**Modulatory effects of instructions on extinction efficacy in appetitive and aversive learning: A registered report**

Lea Busch1\*, Katja Wiech2, Matthias Gamer3, Balint Kincses1, Tamas Spisak4, Katharina Schmidt1#, Ulrike Bingel1#

\* First author

# shared last author

1 Department of Neurology, Center for Translational Neuro- and Behavioural Sciences, University Medicine Essen, Essen, Germany

2 Wellcome Centre for Integrative Neuroimaging (WIN), Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, OX3 9DU, United Kingdom.

3 Department of Psychology (Experimental Clinical Psychology), University of Würzburg, Würzburg, Germany

4 Institute for Diagnostic and Interventional Radiology and Neuroradiology, University Medicine Essen, Essen, Germany

Abstract

In the context of pain, extinction learning has been shown to be slower or incomplete for aversive compared to appetitive cues (i.e., cues signaling pain exacerbation and pain relief, respectively), potentially due to their higher biological relevance. In a therapeutic context, this reluctant extinction that has been discussed as indicative of a ‘better safe than sorry strategy’ could potentially be reduced by making patients aware of the change in contingency. Using a classical conditioning paradigm, we will test whether slower or incomplete extinction learning from aversive than appetitive cues can be prevented when participants are explicitly informed about contingency changes using verbal instructions. Geometric figures will serve as conditioned stimuli, and individually calibrated temperature changes applied to capsaicin-pretreated skin, inducing pain exacerbation or pain relief, respectively, from tonic pain levels will serve as unconditioned stimuli. Behavioral measures (expectancy and valence ratings) and physiological measures (pupillometry, skin conductance responses) will be collected as outcome measures. Task-independent neural activity (resting-state functional magnetic resonance imaging) will be assessed prior to the behavioral paradigm. We expect stronger acquisition and extinction learning with instruction and will assess whether this effect is more pronounced for appetitive stimuli. Moreover, we intend to identify neural markers that are associated with the modulatory effects of expectations on appetitive and aversive learning.

Keywords: Instructed extinction; expectation; pain conditioning; eye tracking; rsfMRI

1. **Introduction**

Learning which stimuli or events precede pain is amongst the most important forms of learning, as it allows us to avoid similar painful experiences in the future. Experimental studies have shown that - due to their pain-predictive nature - these cues begin to elicit a learned response, e.g., fear of movement-related pain (Meulders & Vlaeyen, 2013). However, similarly important is the updating of these learned associations when the cue is no longer followed by pain, as such extinction learning ensures that cognitive resources are allocated primarily to cues signaling impending threat. Acquisition and extinction can be investigated experimentally using a classical conditioning paradigm and within different aversive modalities including pain. In a previous study including (aversive) increases and (appetitive) decreases in pain, we found stronger acquisition learning of associations between unconditioned stimuli (US) and conditioned stimuli (CS) for aversive as compared to appetitive CS, as indicated by differences in CS valence ratings (van der Schaaf & Schmidt et al., 2022). These observations may reflect a ‘better safe than sorry strategy’ in the context of pain conditioning (Solomon & Wynne, 1954), which implies that the anticipation and avoidance of potential threat may be prioritized to ensure the survival of the organism. However, our data showed no differences in the changes in differential CS valence ratings over the course of the extinction training, i.e., in extinction slopes. Instead, we found incomplete extinction at the end of the phase in absolute aversive CS valence ratings only, as also previously reported in a study comparing appetitive and aversive effects of food and painful electrical stimulation (Andreatta & Pauli, 2015). To further test extinction efficacy, a reinstatement manipulation and test phase, which includes the unannounced presentation of US without CS, followed by a test phase without reinforcement, can be used. Reinstatement effects have been shown for pain-related learning of different modalities (Icenhour et al., 2015; Meulders et al., 2015; Schmidt et al., 2020) but have not been studied for pain relief-related learning.

