero. We also provided qualitative descriptions of coefficient size based on commonly applied recommendations (see Matheson, 2019), although we caution against stringent application of these standards for conditioning tasks given the limited work in this area and disagreement on firm guidelines regarding interpretation of within-person reliability (Matheson, 2019).

To supplement our G coefficients and to provide an additional point of reference of comparison to prior studies, we also calculate row-wise correlations that reflect the overall paired correlation between the two sessions across all stimuli. These correlations can be compared to the intraclass correlation coefficients (ICCs) reported in the broader test-retest literature (Fisher, 1992), but are a less precise measure for the current effort than the G coefficients.

Results

Differential Threat Conditioning

SCR

Successful differential conditioning, operationalized as significantly larger CS+ responses compared with CS- responses, was evident during acquisition at both timepoints (initial test: $t(2171) = 12.7$, $p < .001$; follow-up test: $t(1724) = 15.3$, $p < .001$). Participants continued to respond more strongly to the CS+ compared with the CS- during generalization at both timepoints (initial test: $t(1287) = 2.8$, $p = .005$; follow-up test*:* $t(1021) = 15.3$, $p < .001$).

Ratings

Successful differential conditioning, operationalized as significantly larger CS+ risk ratings compared with CS- ratings, was evident during acquisition at both timepoints (session 1: $t(2131) = 54.6$, $p < .001$; session 2: $t(1718) = 58.6$, $p < .001$). Participants continued to expect the US more for the CS+ compared with the CS- during generalization at both timepoints (session 1: $t(1268) = 16.3$, $p < .001$; session 2: $t(1015) = 22.6$, $p < .001$).

Generalization Gradients

SCR

A model with a random-effect of testing session was the best fit for SCR data, $\chi^2(2) = 286.49$, $p < .001$. In this model, both stimulus, $\beta = .08$, $t(7430) = 7.94$, $p < .001$, 95% CI [.06, .10], and session, $\beta = -.07$, $t(7430) = -2.28$, $p = .023$, 95% CI [-.13, -.01], predictors were significant, but the Stimulus x Session interaction was not significant $\beta = -.007$, $t(7429) = 1.86$, $p = .063$, 95% CI [0, .02]. When examining difference slopes at the individual GS level across sessions, only the CS+, $\beta = .13$, $t(7417) = 3.66$, $p < .001$, 95% CI [.06, .20], and $GS_{1100,}$ $\beta = .12$, $t(7417) = 3.04$, $p =. 002$, 95% CI [.064 .20], slopes significantly differed from the CS- slopes, with almost no change across sessions for CS+ and GS_{1100}. See Figure 2A for visualized

generalization gradients at each testing session. The CBM intervention did not significantly affect gradients, as assessed through a Stimulus x Session x CBM Group interaction, β = .0006, $t(7425) = -0.26$, $p = .798$, 95% CI [-.1, 0] and rerunning models with CBM Group included as a separate fixed-effect term, which yielded almost no change in the reported coefficients and did not change their significance.

Ratings

A model with a random-effect of testing session was also the best fit for risk rating data, $\chi^2(2) = 55.61$, $p < .001$. In this model, both stimulus, β = .14, $t(7378) = 13.02$, $p < .001$, 95% CI [.12, .16], and session, β = -.10, $t(7378) = -4.89$, $p < .001$, 95% CI [-.14, -.06], predictors were significant, but as with the SCR data, the Stimulus x Session interaction was not significant β = -.005, $t(7377) = -1.14$, $p = .253$, 95% CI [-.01, .0]. When examining difference slopes at the individual GS level across sessions, only the CS+, β = .17, $t(7365) = 4.44$, $p < .001$, 95% CI [.10, .25] slope significantly differed from the CS- slope, with almost no change across sessions for CS+. See Figure 2B for visualized generalization gradients at each testing session. The CBM intervention did not significantly affect gradients, as assessed through a Stimulus x Session x CBM Group interaction, β = .0006, $t(7373) = 0.23$, $p = .816$, 95% CI [-.01, .01] and rerunning models with CBM Group included as a separate fixed-effect term, which yielded almost no change in the reported coefficients and did not change their significance

