Received: 18 March 2022 | Revised: 28 November 2022 | Accepted: 2 December 2022

DOI: 10.1111/psyp.14242

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Test–retest reliability of human threat conditioning and generalization across a 1-to-2-week interval

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, Austin, Texas, USA

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

[3]Department of Psychology, West Virginia University, Morgantown, West Virginia, USA

Correspondence
Samuel E. Cooper, Dell Medical School, University of Texas at Austin, 1601 Trinity St Bldg B, Austin, TX 78701, USA.
Email: samuel.cooper@austin.utexas.edu

Funding information
National Institute of Mental Health, Grant/Award Number: F32 MH129136 and R01 MH122387; National Science Foundation, Grant/Award Number: CAREER Award 1844792

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, 46 human volunteers participated in an identical auditory threat acquisition and generalization protocol at two sessions separated by 1-to-2 weeks. Skin conductance responses (SCR) and trial-by-trial shock risk ratings served as primary measures. We used linear mixed effects modeling to test differential threat responses and generalization gradients, and Generalizability (G) theory coefficients as our primary formal assessment of test–retest reliability of intraindividual stability and change across time. Results showed largely invariant differential conditioning and generalization gradients across time. G coefficients indicated fair reliability for acquisition and generalization SCR. In contrast, risk rating reliabilities were mixed, and reliability was particularly low for acquisition risk ratings. Our findings generally support reliability of the threat conditioning and generalization paradigm for shorter test–retest intervals and highlight their utility for assessments of behavioral interventions in mental health research, but challenges remain and further work is needed. Threat conditioning and generalization tasks are increasingly used for translational efforts to improve behavioral interventions, and thus test–retest reliability for these tasks needs to be established. Our results support the test–retest reliability of threat conditioning and generalization over a relatively short (1-to-2 week) interval, but this depends on the measure used (physiological vs. self-report). Overall, these tasks could be appropriate for repeated testing over the course of a short-duration intervention study, but more research is needed, particularly in regard to longer-duration studies.

KEYWORDS
associative learning, psychometrics, reliability, skin conductance, threat generalization

1 | INTRODUCTION

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 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 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, van 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 psychophysiological 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 is not 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 provides mixed evidence for test–retest reliability within the same individuals. The majority of this work does not include a generalization test and focuses on differential threat acquisition and, in some cases, extinction (for a detailed survey of study parameters and results, see Klingelhöfer-Jens et al., 2022). The

earliest example, by Fredrikson et al. (1993), tested participants 20 days apart and only assessed SCR reliability using simple Pearson correlation for CS+ and CS− separately. Although test–retest was in the moderate-to-strong range, these correlations are sub-optimal for determining test–retest reliability (Heise, 1969). Other studies added to the literature with more modern test–retest metrics: Zeidan et al. (2012) found moderate-to-strong test–retest reliability for SCRs across three identical test sessions separated by up to 3 months each, and Ridderbusch et al. (2021) reported relatively weaker reliability for rating and neural measures across two testing sessions separated by 13 weeks. More recently, a comprehensive study by Klingelhöfer-Jens et al. (2022) tested participants six months apart and found fair-to-moderate test–retest reliability for multiple behavioral (SCR, fear ratings) and neural measures using several different quantification approaches (an important and timely topic in the literature, see Kuhn et al., 2022). This effort suggests that threat acquisition reliabilities are generally modest and can substantially differ by type of dependent measure. There is only one prior study directly testing test–retest reliability of generalization. Torrents-Rodas et al. (2014) also found mixed test–retest reliability for acquisition and threat generalization using visual stimuli (shapes) in the same individuals across two test sessions separated by 8 months. Measures included SCR, FPS, and shock expectancy risk ratings. Generalization test–retest reliability was highest for SCR, but notably lower for FPS and risk ratings. In a notable departure from prior studies, Torrents-Rodas et al. (2014) employed more complex test–retest reliabilities estimates that were appropriate for assessing the stability of generalization patterns, which are necessary given that generalization tests typically include more than 2 stimulus classes.

The above discussed studies provide promising, if mixed, evidence of test–retest reliability across conditioning indices. However, work in this area is scarce, and the majority used 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 1-to-2-week 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, & Phelps, 2017). Informed by Torrents-Rodas et al. (2014), we predicted that these generalization gradients would show significant test–retest reliability (i.e., coefficient 95% confidence intervals [CIs] do not contain zero) across a 1-to-2-week period.

2 | METHOD

2.1 | 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, five who returned for their second session after 15 days or longer, 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 $N = 46$ for the analyses described in the current effort. We did not apply any

performance-based exclusions (such as the common practice of excluding based on poor discrimination during acquisition, for critical discussion and compelling argument against this practice see Lonsdorf et al., 2019). 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 and characteristics ($N = 46$)

Age	
Mean (SD)	20.2 (2.97)
Gender	
Female	35 (76.1%)
Male	11 (23.9%)
Race	
American Indian/Alaska Native	1 (2.2%)
Black/African origin	1 (2.2%)
East Asian	2 (4.3%)
Other or Unknown	6 (13.0%)
White/European origin	36 (78.3%)
Ethnicity	
Hispanic or Latino	8 (17.4%)
Not Hispanic or Latino	38 (82.6%)
Education	
Associate's Degree	1 (2.2%)
Bachelor's Degree	3 (6.5%)
High School Graduate	4 (8.7%)
Some College	38 (82.6%)

2.2 | 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 threat 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;