Clinically, the development and maintenance of maladaptive responses to pain after tissue healing have been discussed to drive pain chronification as, for instance, described in the fear-avoidance model of pain (Flor, 2012; Linton, 2000; Vlaeyen et al., 1995). Thus, swift updating of learned representations is important, but remains a key challenge in the treatment of chronic pain (Flor, 2012; Leeuw et al., 2007; Vlaeyen & Linton, 2012).

Besides learning from first-hand experience, external information informs our representation of cue-outcome associations. In pain treatment, for instance, psychoeducation is used to point out triggers of pain alleviation and exacerbation and changes in their association with pain over time. Experimentally, this influence of external information can be investigated using instructed extinction, i.e., by informing participants about the subsequent omission of the US prior to the extinction training (Luck & Lipp, 2016). Instructions have been shown to reliably facilitate extinction learning across different stimuli and outcomes (Luck & Lipp, 2016). In a study by Sevenster et al. (2012), instructions reduced US expectancy ratings already at the first trial of the extinction training and during the subsequent reinstatement phase. This suggests that explicit information about a change in contingencies might be a promising method to prevent the incomplete extinction of aversive associations as observed by van der Schaaf & Schmidt et al. (2022).

So far, the effects of instructed extinction on appetitive and aversive learning have not been compared directly. Here we use a differential conditioning paradigm featuring both appetitive and aversive US in the context of pain (compare van der Schaaf & Schmidt et al., 2022) which allows to test for potentially different effects of instruction on extinction in these two opposing domains.

US expectancy ratings will be the main outcome and will therefore determine the overall evaluation of our hypotheses. Additionally, we will assess CS valence ratings as proxies for the emotional learning component. These ratings have been shown to extinguish more slowly than US expectancy ratings, and CS valence following extinction predicted the degree of reinstatement (e.g., Dirikx et al., 2004; Zbozinek, Hermans, et al., 2015; Zbozinek, Holmes, et al., 2015). We will furthermore measure SCR as an established autonomic physiological readout in conditioning studies (e.g., Andreatta & Pauli, 2015; Jentsch et al., 2020; Schlitt et al., 2022; van der Schaaf et al., 2022) and studies investigating instruction effects in conditioning paradigms (Atlas et al., 2016; Atlas & Phelps, 2018; Costa et al., 2015; Javanbakht et al., 2017; Mertens et al., 2018; Mertens & De Houwer, 2016; Sevenster et al., 2012; Wendt et al., 2020). Pupil dilation will be assessed as a second physiological measure as it has recently been shown to be less prone to habituation effects, thus potentially making it more sensitive to extinction effects than SCR (Leuchs et al., 2019). A recent meta-analysis suggested this measure for both appetitive and aversive conditioning (Finke et al., 2021).

In addition to the investigation of behavioral and autonomic physiological parameters, the neural basis of fear conditioning has been subject of extensive research. In a meta-analysis, Fullana et al. (2016) identified a large functional brain network that responds to aversive events, comprising areas such as the anterior insular cortex, dorsal anterior cingulate cortex (dACC), dorsomedial prefrontal cortex (dmPFC) and ventral striatum. Importantly, a comparison of these circuitries with appetitive neural networks identified using a monetary reward task revealed very similar activations (Klein et al., 2022). This finding was further substantiated by another meta-analysis comparing studies investigating the anticipation of monetary reward or loss which revealed shared activation in the striatum, amygdala, and insula (Oldham et al., 2018). Furthermore, Hayes et al. (2014) concluded from their more general cross-species meta-analysis of neuronal activity to appetitive and aversive stimuli that many identified regions were associated with both appetitive and aversive processes. Together with a study that revealed both shared and specific regions for pain relief learning compared to regular appetitive learning (Leknes et al., 2011), these studies suggest a common neural system for appetitive and aversive learning mechanisms.