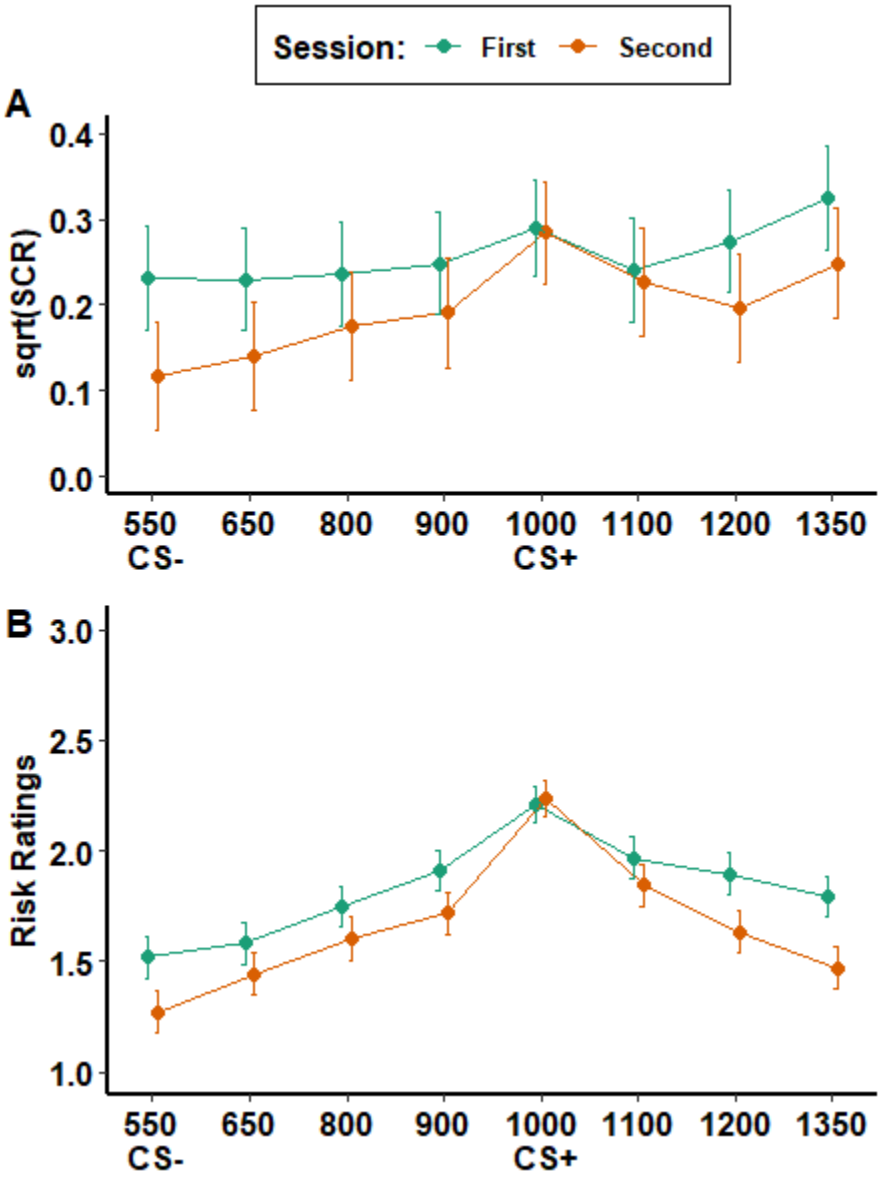

Figure 2. Conditioned generalization gradients at initial and follow-up session. All plotted values are fitted values from the linear mixed effects models described in text. Error bars represent 95% confidence intervals adjusted for random effects of the model. Panel A displays SCR generalization gradients; Panel B demonstrates risk rating gradients. CS- = conditioned safety cue; CS+ = conditioned threat cue; SCR = skin conductance response; US = unconditioned stimulus.

Test-Retest Reliability

Table 2 displays all variance components for each phase. Zero-order correlations between each stimulus at session 1 and session 2 can be found in Supplemental Material Figures S1-S3. Plots documenting individual-level change over time for each stimulus can be found in Supplemental Material Figures S4-S7.

Table 2. Variance component analysis results

	Acquisition				Generalization			
	SCR		Ratings		SCR		Ratings	
Component	Variance	%	Variance	%	Variance	%	Variance	%
Participant	0.02	0.14	0	0	0.02	0.37	0.05	0.19
Session	0	0.02	0	0	0	0.03	0.01	0.05
Stimulus	0.03	0.26	0.96	0.91	0	0.02	0.06	0.23
Participant x Session	0.02	0.18	0	0	0.02	0.25	0.02	0.08
Participant x Stimuli	0.03	0.21	0.04	0.04	0.01	0.13	0.05	0.16
Session x Stimuli	0	0.01	0	0	0	0.01	0.01	0.02
Residual	0.02	0.18	0.05	0.05	0.01	0.2	0.08	0.28

Notes: Each variance component was extracted from a linear mixed model constructed for each dependent variable in each phase. Here, we report both variance values and percentage of total variance for each component. ACQ = acquisition; GEN = generalization; SCR = skin conductance response.