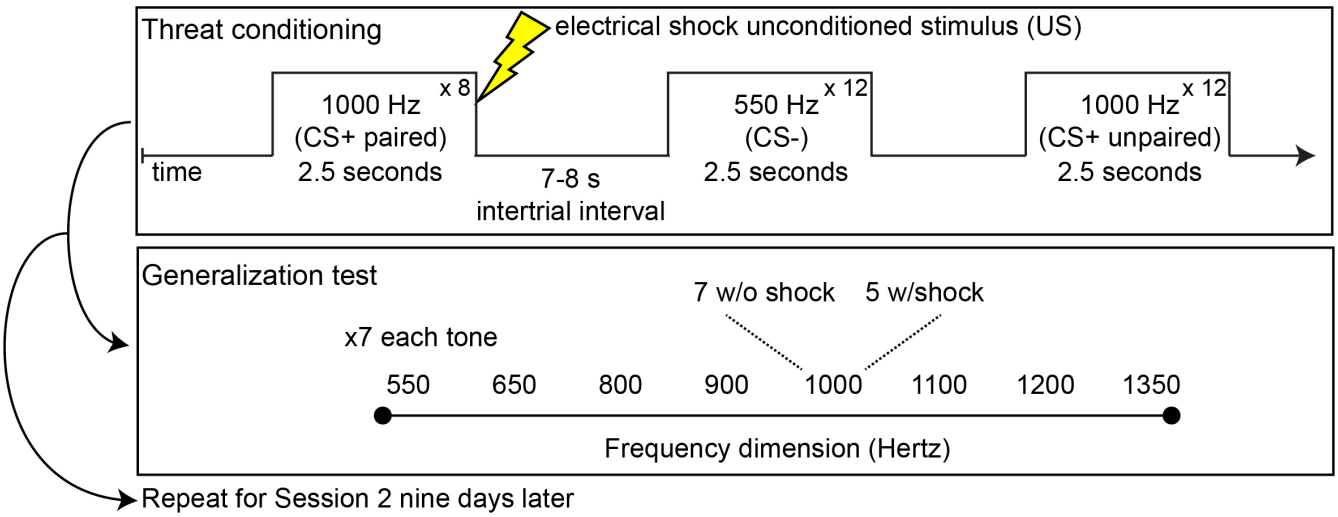

FIGURE 1 Threat conditioning design. Discriminative threat conditioning included pure tone conditioned stimuli paired (CS+, 1000 Hz) or unpaired (CS−, 550 Hz) with an aversive US. Generalization stimuli were novel tones spanning a frequency continuum between the CS− and CS+ and beyond the CS+. CS−, conditioned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

see also Blough, 1975; Dunsmoor et al., 2009; Lissek et al., 2008).

In all phases, we collected SCRs and trial-by-trial shock expectancy risk ratings. These ratings consisted of a three alternative-forced-choice scale corresponding to '1/no risk,' '2/moderate risk,' and '3/high risk' for receiving the US, based on prior generalization studies (scored as 1–3, 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.

2.3 | 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). We did not filter SCR data. 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). This analytic approach was chosen to align our test–retest work with the bulk of prior generalization studies, and threat conditioning studies in general, that use the same approach to quantifying SCR (Lonsdorf et al., 2017).

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.

2.4 | 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 IU. The primary aims and outcomes of this intervention are described elsewhere; briefly, only participants with higher IU were recruited, and results from this intervention revealed that the active CBM-I condition showed improvements compared to the control condition for task-assessed interpretation bias and self-reported symptoms. Preliminary analyses determined the intervention did not influence any conditioning task variables. This manuscript describes data from the threat generalization task completed at timepoint 1 (here, "initial session/session 1") and timepoint 4 (here, "follow-up session/session 2"), which were 1-to-2-weeks apart (median days = 9, mean days = 9.43, SD = 1.41, range = 7–14, IQR = 2). Trained researchers attached SCR and shock electrodes to participants and then guided them 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.

2.5 | Analytic plan

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

2.5.1 | Linear mixed models

We used linear mixed models (i.e., linear mixed-effects regression) 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, 2022). For both dependent variables, trial-level data for each stimulus was averaged and submitted to analysis. All models contained a random-intercept of participant and fixed effects of stimulus, session, and the Stimulus × 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.

2.5.2 | 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 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}:

$$R_{\mathrm{IRS}} = \frac{\sigma^2_{\mathrm{Participant\times Stimulus}}}{\left(\sigma^2_{\mathrm{Participant\times Stimulus}} + \sigma^2_{\mathrm{Residual}}\right)} \quad (1)$$

In Equation (1), $\sigma^2_{\mathrm{Participant\times Stimulus}}$ refers to individual variability in response to the experimental stimuli, and $\sigma^2_{\mathrm{Residual}}$ 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 timepoints, as opposed to stability of a particular response pattern:

$$R_{\mathrm{C}} = \frac{\sigma^2_{\mathrm{Participant\times Session}}}{\left(\sigma^2_{\mathrm{Participant\times Session}} + \left[\sigma^2_{\mathrm{Residual}} / m\right]\right)} \quad (2)$$

In Equation (2), $\sigma^2_{\mathrm{Participant\times Session}}$ refers to individual variability at each session (i.e., across time). The $\sigma^2_{\mathrm{Residual}}$ 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 would provide initial support for a measure being useful to track systematic changes in response over time, as opposed to change as a result of random error (which is represented by the $\sigma^2_{\mathrm{Residual}}$ residual term in Formula 2). This coefficient is perhaps most applicable to intervention work, as it is vital to ensure intervention change (i.e., systematic change) is not conflated with error-related change.