Studies investigating the effect of instructions on brain processes of acquisition and extinction learning are still sparse. Atlas et al. (2016) reported effects of instructions on activity in the ventromedial PFC (vmPFC) / orbitofrontal cortex (OFC) and the striatum, but not the amygdala. Both the OFC and striatum contain distinct regions for encoding explicit contingency information and value (Pauli et al., 2019) and amygdala engagement has been linked to the value of a CS for both appetitive and aversive stimuli (Belova et al., 2008; for a review, see Sharpe & Schoenbaum, 2016). Thus, contingency-encoding regions might update to instructions more easily, especially when instructions are contingency-focused (Atlas et al., 2016), which might be decoupled from affective ratings (Luck & Lipp, 2016).

Cognitive function can also be linked to task-independent brain connectivity (Smith et al., 2009). Networks identified using functional activation showed a high similarity to those at rest (Smith et al., 2009). It has been argued that resting state connectivity may reflect preparatory states that allow efficient processing of stimuli relevant to the respective neural system (Hashmi et al., 2014). Beyond cognitive function, pain-related measures, such as pain sensitivity, can be predicted from pain-free resting states (Spisak et al., 2020), and pain chronification has been predicted based on the connectivity of the reward system and default mode network (Pfannmöller & Lotze, 2019). In relation to learning, connectivity (particularly of the (v)mPFC, amygdala) has been associated with processes such as renewal (Lissek & Tegenthoff, 2021), fear generalization, clinical anxiety (Cha et al., 2014) and treatment outcome (Klumpp et al., 2014). However, interindividual differences in acquisition and extinction have not been linked to pre-conditioning resting state connectivity so far, but only to changes in connectivity or connectivity acquired post-conditioning (Belleau et al., 2018; Feng et al., 2016; Martynova et al., 2020; Schultz et al., 2012). Such an approach could allow for the extraction of specific markers that are associated with different patterns of learning, which, in the future, could be utilized to identify individuals who are more prone to chronification from an acute injury.

Based on previous results, we aim to investigate in a first step (manuscript 1) whether providing instructions before extinction training affects extinction efficacy following appetitive and aversive conditioning in the context of pain. To this end, we will first investigate whether we can replicate our previous findings of faster acquisition of aversive than appetitive CS, and incomplete extinction of aversive CS in the uninstructed group (van der Schaaf & Schmidt et al. (2022). Second, we will study the effect of instructions on extinction and expect an overall facilitating effect, and we will test whether this effect is stronger for appetitive than aversive stimuli in the instructed group. Extinction efficacy will also be tested using a reinstatement manipulation and test phase. For statistical analyses, we will apply classical frequentist statistics as well as calculate Bayes factors. This approach will enable us to rate the strength of evidence in favor or against our proposed hypotheses on a continuous scale as the Bayes factor allows to discriminate evidence of absence from absence of evidence.

We state the following hypotheses, which are also displayed in Table 1 along with the respective research question, analysis plan, and interpretation:

H1: We expect steeper acquisition slopes of conditioned responses (CR) to aversive than appetitive CS, indicated by a *CS type* (appetitive, aversive) × *time* interaction, with US expectancy ratings (H1a), SCR (H1b), pupil dilation (H1c), and CS valence (H1d) as outcome measures.

H2: We expect steeper extinction slopes of CR to appetitive than aversive CS in the uninstructed group, indicated by a *CS type* (appetitive, aversive) × *time* interaction, with US expectancy ratings (H2a), SCR (H2b), pupil dilation (H2c), and CS valence (H2d), as outcome measures.

H3: We expect steeper extinction slopes of CR in the instructed than the uninstructed group, indicated by an *instruction group* (instructed, uninstructed) × *time* interaction, with US expectancy ratings (H3a), SCR (H3b), pupil dilation (H3c), and CS valence (H3d), as outcome measures.