Acquisition

Although the largest source of variance at acquisition was the stimulus component for both SCR and ratings, the magnitude of this component notably varied. For SCR, the stimulus component accounted for 26.2% of variance, with the Participant x Stimulus (21.2%) and Participant x Time (17.5%) interactions accounting for slightly smaller proportions of variance. These components indicate that a modest majority of the variance resulted from differences in the average responding to each stimulus (i.e., "main effect"), as would be expected during differential conditioning, but that responses also notably varied depending on the person and the testing session, as would be expected of a psychophysiological variable. Residual variance was also comparable to these components (17.5%), indicating a notable proportion of error variance in SCR measurements. In contrast, the largest variance component for risk ratings was also stimulus, but with this component accounting for 90.6% of variance, with the negligible remainder mostly accounted for by the Participant x Stimulus interaction (3.9%) and residual (5%) terms. Accordingly, variance in risk ratings was almost entirely accounted for by the difference in stimuli and was consistent across all participants.

Test-retest coefficients for this phase also differed depending on the measure (see Figure 3). The reliability of within-person patterns of responding was fair-to-good for both measures (SCR: R_{IRS} = .50, 95% CI [.26, .68]; risk ratings: R_{IRS} = .47, 95% CI [.23, .66]) but reliability of change across time notably differed by measure. For SCR, reliability of change was good (R_C = .67, 95% CI [.49, .8]), whereas for risk ratings it was poor (R_C = .14, 95% CI [-.14, .4]) and was the only generalizability coefficient calculated in the current effort that was not significant. The row-wise correlation between acquisition SCRs at each testing session was significant, r = .45, p

<.001, as was the correlation between acquisition risk ratings at each testing session, $r = .44$, $p =$ 001.

Generalization

The pattern of variance components for generalization markedly differed from those from acquisition (see Table 2). For SCR, the largest predictor variance component was the participant component (37.3%), followed by the Participant x Time (24.5%) and Participant x Stimulus (13%) interactions. This indicates that the majority of variation was across participants (i.e., differences in average physiological responding), but also dependent on the testing session and, to a lesser extent, each persons' pattern of responding to each stimulus. Residual variance also accounted for a notable proportion of variance (19.8%), indicating marked error variance in SCR at this phase. In contrast, the largest predictor variance component for risk ratings was the stimulus component (22.6%), followed by participant (18.9%) and the Participant x Stimulus interaction (16.3%). Of note is that for risk ratings, residual variance also accounted for the overall largest proportion of variance (28.4%), indicating a substantial amount of variance could not be explained by the predictors.

Test-retest was largely similar across both measures during generalization, and all coefficients were significant (see Figure 3). The reliability of within-person patterns of responding was fair for both measures (SCR: $R_{IRS} = .39$, 95% CI [.13, .60]; risk ratings: $R_{IRS} =$ .36, 95% CI [.10, .58]), although lower than reliability during acquisition. Reliability of change was good for both measures (SCR: $R_{IRS} = .91$, 95% CI [.85, .95]; risk ratings: $R_{IRS} = .68$, 95% CI [.50, .80]). The row-wise correlation between generalization SCRs at each testing session was significant, $r = .59$, $p <.001$, as was the correlation between generalization risk ratings at each testing session, $r = .64$, $p <.001$,