R_C 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

[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 instead 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.

very high R_C and R_{IRS} simultaneously. Of note is that high values for both R_C and R_{IRS} simultaneously would not be desirable for a treatment measure, because it would suggest that measure is not amenable to intervention-related change.

For both types of coefficients, we constructed 95% confidence intervals using the method provided in tab. 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 provide qualitative descriptions of coefficient size based on commonly applied recommendations (i.e., .4–0.75 is considered "fair-to-good" reliability, >0.75 considered "excellent" reliability, see Matheson, 2019), although we caution against stringent application of these standards for conditioning tasks for two reasons. First, there is limited work in this area and disagreement on firm guidelines regarding interpretation of within-person reliability (Matheson, 2019). Second, reliability cut-offs are typically based on psychometric work on self-report measures of psychological traits and states and thus are likely overly conservative for metrics with additional potential sources of error, including psychophysiological measurements.

To supplement our G coefficients and to provide additional points of comparison to prior studies, we provide individual stimuli intraclass correlation coefficients (ICCs) as commonly reported in the broader test–retest literature (Fisher, 1992). Specifically, we provide ICCs that measure absolute agreement between timepoints and assumes a random interval between timepoints (ICC2 in McGraw & Wong, 1996), which is appropriate given the variable number of days between Session 1 and Session 2 for some participants.

3 | RESULTS

3.1 | Differential threat conditioning

3.1.1 | SCR

Successful differential conditioning, operationalized as significantly larger CS+ responses (session 1: M = 0.51, SD = 0.33; session 2: M = 0.51, SD = 0.44) compared with CS− responses (session 1: M = 0.28, SD = 0.17; session 2: M = 0.17, SD = 0.14), was evident during acquisition at both sessions (session 1: $\beta = 1.3$, $t(87) = 6.26$, $p < .001$, 95% CI [0.89, 1.72]; session 2: $\beta = 2.37$, $t(88) = 6.13$, $p < .001$, 95% CI [1.6, 3.13]). Participants continued to respond more strongly to the CS+ compared with the CS− during generalization at both sessions (session 1: $\beta = .36$, $t(88) = 2.25$, $p = .027$, 95% CI [0.04, 0.67]; session 2: $\beta = 1.45$, $t(88) = 3.92$, $p < .001$, 95% CI [0.72, 2.19]).

3.1.2 | Ratings

Successful differential conditioning, operationalized as significantly larger CS+ risk ratings (session 1: M = 2.57, SD = 0.36; session 2: M = 2.63, SD = 0.34) compared with CS− ratings (session 1: M = 1.24, SD = 0.22; session 2: M = 1.18, SD = 0.21), was evident during acquisition at both sessions (session 1: $\beta = 5.91$, $t(72) = 19.58$, $p < .001$, 95% CI [5.31, 6.51]; session 2: $\beta = 6.76$, $t(84) = 30.27$, $p < .001$, 95% CI [6.31, 7.2]). Participants continued to expect the US more for the CS+ compared with the CS− during generalization at both sessions (session 1: $\beta = 1.46$, $t(76) = 6.67$, $p < .001$, 95% CI [1.02, 1.89]; $\beta = 2.26$, $t(83) = 11.31$, $p < .001$, 95% CI [2.19, 3.12]).

3.2 | Generalization gradients

3.2.1 | SCR

A model with a random-effect of testing session was the best fit for SCR data, $\chi^2(2) = 277.34$, $p < .001$. In this model, both stimulus, $\beta = .13$, $t(729) = 5.88$, $p < .001$, 95% CI [0.09, 0.17], and session, $\beta = -.11$, $t(729) = -2$, $p = .036$, 95% CI [−0.22, −0.01], predictors were significant. The Stimulus × Session interaction was not significant, $\beta = .02$, $t(728) = 1.84$, $p = .066$, 95% CI [0, 0.03]. Follow-up estimated marginal means analyses revealed that the CS− was the only stimulus to significantly differ across sessions, $b = 0.11$, $t(188) = 2.85$, $p_{bonferroni} = .034$, 95% CI [0.03, 0.19], with CS− magnitude larger at session 1 compared with session 2. See Figure 2a for visualized generalization gradients at each testing session.

The CBM intervention did not significantly affect gradients, as assessed through an additional model that included a Stimulus × Session × CBM Group interaction, $\beta = .001$, $t(724) = 1.56$, $p = .876$, 95% CI [−0.01, 0.01]. Mean IU interacted with the Stimulus × Session term, while controlling for change in IU between sessions, was also not significant, $\beta = .005$, $t(723) = -0.919$, $p = .358$, 95% CI [−0.01, 0]. Rerunning primary models with CBM Group and IU terms included as a separate fixed-effects yielded almost no change in the reported coefficients and did not change their significance.

3.2.2 | 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, $\beta = .19$, $t(700) = 6.03$, $p < .001$, 95% CI [0.13, 0.25], and session, $\beta = -.17$, $t(700) = -4.44$, $p < .001$, 95% CI [−0.24, −0.09], predictors were significant,