H4: We will test whether the instruction differentially affects the extinction slopes of CR to appetitive and aversive CS. Such an effect would be indicated by a *CS type* × *instruction group* × *time* interaction, with US expectancy ratings (H4a), SCR (H4b), pupil dilation (H4c), and CS valence (H4d) as outcome measures. Based on the idea of a ‘better safe than sorry strategy’ (Solomon & Wynne, 1954; van der Schaaf & Schmidt et al., 2022), we expect the facilitating effect of instructions to be stronger for appetitive than for aversive CS.

In another manuscript (manuscript 2), we aim to identify functional connectivity-based brain markers assessed with resting state fMRI acquired prior to task performance that are associated with an individual’s aversive and appetitive learning during acquisition and extinction training, and the effect of the instruction. US expectancy will serve as the main behavioral outcome measure as it constitutes the main outcome in the behavioral manuscript (manuscript 1) and as it has shown conditioning effects and effects of instructions on conditioning in previous studies (e.g., Duits et al., 2017; Mertens et al., 2016; Scheveneels et al., 2019; Sevenster et al., 2012). Since studies on the association between resting state connectivity and appetitive and aversive learning and instructions are scarce, we based our hypotheses on task-based studies, assuming that the respective regions and their connectivity will also be relevant at rest (Hashmi et al., 2014; Smith et al., 2009). VmPFC, amygdala, and striatum have been related to appetitive and aversive learning mechanisms including acquisition and extinction efficacy (Battaglia et al., 2022; Becerra et al., 2013; Belleau et al., 2018; Doll et al., 2009; Fullana et al., 2016; Klein et al., 2022; Leknes et al., 2011; Leknes & Tracey, 2008; Martynova et al., 2020; Milad & Quirk, 2012; Oldham et al., 2018; Sescousse et al., 2013; Seymour et al., 2005; Wendt & Morriss, 2022). Due to their different functional roles in learning (Balleine et al., 2007; Liljeholm & O’Doherty, 2012; Stanley et al., 2021), we will differentiate between dorsal and ventral aspects of the striatum. Both regions will further be referenced to as striatum.

We will investigate whether connectivity between the listed regions of interest (ROI) is associated with indices of appetitive and aversive acquisition and extinction learning. Finally, we want to test which brain regions are involved in mediating the effect of instruction, as would be evident in a stronger association of resting-state brain connectivity with an individual’s extinction efficacy in the instructed as compared to the uninstructed group. Such a finding might help tailoring individual therapy plans for chronic pain conditions. Regarding instruction effects, the dorsolateral PFC (dlPFC) has been shown to affect activity in the striatum and vmPFC in the context of reward and aversive reversal learning (Atlas et al., 2016; Li et al., 2011). Following a conservative approach, we focus on these key ROIs in our seed-based functional connectivity (SBFC) analyses.

We state the following hypotheses (see Table 2 for the respective analysis plan and interpretation):

H1+2: We expect higher functional connectivity of the stated ROIs (i.e., vmPFC, amygdala, and striatum), extracted performing SBFC analyses based on Pearson correlations, to be associated with an individual’s acquisition and extinction of both aversive (H1) and appetitive (H2) CR, as assessed by the slopes of US expectancy.

H3+4: We expect connectivity of the stated ROIs (i.e., dlPFC, vmPFC, and striatum), extracted performing SBFC analyses based on Pearson correlations, to be associated with the effect of instruction as assessed immediately before and after instruction (H3) and over the course of the extinction training (H4). This would be evident in steeper slopes of the association between resting-state brain connectivity and extinction efficacy in the instructed compared to the uninstructed group.