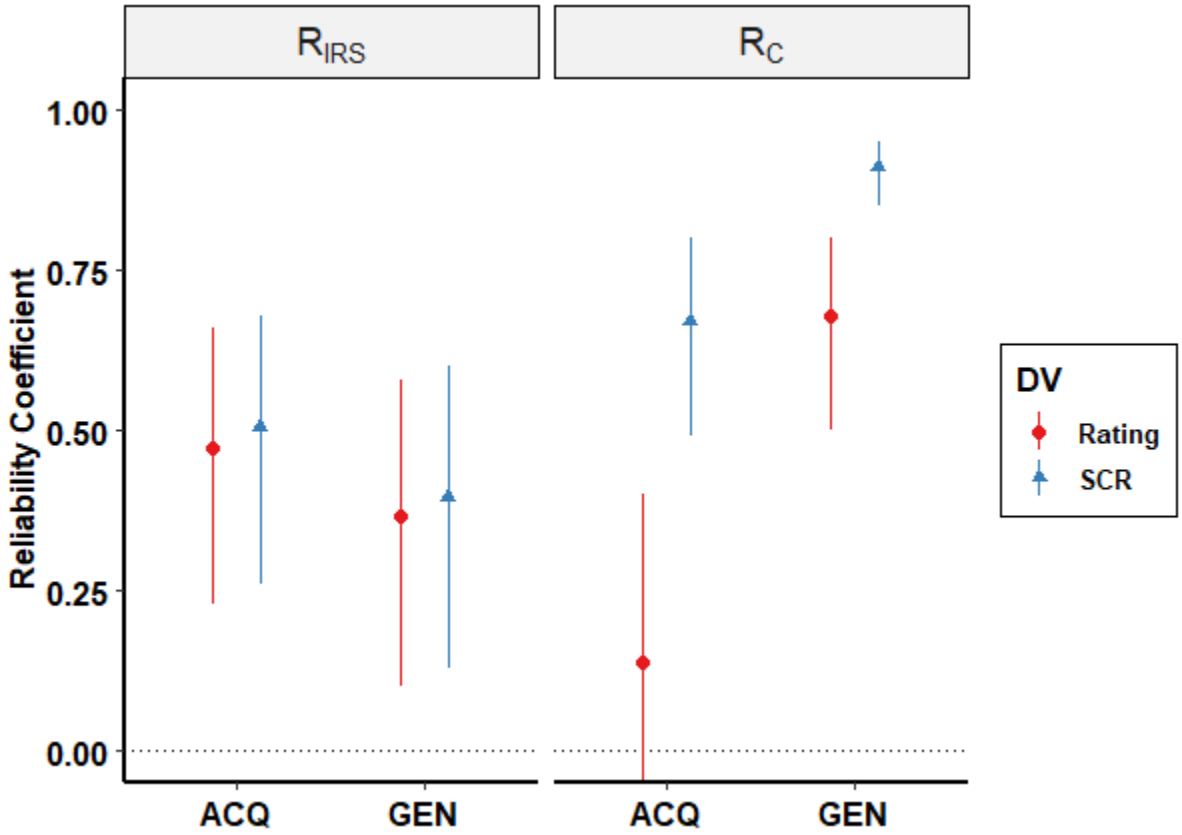

Figure 3. **Reliability of SCR and risk rating data across testing phases and sessions.** The left panel displays R_{IRS} (within-subject stability of response pattern) coefficients; the right panel displays R_C (within-subject reliability of change) coefficients. Error bars reflect 95% confidence intervals; CIs that do not overlap with zero indicate the coefficient is significantly different from zero. ACQ = acquisition phase; CS- = conditioned safety cue; CS+ = conditioned threat cue; GEN = generalization phase; SCR = skin conductance response; US = unconditioned stimulus.

Discussion

Given the prominent role of conditioning paradigms in preclinical and clinical-translational research, it is important to assess the reliability of these protocols over time in the same individuals. Here, we investigated the test-retest reliability of two behavioral measures during auditory fear acquisition and stimulus generalization tests, SCR and risk ratings, across a 9-day interval. In line with our hypotheses, test-retest reliability was generally adequate across two types of G coefficients, one indexing within-person stability of response patterns and the other indexing reliability of change across time. Further, generalization gradients did not significantly differ across sessions, although there was limited across-session variability at the level of individual stimuli. These results support the conclusion that threat generalization gradients remain relatively stable using an identical protocol at two time points. These findings might further encourage the use of these paradigms for pre-to-post measures on the efficacy of behavioral interventions aimed at reducing generalized fear and arousal (e.g., cognitive behavioral therapy or cognitive bias modification for anxiety disorders; Cristea et al., 2015; Steinman et al., 2021).

We measured two types of reliability in our analyses: stability of responses within individuals, and reliability of change across sessions. The first is most important for understanding how generalization profiles are stable over time, the second for clarifying if changes across time are systematic (e.g., related to between-session interval or intervention) or random error. In this study, we found that within-person stability of behavioral generalization was fair and relatively invariant across dependent measures of SCR and risk ratings. Interestingly, these coefficients were larger than those found in another reliability study of threat generalization (Torrent-Rodas et al. 2014) that retested generalization at an 8-month interval.