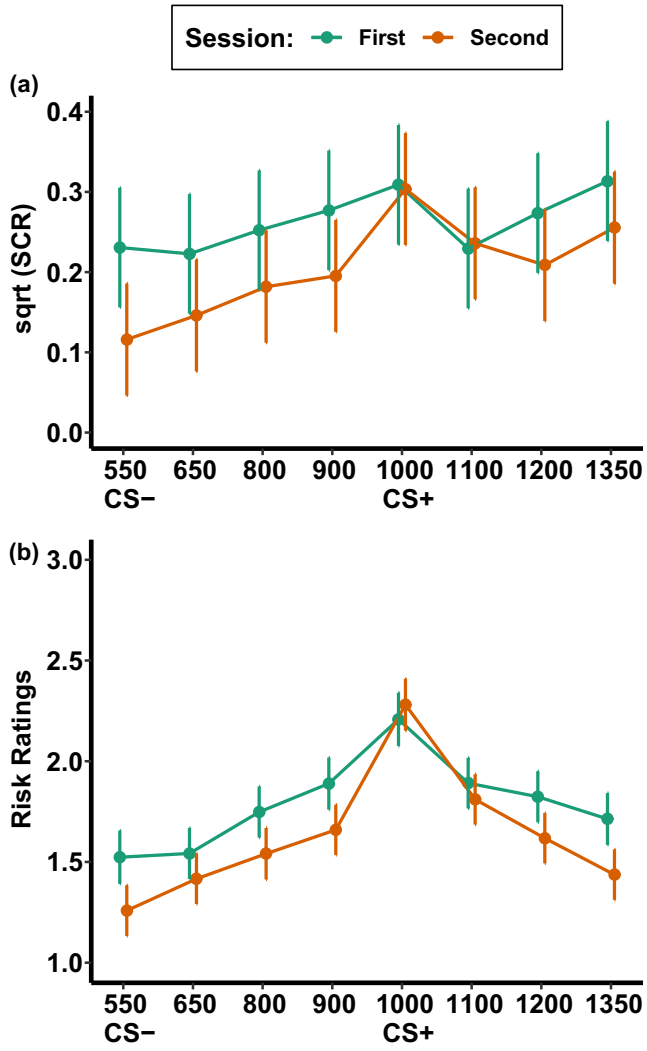

FIGURE 2 Conditioned generalization gradients at initial and follow-up session. All plotted values are fitted values from the linear mixed effects models described in text. Error bars represent 95% confidence intervals adjusted for random effects of the model. Panel a displays square-root transformed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

but as with the SCR data, the Stimulus × Session interaction was not significant $\beta = -.001$, $t(699) = -0.07$, $p = .938$, 95% CI [−0.02, 0.02]. Follow-up estimated marginal means analyses revealed that ratings for the CS−, $b = 0.26$, $t(506) = 3.29$, $p_{\text{bonferroni}} = .007$, 95% CI [0.1, 0.42], CS_{900}, $b = 0.23$, $t(491) = 2.92$, $p_{\text{bonferroni}} = .025$, 95% CI [0.07, 0.38], and CS_{1350}, $b = 0.27$, $t(491) = 3.52$, $p_{\text{bonferroni}} = .003$, 95% CI [0.12, .43], significantly differed across sessions, with ratings for all three of these stimuli significantly higher at session 1 compared with session 2. See Figure 2b for visualized generalization gradients at each testing session.

The CBM intervention did not significantly affect gradients, as assessed through an additional model that included a Stimulus × Session × CBM Group interaction, $\beta = -.003$, $t(695) = -0.416$, $p = .677$, 95% CI [−0.01, 0.01]. Mean IU interacted with the Stimulus × Session term, while controlling for change in IU between sessions, was also not significant, $\beta = -.004$, $t(694) = -0.589$, $p = .556$, 95% CI [−0.02, 0.01]. Rerunning primary models with CBM Group and IU terms included as a separate fixed-effects yielded almost no change in the reported coefficients and did not change their significance.

3.3 | Test–Retest reliability

Table 2 displays all variance components for each phase which were submitted to generalization coefficient calculations. Figures S1–S4 plot individual-level raw values and change slopes for all dependent variables, which can be used to visually assess variability of cross-session effects within each participant.

3.3.1 | 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 30% of variance, with the Participant × Stimulus (23%) and Participant × Session (13%) interactions accounting for 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%), 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 92% of variance, with the negligible remainder mostly accounted for by the Participant × Stimulus interaction (2%) and residual (7%) 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 (similarity, i.e., R_{IRS}) slightly varied depending on measure, with SCR demonstrating higher reliability, $R_{\text{IRS}} = .44$, 95% CI [0.17,

TABLE 2 Variance component analysis results

	Acquisition				Generalization			
	SCR		Ratings		SCR		Ratings	
Component	Variance	%	Variance	%	Variance	%	Variance	%
Participant	0.02	15	0.011	1	0.025	38	0.051	18
Session	0	0	0	0	0.001	2	0.013	5
Stimulus	0.04	30	0.97	92	0.001	2	0.064	23
Participant × Session	0.017	13	0.004	0	0.016	24	0.021	8
Participant × Stimuli	0.03	23	0.016	2	0.009	14	0.04	15
Session × Stimuli	0.002	2	0	0	0.001	1	0.005	2
Residual	0.022	17	0.058	6	0.013	20	0.082	30

Note: 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.

Abbreviations: ACQ, acquisition; GEN, generalization; SCR, skin conductance response.