Table 1. Study design table with research questions and corresponding hypotheses, analysis plan and interpretation for manuscript 1.

| **Question** | **Hypothesis** | **Sampling plan** | **Analysis Plan** | **Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis** | **Interpretation given different outcomes** | **Theory that could be shown wrong by the outcomes** |
| --- | --- | --- | --- | --- | --- | --- |
| Manipulation check: Was the perception of US painfulness [and US (un‑)pleasantness] during the acquisition training manipulated successfully? | We expect higher US painfulness [and US (un‑)pleasantness] ratings for the USincrease than the USmedium, and higher ratings for the USmedium than the USdecrease during acquisition: USincrease > USmedium > USdecrease | N/A | Linear mixed model (LMM) in a 3 (US type, categorical) × 3 (time, continuous) design with   1. US painfulness ratings, and 2. US (un‑)pleasantness ratings   as the outcome (main outcome: a). The influence of potentially confounding variables such as gender or age will be tested by adding them to the model one-by-one and testing model improvement. For further details, see 2.4. | N/A | Relevant effect: A statistically significant main effect of the factor *US type*, with post-hoc tests indicating that ratings are higher for the USincrease than the USmedium, and higher for the USmedium than the USdecrease during acquisition would suggest successful manipulation. | Manipulation check. Will be met for US painfulness ratings due to the exclusion criterion (USincrease > USmedium > USdecrease). We will additionally account for potential differences by adding the covariate to the model. |
| Manipulation check: Do both groups learn CS-US contingencies until the end of acquisition? | We do not expect to find group differences after the acquisition, since there will be no procedural differences until then. We expect higher estimations of coupling to the CSincrease than the CSmedium, and to the CSdecrease than to the CSmedium, respectively. | N/A | LMM in a 2 (instruction group, categorical) × 3 (CS type, categorical) × 2 (time, categorical) design with CS-US contingency ratings between phases as the outcome. Values will be scored in a way that positive values represent a high association with the correct US change, while zero represents no learnt change expectation and negative values learning of the incorrect association. | N/A | No statistically significant interactions between *instruction group* and *time*, and a significant interaction between *CS type* and *time*, with differences only after the acquisition training, would indicate a successful manipulation. | Manipulation check. Will test whether participants learned the contingencies. |
| Manipulation check: Did the US reliably evoke a response in autonomic physiological measures? | We expect differences based on US type, with larger reactions to both the USincrease and USdecrease than to the USmedium. | N/A | Analysis of variance (ANOVA) comparing reactions during acquisition training between CS types. | N/A | Relevant effect: Main effect of CS type, plus post-hoc test indicating that USincrease > USmedium > USdecrease for all unprocessed ratings and USincrease > USmedium < USdecrease for autonomic physiological responses. | Manipulation check. Will test whether US application reliably evoked unconditioned responses. |
| Do slopes of outcome measures differ between appetitive and aversive CS during acquisition?  Please note that CS valence will be shown in a separate row, since it is assessed in a different model due to fewer ratings. | H1: We expect faster acquisition of CS-US associations (i.e., US expectancy) and physiological responses (i.e., SCR and pupil dilation) to the CSincrease than the CSdecrease. | *N* = 150, or until a BF10 > 6 or BF10 < 1/6 is reached for all main hypotheses (H1-4a). | LMM in a 2 (CS type) × 5 (time, continuous) design with US expectancy ratings (H1a), SCR (H1b), and pupil dilation (H1c)  as the outcome (main outcome: a). Pupillometry will initially be analyzed with (i) an additional factor *bin* (three levels), and (ii) with only the selected bin. US painfulness ratings will be assessed as a potential covariate in all non-MRI analyses. | Our previous study found a steeper course of acquisition of CS valence for CSincrease compared to the CSdecrease (Δβ: 2.91 ± 1.14; t(224.80) = 2.56, p = 0.01, d = 0.34) in *N* = 36 participants (van der Schaaf & Schmidt et al., 2022). Decisions will be made based on statistical significance since also small effect sizes are of interest. Evidence will be evaluated based on how much the alternative hypothesis is favored over the null hypothesis (BF10). For details, see section 2.5.3. | A statistically significant interaction effect of *CS type* × *time* with steeper acquisition slopes of US expectancy for CSincrease than for CSdecrease would be in line with previous results (van der Schaaf & Schmidt et al., 2022). A non-significant interaction effect would mean that our previous findings could not be replicated. | The expected interaction may reflect a ‘better safe than sorry strategy’ (Solomon & Wynne, 1954), motivating participants to particularly anticipate the biologically more relevant event signaling potential harm. No difference might indicate that, while the aversive CS were emotionally more relevant (i.e., valence ratings), participants still learned the objective probability of each event. |