Specifically, test-retest stability for the 8-month interval ranged from R_{IRS} =.23 to R_{IRS} =.34, compared to R_{IRS} =.36 to R_{IRS} =.39 in the current study. Differences in reliability at different intervals is a key issue in determining the utility of repeated testing of conditioned generalization. One possibility is that generalization stability is improved over a short test-retest interval compared with longer intervals. In contrast, a longer interval between testing sessions could result in forgetting the details of stimulus attributes and the experimental procedure, and subsequently promote increased generalization (Jasnow et al., 2016; Riccio & Joynes, 2007). Also possible is that test-retest at an even shorter interval, such as 24 or 48 hours, would result in poorer reliability due to participants likely having strong explicit memory that would bias their responding relative to the initial, naïve testing session. Taken together, the current results suggest that a 9-day interval is adequate for obtaining stable generalization over time, and theory and limited prior results suggest that much shorter or longer intervals might pose some issues for reliability (as is seen in other memory-related tasks, for meta-analysis see Scharfen et al., 2018). However, further reliability studies at varying time-intervals are warranted to detail the optimal interval for pre-to-post testing of generalization protocols.

The current effort is the first to investigate the reliability of change across testing sessions in threat generalization tasks. We observed good reliability for generalization measures, which suggests testing of generalization across a short-term interval is viable for intervention studies seeking to measure treatment-related change over time. Contrary to the reliability observed for generalization, change in acquisition risk rating across testing sessions evidenced poor reliability. Interindividual differences in memory for the US contingency learned in the first session might explain poor change reliability here: some participants will perfectly remember the relatively simple CS/US association and provide invariant risk ratings for acquisition during the second

test. Others might have more variable ratings during this phase, either due to poorer retention or more elaborate reasons (e.g., expecting a change in contingency or stimuli). Regardless of reason for this pattern, a subgroup of participants with near invariant responding at one timepoint will negatively bias change reliability scores, as there is essentially no change to measure (Shrout & Lane, 2012).

One limitation of the current study is that we did not collect qualitative or quantitate data on participants' explicit memory for their prior testing session. Therefore, we could not account for whether performance at the second test was affected by participants' ability to recall explicit details of the task structure. Another limitation is that the sample was constrained to those with relatively higher IU scores. Although this might limit the generalizability of our findings, the levels of IU in the current study still likely reflect a sizable proportion of the population and we contend our results are still broadly applicable. Additionally, meta-analyses find that test-retest reliability does not differ by clinical status in several commonly used cognitive neuroscience and neuropsychology tasks (Calamia et al., 2013; Elliott et al., 2020). This suggests that those with markedly higher IU scores compared with those with lower scores would yield similar test-retest reliability on the threat generalization task. Lastly, participants in the current sample identified as primarily White and non-Hispanic or Latino, college-aged, and female. Replication with more demographically diverse samples is needed, particularly given evidence of demographic differences in fear conditioning metrics (e.g., Cooper, Hunt, et al., 2022; Rosenbaum et al., 2015).

Future work is needed to add to the growing evidence base of reliability studies of threat conditioning tests, particularly those testing generalization. For generalization studies, a study spanning multiple timepoints is the next step to determine differences in short and long-term

reliability. Further, the current study and prior work can only speak to generalization of passive-emotional Pavlovian learning. There has been substantial recent empirical attention on the overt behavioral consequences of threat generalization, most notably avoidance of threat (Pittig et al., 2020; Wong et al., 2022), which suggests that studies will be needed to clarify the reliability of generalized avoidance over time. More evidence is likely needed to form a strong conclusion on the utility of repeated generalization tests for intervention research. Thus, the next reliability studies of threat generalization would benefit from testing participants with diagnosed psychopathology across multiple timeframes and utilizing a design that resembles those from intervention studies, such as weekly testing sessions. However, the current study suggests that fear generalization paradigms are reliable over a short interval and can be appropriate for assessing behavioral intervention effects.