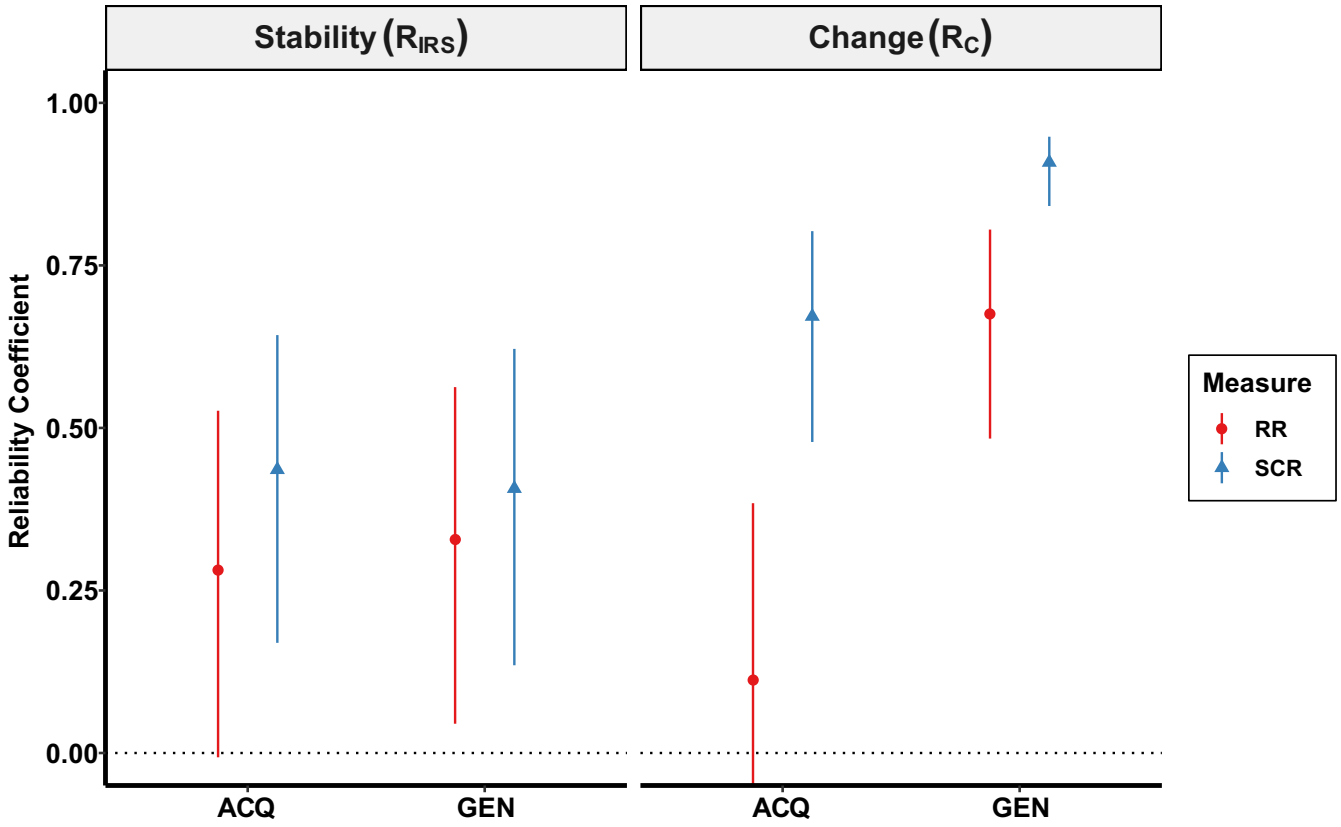

FIGURE 3 Test–retest reliabilities of SCR and risk rating data at each testing phase. The left panel displays stability (i.e., within-subject stability of response pattern) coefficients (R_{IRS}); the right panel displays change (i.e., within-subject reliability of change) coefficients (R_C). Error bars 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; RR, risk rating; SCR, skin conductance response; US, unconditioned stimulus.

0.64], than risk ratings, $R_{IRS} = 0.28$, 95% CI [−0.01, 0.53]. Reliability of change (i.e., R_C) across time also differed by measure. SCR reliability, $R_C = 0.67$, 95% CI [0.47, 0.80], was notably higher than risk rating reliability, $R_C = 0.11$, 95% CI [−0.17, 0.38]. Of note is that the risk rating R_{IRS} and R_C coefficients were the only generalizability coefficients calculated in the current effort that were not significant. In terms of individual stimulus reliabilities, those