| Do appetitive and aversive CS differ in the slope of extinction in the uninstructed group? | H2: We expect faster extinction of CS-US associations and physiological responses to the CSdecrease compared to CSincrease. | N = 150, or until a BF10 > 6 or BF10 < 1/6 is reached for all pre-specified hypotheses. | LMM with data of the uninstructed group with a 2 (CS type) × 6 (time, continuous) design with US expectancy ratings (H2a), SCR (H2b), and pupil dilation (H2c)  as the outcome (main outcome: a). Pupillometry will be analyzed as described for H1. | Decisions will be made based on statistical significance since also small effect sizes are of interest. Evidence will be evaluated based on how much the alternative hypothesis is favored over the null hypothesis (BF10). For details, see section 2.5.3. | A statistically significant interaction effect of *CS type* × *time* with steeper extinction slopes of US expectancy for CSdecrease would extend the previous findings by van der Schaaf and Schmidt et al. (2022), who reported incomplete extinction of the CSincrease for CS valence. No difference between the slopes would indicate similar processes for either CS type with respect to US expectancy. | The expected interaction effect would suggest that the different biological relevance of the CS still affects learning processes of stimuli that are not presented any more. A non-significant interaction would indicate a general independence from biological relevance, also regarding US expectancy, or that extinction of either stimulus has the same relevance. |
| Does instructed extinction affect extinction slopes of appetitive and aversive CS, and does it affect them differently? | H3: Instructed extinction facilitates the extinction of CR.  H4: Instructed extinction facilitates the extinction of CR, especially pronounced for appetitive compared to aversive conditioning. | *N* = 150, i.e., 75 per group, or until a BF10 > 6 or BF10 < 1/6 is reached for all pre-specified hypotheses. | LMM in a 2 (instruction group) × 2 (CS type) × 6 (time, continuous) design with US expectancy ratings (H3a, H4a), SCR (H3b, H4b), and pupil dilation (H3c, H4c) as the outcome (main outcome: a). Pupillometry analyses will be carried out as explained for acquisition (H1). | Decisions will be made based on statistical significance since also small effect sizes are of interest. Evidence will be evaluated based on how much the alternative hypothesis is favored over the null hypothesis (BF10). For details, see section 2.5.3. | H3: A statistically significant *instruction group* × *time* interaction, with faster extinction in the instructed group would signal a successful modulation by instructions. No difference would indicate that verbal information does not affect the respective outcome measure.  H4: A statistically significant *instruction group* × *CS type* × *time* interaction, with particularly faster extinction of CR to the CSdecrease in the instructed group would imply the expected different susceptibility of the CS types to the instruction. No differences would indicate a similar effect. | H3+4: The expected interaction (H4) could reflect less learning of the absence of biologically more relevant or salient events in extinction training, even due to verbal information. No difference between appetitive and aversive extinction in the instructed group, but a significant interaction as expected in H3, might indicate a general underlying effect independent of biological relevance. Higher expectancy of the CS in the instructed group might reflect doubts regarding the believability of the instruction. No instruction effect at all would contrast the findings by Sevenster et al., 2012, who reported a large effect on US expectancy ratings. |
| Is CS valence of aversive vs. appetitive pain-related stimuli differently affected by experimental factors? | H1d: We expect a higher differential valence of the CSincrease than the CSdecrease after the acquisition training compared to the habituation phase.  H2d: In a subsample, i.e., the uninstructed group, we expect a higher differential valence of the CSincrease than the CSdecrease after the extinction training compared to after acquisition training.  H3d: We expect the instruction to facilitate the extinction of CR, i.e., we expect lower differential valences after the extinction training compared to after acquisition training in the instructed group.  H4d: We expect lower differential CS valence after the instruction (i.e., the ratings after extinction training compared to after acquisition training) in the instructed group, particularly for appetitive CS. | N = 150, or until a BF10 > 6 or BF10 < 1/6 is reached for all pre-specified hypotheses. | H1d: LMM in a 2 (CS type) × 2 (time, categorical) design with differential CS valence ratings during habituation and after acquisition as the outcome.  H2d: LMM with only the uninstructed group in a 2 (CS type) × 2 (time, categorical) design with differential CS valence ratings after acquisition and extinction as the outcome.  H3d: LMM in a 2 (instruction group) × 2 (time, categorical) design with differential CS valence ratings after acquisition and extinction as the outcome.  H4d: LMM in a 2 (instruction group) × 2 (CS type) × 2 (time, categorical) design with differential CS valence ratings after acquisition and extinction as the outcome. | Decisions will be made based on statistical significance since also small effect sizes are of interest. Evidence will be evaluated based on how much the alternative hypothesis is favored over the null hypothesis (BF10). For details, see section 2.5.3. | H1d: See H1(a-c).  H2d: See H2(a-c).  H3d: See H3(a-c).  H4d: See H4(a-c). | H1d: See H1.  H2d: See H2.  H3d: See H3.  H4d: The interaction could show that instructed extinction can in fact affect CS valence ratings, as opposed to the interpretation by Luck and Lipp (2016). |