References

Adolph, D., Flasinski, T., Lippert, M. W., Pflug, V., Hamm, A., O., R., Jan, Margraf, J., & Schneider, S. (2022). Measuring Extinction Learning across the Lifespan – Adaptation of an optimized paradigm to closely match exposure treatment procedures. *Biological Psychology*, 108311. https://doi.org/10.1016/j.biopsycho.2022.108311

Ball, T. M., Knapp, S. E., Paulus, M. P., & Stein, M. B. (2017). Brain activation during fear extinction predicts exposure success. *Depression and Anxiety*, *34*(3), 257–266. https://doi.org/10.1002/da.22583

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Barrett, J., & Armony, J. L. (2009). Influence of trait anxiety on brain activity during the acquisition and extinction of aversive conditioning. *Psychological Medicine*, *39*(2), 255–265. https://doi.org/10.1017/S0033291708003516

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

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2003). Test-retest reliability, practice effects and reliable change indices for the recognition memory test. *The British Journal of Clinical Psychology*, *42*(Pt 4), 407–425. https://doi.org/10.1348/014466503322528946

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*(1), 3.

Brennan, R. L. (2001). *Generalizability theory* (pp. xx, 538). Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4757-3456-0

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

Calamia, M., Markon, K., & Tranel, D. (2013). The Robust Reliability of Neuropsychological Measures: Meta-Analyses of Test–Retest Correlations. *The Clinical Neuropsychologist*, *27*(7), 1077–1105. https://doi.org/10.1080/13854046.2013.809795

Cooper, S. E., Dis, E.-A. van, Hagenaars, M., Krypotos, A.-M., Nemeroff, C., Lissek, S., Engelhard, I., & Dunsmoor, J. E. (2022). *A Meta-Analysis of Conditioned Fear Generalization in Anxiety-Related Disorders*. PsyArXiv. https://doi.org/10.31234/osf.io/q6zfh

Cooper, S. E., & Dunsmoor, J. E. (2021). Fear conditioning and extinction in obsessive-compulsive disorder: A systematic review. *Neuroscience & Biobehavioral Reviews*, *129*, 75–94. https://doi.org/10.1016/j.neubiorev.2021.07.026

Cooper, S. E., Hunt, C., Ross, J. P., Hartnell, M. P., & Lissek, S. (2022). Heightened generalized conditioned fear and avoidance in women and underlying psychological processes. *Behaviour Research and Therapy*, *151*, 104051. https://doi.org/10.1016/j.brat.2022.104051

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A Procedure for Evaluating Sensitivity to Within-Person Change: Can Mood Measures in Diary Studies Detect Change Reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929. https://doi.org/10.1177/0146167206287721

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

Cristea, I. A., Kok, R. N., & Cuijpers, P. (2015). Efficacy of cognitive bias modification interventions in anxiety and depression: Meta-analysis. *The British Journal of Psychiatry*, *206*(1), 7–16. https://doi.org/10.1192/bjp.bp.114.146761

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.

Dunsmoor, J. E., Campese, V. D., Ceceli, A. O., LeDoux, J. E., & Phelps, E. A. (2015). Novelty-Facilitated Extinction: Providing a Novel Outcome in Place of an Expected Threat Diminishes Recovery of Defensive Responses. *Biological Psychiatry*, *78*(3), 203–209. https://doi.org/10.1016/j.biopsych.2014.12.008

Dunsmoor, J. E., Cisler, J. M., Fonzo, G. A., Creech, S. K., & Nemeroff, C. B. (2022). Laboratory models of post-traumatic stress disorder: The elusive bridge to translation. *Neuron*, 24.

Dunsmoor, J. E., Kroes, M. C. W., Braren, S. H., & Phelps, E. A. (2017). Threat intensity widens fear generalization gradients. *Behavioral Neuroscience*, *131*(2), 168–175. https://doi.org/10.1037/bne0000186

Dunsmoor, J. E., Mitroff, S. R., & LaBar, K. S. (2009). Generalization of conditioned fear along a dimension of increasing fear intensity. *Learning & Memory*, *16*(7), 460–469. https://doi.org/10.1101/lm.1431609

Dunsmoor, J. E., Otto, A. R., & Phelps, E. A. (2017). Stress promotes generalization of older but not recent threat memories. *Proceedings of the National Academy of Sciences*, 201704428. https://doi.org/10.1073/pnas.1704428114

Dunsmoor, J. E., & Paz, R. (2015). Fear Generalization and Anxiety: Behavioral and Neural Mechanisms. *Biological Psychiatry*, *78*(5), 336–343. https://doi.org/10.1016/j.biopsych.2015.04.010

Dymond, S., Dunsmoor, J. E., Vervliet, B., Roche, B., & Hermans, D. (2015). Fear Generalization in Humans: Systematic Review and Implications for Anxiety Disorder Research. *Behavior Therapy*, *46*(5), 561–582. https://doi.org/10.1016/j.beth.2014.10.001

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 0956797620916786. https://doi.org/10.1177/0956797620916786

Elsner, B., Reuter, B., Said, M., Linnman, C., Kathmann, N., & Beucke, J.-C. (2022). Impaired differential learning of fear versus safety signs in obsessive-compulsive disorder. *Psychophysiology*, *59*(2), e13956. https://doi.org/10.1111/psyp.13956

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics* (pp. 66–70). Springer.