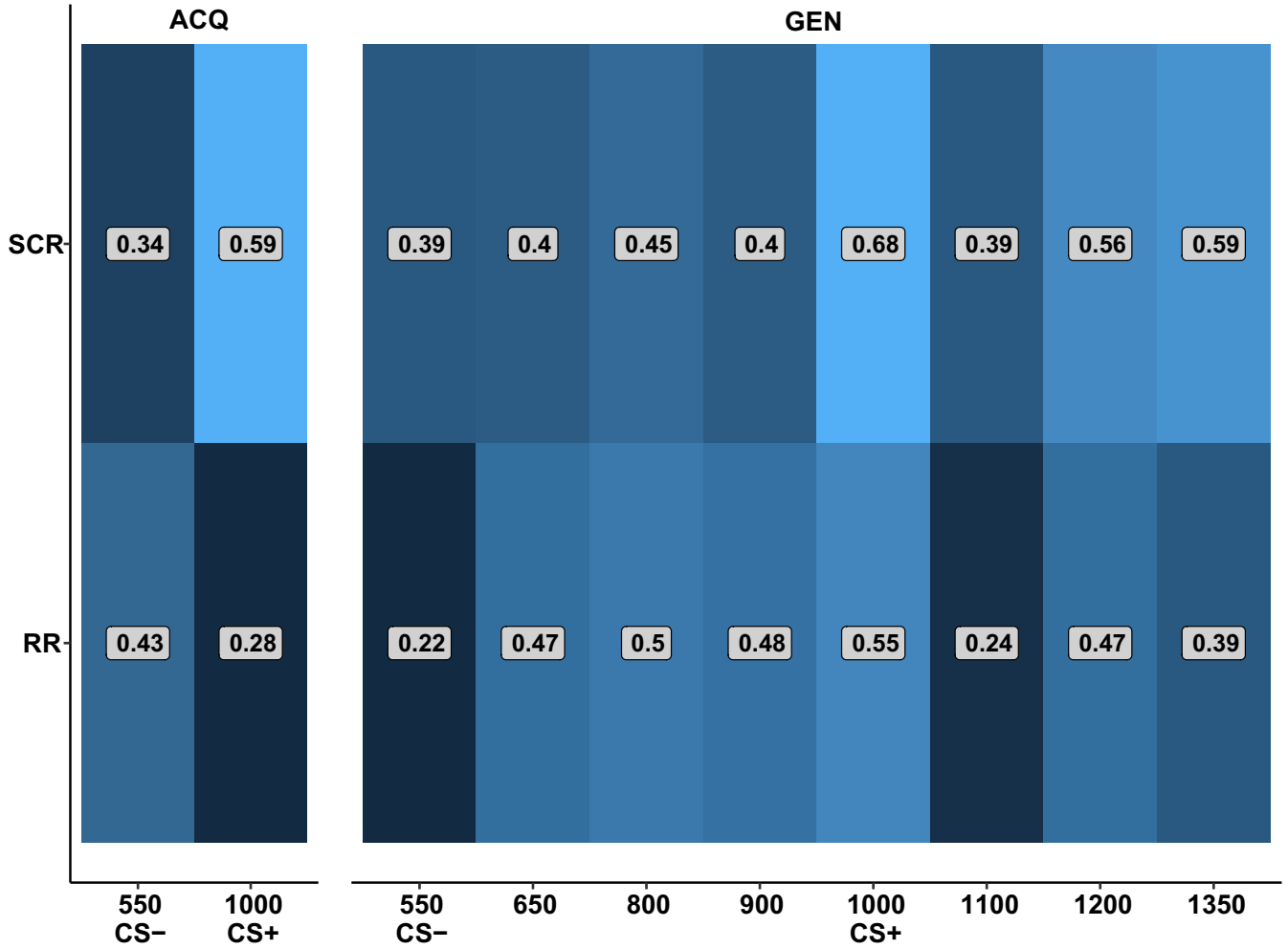

FIGURE 4 Individual stimuli test–retest ICCs for SCR and risk rating data across testing phases and sessions. Lighter blue panels indicate larger ICCs (i.e., higher test–retest reliability). ICC values are located in the gray boxes within each panel. ACQ, acquisition phase; CS−, conditioned safety cue; CS+, conditioned threat cue; GEN, generalization phase; ICC, intraclass correlation coefficient; RR, risk rating; SCR, skin conductance response.

for SCRs were lower for the CS−, ICC = 0.34, 95% CI [0.04, 0.58], compared with the CS+, ICC = 0.59, 95% CI [0.36, 0.75]. This pattern was reversed for risk ratings, with CS− reliability, ICC = 0.43, 95% CI [0.17, 0.64], higher than CS+ reliability, ICC = 0.28, 95% CI [−0.01, 0.52], which was not significantly different than zero. See Figure 4 for visualized individual stimulus ICCs.

3.3.2 | Generalization

The pattern of variance components for generalization markedly differed from those from acquisition (see Table 2). For SCR, the largest variance component was the participant component (38%), followed by the Participant × Time (24%) and Participant × Stimulus (14%) 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 (20%), indicating marked error variance in SCR at this phase. In contrast, the largest predictor variance component for risk ratings was the stimulus component (23%), followed by participant (18%) and the Participant × Stimulus interaction (15%). Of note is that for risk ratings, residual variance also accounted for the overall largest proportion of variance (30%), indicating a substantial amount of variance could not be explained by the predictors.

The pattern of test–retest generalizability coefficients for the generalization phase was largely similar to the pattern observed in acquisition. However, all coefficients were significant for this phase (see Figure 3). The reliability of within-person patterns of responding was again higher for SCR, R_{IRS} = 0.41, 95% CI [0.13, 0.62], compared with risk ratings, R_{IRS} = 0.33, 95% CI [0.05, 0.56]. This was also

the case for reliability of change (SCR: $R_{IRS} = 0.9$, 95% CI [0.84, 0.94]; risk ratings: $R_{IRS} = 0.67$, 95% CI [0.48, 0.80]).

In terms of individual stimulus reliabilities, SCR reliability was highest for the CS+, CS_{1200}, and CS_{1350} (ICCs ≥ 0.56), while the CS− and CS_{1100} demonstrated the lowest reliability (ICCs = 0.39). For risk ratings, the CS+ again demonstrated the highest reliability, ICC = 0.55, 95% CI [0.31, 0.72], and the CS− again demonstrated the lowest reliability (ICC = 0.22). Additionally, the CS− and CS_{1100} risk ratings were the only individual stimuli reliabilities to be non-significantly different from zero. See Figure 4 for visualization of all individual stimulus ICCs.

4 | DISCUSSION

Given the prominent role of threat 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 threat acquisition and stimulus generalization tests, SCR and risk ratings, across a 1-to-2-week interval. Our primary goal was to assess the test–retest of threat generalization. For SCR, reliability was generally fair across two types of G coefficients, one indexing within-person stability of response patterns and the other indexing reliability of change across time. However, reliability was notably poorer for risk ratings. Additionally, generalization gradients did not significantly differ at each time-point, although there was some across session variability at the level of individual stimuli. These results provide moderate support for the conclusion that threat generalization gradients remain relatively stable using an identical protocol at two timepoints, but in line with prior work, that reliability is modest in magnitude and related to the specific type of measure. These findings provide useful information regarding the utility 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 for SCR, and poorer for risk ratings. Despite the relatively lower risk rating coefficients in the current study, all coefficients were larger than those found in the other reliability study of threat generalization (Torrents-Rodas et al., 2014) that retested generalization after an 8-month interval. Specifically, test–retest stability across the 8-month interval ranged from $R_{IRS} = 0.23$ to $R_{IRS} = 0.34$, compared to $R_{IRS} = 0.28$ to $R_{IRS} = 0.44$ 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 h, 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.