Table 2. Study design table displaying research questions and corresponding hypotheses, analysis plan and interpretation for manuscript 2.

| **Question** | **Hypothesis** | **Sampling plan** | **Analysis Plan** | **Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis** | **Interpretation given different outcomes** | **Theory that could be shown wrong by the outcomes** |
| --- | --- | --- | --- | --- | --- | --- |
| Can acquisition and extinction of aversive and appetitive CR, represented by behavioral ratings, be related to resting state brain connectivity obtained prior to the behavioral task? | H1+H2: Behavioral indices of aversive, and appetitive acquisition and extinction correlate with resting state brain connectivity from the stated ROIs. | Acquisition max. *N* = 150, extinction max. *n* = 75. | An index of behavioral acquisition and extinction of aversive (H1) and appetitive (H2) CS-US associations, US expectancy, will be correlated with resting state-brain connectivity derived from ROI analyses. | Decisions will be made based on statistical significance since also small effect sizes are of interest. | If a statistically significant correlation will be found, individual levels of aversive and appetitive acquisition and extinction can be associated with differences in resting state connectivity of the expected ROIs. If no correlation will be found, this is not the case. | The expected association will expand previous findings by (Belleau et al., 2018), who reported an association between extinction learning and connectivity changes between a subregion of the amygdala and hippocampus. No association would imply that the relation cannot be found in solely pre-learning resting state data. |
| Can the effects of instructed extinction, as reflected by US expectancy ratings, be related to resting state brain connectivity? | Behavioral indices of the immediate effects of the instruction (H3) and instructed extinction slopes (H4) correlate with resting state brain connectivity from the stated ROIs specifically in the instructed group. | Max. *N* = 150. | An index of behavioral extinction of CS-US associations, derived from US expectancy ratings, i.e., (a) the immediate effect of the instruction, and (b) the interaction of the instruction with experience, will be correlated with resting state-brain connectivity derived from ROI analyses. | Decisions will be made based on statistical significance since also small effect sizes are of interest. | If a statistically significant interaction between instruction group and behavioral index (change in US expectancy) is found, with a different correlation between brain connectivity and the behavioral index in the instructed group compared to the uninstructed group, individual effectiveness of an instruction on extinction can be associated with differences in resting state connectivity of the expected ROIs. If no correlation will be found, this is not the case. | The expected association would be in line with previous findings using fMRI (Atlas et al., 2016). No association would suggest that the relation does not transfer to connectivity. |