Forcadell, E., Torrents-Rodas, D., Vervliet, B., Leiva, D., Tortella-Feliu, M., & Fullana, M. A. (2017). Does fear extinction in the laboratory predict outcomes of exposure therapy? A treatment analog study. *International Journal of Psychophysiology*, *121*, 63–71. https://doi.org/10.1016/j.ijpsycho.2017.09.001

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology*, *91*(3), 186–193. https://doi.org/10.1016/j.ijpsycho.2013.10.015

Hinz, A., Hueber, B., Schreinicke, G., & Seibt, R. (2002). Temporal stability of psychophysiological response patterns: Concepts and statistical tools. *International Journal of Psychophysiology*, *44*(1), 57–65. https://doi.org/10.1016/S0167-8760(01)00191-X

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

Jasnow, A. M., Cullen, P. K., & Riccio, D. C. (2012). Remembering Another Aspect of Forgetting. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00175

Jasnow, A. M., Lynch, J. F., Gilman, T. L., & Riccio, D. C. (2016). Perspectives on fear generalization and its implications for emotional disorders. *Journal of Neuroscience Research*, n/a-n/a. https://doi.org/10.1002/jnr.23837

Kothgassner, O. D., Goreis, A., Kafka, J. X., Van Eickels, R. L., Plener, P. L., & Felnhofer, A. (2019). Virtual reality exposure therapy for posttraumatic stress disorder (PTSD): A meta-analysis. *European Journal of Psychotraumatology*, *10*(1), 1654782. https://doi.org/10.1080/20008198.2019.1654782

Koval, K. A., Dunsmoor, J. E., Pino, E. R., Edwards, C., & Steinman, S. A. (2022). Cognitive bias modification for intolerance of uncertainty. Manuscript in progress.

Lissek, S. (2012). Toward an Account of Clinical Anxiety Predicated on Basic, Neurally Mapped Mechanisms of Pavlovian Fear-Learning: The Case for Conditioned Overgeneralization. *Depression and Anxiety*, *29*(4), 257–263. https://doi.org/10.1002/da.21922

Lissek, S., Biggs, A. L., Rabin, S. J., Cornwell, B. R., Alvarez, R. P., Pine, D. S., & Grillon, C. (2008). Generalization of conditioned fear-potentiated startle in humans: Experimental validation and clinical relevance. *Behaviour Research and Therapy*, *46*(5), 678–687. https://doi.org/10.1016/j.brat.2008.02.005

Lonsdorf, T. B., & Merz, C. J. (2017). More than just noise: Inter-individual differences in fear acquisition, extinction and return of fear in humans - Biological, experiential, temperamental factors, and methodological pitfalls. *Neuroscience & Biobehavioral Reviews*, *80*, 703–728. https://doi.org/10.1016/j.neubiorev.2017.07.007

Lykken, D. T., & Venables, P. H. (1971). Direct Measurement of Skin Conductance: A Proposal for Standardization. *Psychophysiology*, *8*(5), 656–672. https://doi.org/10.1111/j.1469-8986.1971.tb00501.x

Mataix-Cols, D., Fernández de la Cruz, L., Monzani, B., Rosenfield, D., Andersson, E., Pérez-Vigil, A., Frumento, P., de Kleine, R. A., Difede, J., Dunlop, B. W., Farrell, L. J., Geller, D., Gerardi, M., Guastella, A. J., Hofmann, S. G., Hendriks, G.-J., Kushner, M. G., Lee, F. S., Lenze, E. J., … Thuras, P. (2017). D-Cycloserine Augmentation of Exposure-Based Cognitive Behavior Therapy for Anxiety, Obsessive-Compulsive, and Posttraumatic Stress Disorders: A Systematic Review and Meta-analysis of Individual Participant Data. *JAMA Psychiatry*, *74*(5), 501. https://doi.org/10.1001/jamapsychiatry.2016.3955