We also found that test–retest reliabilities differed for individual stimuli during generalization. Both risk ratings and SCRs for the CS− were among the lowest test–retest reliabilities (ICC), and mean CS− for both measures also significantly differed between sessions. This result potentially points to participants during Session 1 forming a stable and enduring memory of the CS−, which then facilitates much lower Session 2 responses. Additionally, the CS− is the only unambiguous stimulus, as CS+ is not always reinforced with shock and the intermediary CSs are ambiguous by design due to their increasing similarity to the CS+. Also relevant is that lower reliabilities for some, but not all, stimuli might be a key contributor to lower reliabilities for the full generalization gradients. Given the number of stimulus classes commonly used in generalization tests, this suggests alternative generalization quantification strategies that can mitigate the influence of individual stimuli are needed. One option is to use a limited number of parameters that describe the shape of the gradient (for discussion, see Lee et al., 2020). Another option is to apply a latent variable approach and to assess the reliability of latent generalization variables that underlie manifest generalization indicators (e.g., latent growth curves, for applied example see Gazendam et al., 2020).

Taken together, the current results suggest that a 1-to-2-week interval results in generally stable generalization over time, with some exceptions. 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. It also must be emphasized that the current effort cannot

necessarily be extrapolated into inferences regarding much longer periods of time and refer readers to studies testing over long periods (see Klingelhöfer-Jens et al., 2022).

Contrary to the reliability observed for generalization metrics, both stability and change in acquisition risk rating across testing sessions evidenced relatively poorer reliability. An explanation for the overall poorer reliability for acquisition risk ratings (and for risk ratings in general in this investigation) is that fewer response options (i.e., 3 response options, as used in the current study) tend to be associated with lower reliability (e.g., Weng, 2004). We strongly encourage future studies to use rating scales with more response options to circumvent this issue. Additionally, one plausible explanation for the poor change reliability (which was the lowest generalizability coefficient we found) is related to interindividual differences in memory for the US contingency learned in the first session. 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 relatively 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). Although this could not be tested in the current study given sample size limitations, future studies with larger sample sizes would benefit from subgroup analysis of those with and without near-invariant responding.

One limitation of the current study is that we did not collect qualitative or quantitative 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. 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. We also note that although IU is frequently an individual difference of interest in relation to threat conditioning, work on relating IU to threat generalization has been inconsistent (e.g., Bauer et al., 2020; Hunt et al., 2019; Nelson et al., 2014), and would not necessarily affect test–retest reliability. Further, psychometric approaches to test–retest reliability are predicated on the assumption that between-subjects variance is minimally influential compared with within-subjects variance, and does not require all participants have the same level of a particular trait (most clearly seen in psychometrics applied to self-report questionnaires, where the assumption is that there will be between-subjects variability on multiple traits related to the outcome of interest; e.g., Enkavi et al., 2019). Additionally, meta-analyses find that test–retest reliability does not differ by clinical individual differences in several commonly used cognitive neuroscience and neuropsychology tasks (e.g., Calamia et al., 2013; Elliott et al., 2020). That said, the field would benefit from additional psychometric assessment of conditioning tasks where candidate individual differences, including IU, are comprehensively sampled and systematically tested over different time periods. Further, when possible, we recommend assessing reliability in a fully representative sample that has not been preselected for a certain level of a trait. The possibility remains, both for the current work and similar future studies, that reliability would differ in a sample that has not been constrained based on a psychological variable of interest.

Another limitation in the current study is that 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 threat conditioning metrics (e.g., Cooper, Hunt, et al., 2022; Rosenbaum et al., 2015). Finally, we note that the current investigation used a relatively small sample-size, although it was similar to the prior study of generalization test–retest by Torrents-Rodas et al. (2014). Future studies of generalization test–retest would benefit from increased sample-sizes to facilitate more comprehensive test–retest evaluations, such as those performed with the N = 120 sample of Klingelhöfer-Jens et al. (2022). An additional issue of note is in regards to psychophysiological data (SCR, in the current study). Different strategies for SCR quantification can impact inferential statistics (Kuhn et al., 2022; Lonsdorf et al., 2019) and reliability metrics (Klingelhöfer-Jens et al., 2022). In the current study, we use the most common SCR approach in the generalization literature to maximize the compatibility of our work with the prior literature. That said, the field would benefit from more formal analysis of different physiological quantification pipelines in relation to generalization reliability (in line with a move toward multiverse analyses, e.g., Klingelhöfer-Jens et al., 2022; Kuhn et al., 2022). One particularly important avenue for future research in this area are quantification approaches that minimize trial-by-trial variability (e.g., model-based approaches, see Kuhn et al., 2022), and therefore would potentially limit the impact that initial learning trials during the first session have on overall test–retest reliability.

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 a larger sample of 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 threat generalization paradigms are reliable over a short interval and can be appropriate for assessing behavioral intervention effects.

AUTHOR CONTRIBUTIONS

Samuel E. Cooper: Conceptualization; formal analysis; methodology; visualization; writing – original draft; writing – review and editing. **Joseph E. Dunsmoor:** Conceptualization; investigation; methodology; visualization; writing – original draft; writing – review and editing. **Kathleen A. Koval:** Data curation; writing – review and editing. **Emma R. Pino:** Data curation; writing – review and editing. **Shari A. Steinman:** Conceptualization; data curation; funding acquisition; investigation; methodology; project administration; resources; supervision; writing – review and editing.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in an Open Science Framework repository at https://www.doi.org/10.17605/OSF.IO/ZQFKJ.

ORCID

Samuel E. Cooper https://orcid.org/0000-0001-9563-1750
Joseph E. Dunsmoor https://orcid.org/0000-0002-5448-6873
Shari A. Steinman https://orcid.org/0000-0002-0068-206X

REFERENCES

Adolph, D., Flasinski, T., Lippert, M. W., Pflug, V., Hamm, A. O., Richter, J., 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), 255–278. 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), 1–51. https://doi.org/10.18637/jss.v067.i01

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

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), 3–21. https://doi.org/10.1037/0097-7403.1.1.3

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., & 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

Cooper, S. E., van Dis, E. A. M., Hagenaars, M. A., Krypotos, A.-M., Nemeroff, C. B., Lissek, S., Engelhard, I. M., & Dunsmoor, J. E. (2022). A meta-analysis of conditioned fear generalization in anxiety-related disorders. *Neuropsychopharmacology*, *47*, 1652–1661. https://doi.org/10.1038/s41386-022-01332-2

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*, *110*, 1754–1776. https://doi.org/10.1016/j.neuron.2022.03.001

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*, *114*, 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*, *31*, 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

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10.1073/pnas.1818430116

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

Fredrikson, M., Annas, P., Georgiades, A., Hursti, T., & Tersman, Z. (1993). Internal consistency and temporal stability of classically conditioned skin conductance responses. *Biological Psychology*, *35*(2), 153–163. https://doi.org/10.1016/0301-0511(93)90011-V

Gazendam, F. J., Krypotos, A.-M., Kamphuis, J. H., van der Leij, A. R., Huizenga, H. M. H., Eigenhuis, A., & Kindt, M. (2020). From adaptive to maladaptive fear: Heterogeneity in threat and safety learning across response systems in a representative sample. *International Journal of Psychophysiology*, *158*, 271–287. https://doi.org/10.1016/j.ijpsycho.2020.09.017

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

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

Heise, D. R. (1969). Separating reliability and stability in test-retest correlation. *American Sociological Review*, *34*(1), 93–101. https://doi.org/10.2307/2092790

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*, 175. 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*, *95*, 821–835. https://doi.org/10.1002/jnr.23837

Klingelhöfer-Jens, M., Ehlers, M. R., Kuhn, M., Keyaniyan, V., & Lonsdorf, T. B. (2022). Robust group- but limited individual-level (longitudinal) reliability and insights into cross-phases response prediction of conditioned fear. *eLife*, *11*, e78717. https://doi.org/10.7554/eLife.78717

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

Kuhn, M., Gerlicher, A. M. V., & Lonsdorf, T. B. (2022). Navigating the manyverse of skin conductance response quantification approaches—A direct comparison of trough-to-peak, baseline correction, and model-based approaches in Ledalab and PsPM. *Psychophysiology*, *59*(9), e14058. https://doi.org/10.1111/psyp.14058

Lee, J. C., Mills, L., Hayes, B. K., & Livesey, E. J. (2020). Modelling generalisation gradients as augmented Gaussian functions. *Quarterly Journal of Experimental Psychology*, *74*, 106–121. https://doi.org/10.1177/1747021820949470

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., Klingelhöfer-Jens, M., Andreatta, M., Beckers, T., Chalkia, A., Gerlicher, A., Jentsch, V. L., Meir Drexler, S., Mertens, G., Richter, J., Sjouwerman, R., Wendt, J., & Merz, C. J. (2019). Navigating the garden of forking paths for data exclusions in fear conditioning research. *eLife*, *8*, e52465. https://doi.org/10.7554/eLife.52465

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

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), 501–510. 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

Nelson, B. D., Weinberg, A., Pawluk, J., Gawlowska, M., & Proudfit, G. H. (2014). An event-related potential investigation of fear generalization and intolerance of uncertainty. *Behavior Therapy*, *46*, 661–670. https://doi.org/10.1016/j.beth.2014.09.010

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. (2022). *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

Ridderbusch, I. C., Wroblewski, A., Yang, Y., Richter, J., Hollandt, M., Hamm, A. O., Wittchen, H.-U., Ströhle, A., Arolt, V., Margraf, J., Lueken, U., Herrmann, M. J., Kircher, T., & Straube, B. (2021). Neural adaptation of cingulate and insular activity during delayed fear extinction: A replicable pattern across assessment sites and repeated measurements. *NeuroImage*, *237*, 118157. https://doi.org/10.1016/j.neuroimage.2021.118157

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*, *98*, 59–64. 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. https://doi.org/10.1111/psyp.12213

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*, 652. 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

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability.

Educational and Psychological Measurement, *64*(6), 956–972. https://doi.org/10.1177/0013164404268674

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

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

FIGURE S1 Individual cross-session change slopes for SCR during generalization.

FIGURE S2 Individual cross-session change slopes for risk ratings during generalization.

FIGURE S3 Individual cross-session change slopes for SCR during acquisition.

FIGURE S4 Individual cross-session change slopes for risk ratings during acquisition.

How to cite this article: Cooper, S. E., Dunsmoor, J. E., Koval, K. A., Pino, E. R., & Steinman, S. A. (2022). Test–retest reliability of human threat conditioning and generalization across a 1-to-2-week interval. *Psychophysiology*, *00*, e14242. https://doi.org/10.1111/psyp.14242