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*, *7*, e6918. https://doi.org/10.7717/peerj.6918

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mertens, G., & Morriss, J. (2021). Intolerance of uncertainty and threat reversal: A conceptual replication of Morriss et al. (2019). *Behaviour Research and Therapy*, 103799. https://doi.org/10.1016/j.brat.2020.103799

Moore, C. T. (2016). *gtheory: Apply Generalizability Theory with R*. https://CRAN.R-project.org/package=gtheory

Pittig, A., Treanor, M., LeBeau, R. T., & Craske, M. G. (2018). The role of associative fear and avoidance learning in anxiety disorders: Gaps and directions for future research. *Neuroscience & Biobehavioral Reviews*, *88*, 117–140. https://doi.org/10.1016/j.neubiorev.2018.03.015

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*, 103550. https://doi.org/10.1016/j.brat.2020.103550

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

Raeder, F., Merz, C., Margraf, J., & Zlomuzica, A. (2020). The association between fear extinction, the ability to accomplish exposure and exposure therapy outcome in specific phobia. *Scientific Reports*, *10*, 4288. https://doi.org/10.1038/s41598-020-61004-3

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research*.

Riccio, D. C., & Joynes, R. L. (2007). Forgetting of stimulus attributes: Some implications for hippocampal models of memory. *Learning & Memory*, *14*(6), 430–432. https://doi.org/10.1101/lm.617107

Rosenbaum, B. L., Bui, E., Marin, M.-F., Holt, D. J., Lasko, N. B., Pitman, R. K., Orr, S. P., & Milad, M. R. (2015). Demographic factors predict magnitude of conditioned fear. *International Journal of Psychophysiology*. https://doi.org/10.1016/j.ijpsycho.2015.06.010

Scharfen, J., Jansen, K., & Holling, H. (2018). Retest effects in working memory capacity tests: A meta-analysis. *Psychonomic Bulletin & Review*, *25*(6), 2175–2199. https://doi.org/10.3758/s13423-018-1461-6

Shrout, P. E., & Lane, S. P. (2012). Reliability. In *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 643–660). American Psychological Association. https://doi.org/10.1037/13619-034

Steinman, S. A., Namaky, N., Toton, S. L., Meissel, E. E. E., St. John, A. T., Pham, N.-H., Werntz, A., Valladares, T. L., Gorlin, E. I., Arbus, S., Beltzer, M., Soroka, A., & Teachman, B. A. (2021). Which Variations of a Brief Cognitive Bias Modification Session for Interpretations Lead to the Strongest Effects? *Cognitive Therapy and Research*, *45*(2), 367–382. https://doi.org/10.1007/s10608-020-10168-3

Torrents-Rodas, D., Fullana, M. A., Bonillo, A., Andión, O., Molinuevo, B., Caseras, X., & Torrubia, R. (2014). Testing the temporal stability of individual differences in the acquisition and generalization of fear. *Psychophysiology*, *51*(7), 697–705.

Vanbrabant, K., Boddez, Y., Verduyn, P., Mestdagh, M., Hermans, D., & Raes, F. (2015). A new approach for modeling generalization gradients: A case for hierarchical models. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00652

Vervliet, B., & Boddez, Y. (2020). Memories of 100 years of human fear conditioning research and expectations for its future. *Behaviour Research and Therapy*, *135*, 103732. https://doi.org/10.1016/j.brat.2020.103732

Wong, A. H. K., Wirth, F. M., & Pittig, A. (2022). Avoidance of learnt fear: Models, potential mechanisms, and future directions. *Behaviour Research and Therapy*, *151*, 104056. https://doi.org/10.1016/j.brat.2022.104056

Zeidan, M. A., Lebron-Milad, K., Thompson-Hollands, J., Im, J. J. Y., Dougherty, D. D., Holt, D. J., Orr, S. P., & Milad, M. R. (2012). Test–Retest Reliability during Fear Acquisition and Fear Extinction in Humans. *CNS Neuroscience & Therapeutics*, *18*(4), 313–317. https://doi.org/10.1111/j.1755-5949.2011.00238.x

Author Note

S.E.C. is funded by the NIH (F32 MH129136). J.E.D. is funded by the NIH (R01 MH122387, R00 MH106719) and the NSF (CAREER Award 1844792).

The authors report they have no conflicts of interest to disclose.