1. **Method**

Here, we will employ a similar classical conditioning paradigm including a habituation, acquisition, extinction, and reinstatement phase previously used by van der Schaaf and Schmidt et al. (2022) but combine it with instructed extinction. We will apply three individually calibrated temperature stimuli to capsaicin pre-treated skin at the left volar forearm, which allows the induction of a medium tonic baseline pain sensation with safe low-level heat stimuli (van der Schaaf & Schmidt et al., 2022). Geometric figures will serve as cues predicting either a temperature increase (CSincrease), temperature decrease (CSdecrease) or no temperature change (CSmedium). Before the extinction phase, half of the participants will be instructed about contingencies in this phase, i.e., no further application of temperature changes, while the other half will not. We will collect expectancy, valence, and contingency ratings, as well as SCR and pupillometry during trials as outcome measures.

* 1. **Participants**

A minimum of *N* = 80 and maximum of *N* = 150 healthy individuals will be included in the study. Participants will be recruited through advertisements and existing participant lists. Recruitment will stop once the maximum number of participants has been reached, or the Bayes Factors (BF10) in favor of our (main) hypotheses (i.e., H1-4a) reach BF10>6 or BF10<1/6 (implying evidence for the alternative hypothesis, or the null hypothesis, respectively; see section 2.5.3. for details). We will test whether the stopping criterion has been reached after every tenth participant. Our sample size estimation is based on a previous study by Sevenster et al. (2012) that found a large effect of instructed extinction in an aversive conditioning paradigm (η2=0.41 for the *stimulus* × *trial* × *group* interaction). However, due to our more complex design featuring two types of CS+ and differences in the statistical analyses (e.g., focus on differences between the two CS+), we expect a slightly smaller effect in our study. The maximum sample size is further informed by recent studies using a similar design, which had shown medium to large effects (Schlitt et al., 2022; van der Schaaf et al., 2022). Based on prior experience with similar experiments, we expect a drop-out rate of about 15%.

Prior to the experimental sessions, participants will be screened for exclusion criteria. These comprise age under 18 or over 80 years, no fluency in German, left-handedness, BMI under 18 or over 30, chronic pain, severe diseases (e.g., cancer, migraine, epilepsy), or mental disorders (e.g., depression), skin diseases (e.g., neurodermatitis), or skin damage (e.g., sunburn, wounds), pregnancy or nursing, anisocoria, no normal vision or corrected-to-normal vision with glasses, contraindication to MRI scans, allergic responses to cayenne pepper, participation in a fear conditioning study, and participation in another study involving the use of pharmacological substances within the last three months. On the day of the experiment, participants will not be able to take part if they have an infection, are in pain, have engaged in professional sports, or consumed psychoactive substances (e.g., tobacco, caffeine, or taurine for three hours) or acute pain medication within the last 24 h. Participants will receive monetary compensation for their participation. The study was approved by the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen (16-7248-BO).

* 1. **Materials**
     1. ***Conditioned stimuli***

Conditioned stimuli will consist of three geometrical figures, i.e., a square, rhombus, and rectangle, presented in light blue color (RBG code: 142, 180, 227) with equal luminance against a black background. Stimulus size will be between 1.3° and 2.7° visual angle horizontally and 2° to 2.7° vertically. All CS will be presented centrally on the screen to minimize errors due to the pupil foreshortening effect (Petersch & Dierkes, 2022).

* + 1. ***Unconditioned stimuli***

During the experiment, a moderately painful heat stimulus (baseline temperature, USmedium) will be applied using a thermode (Model ATS, Pathway System, Medoc, Israel, https://www.medoc-web.com/pathway). The USincrease will be realized through a temperature increase relative to the baseline temperature. The USdecrease will consist of a temperature decrease relative to the baseline temperature. All thermal stimuli will be individually calibrated based on participants’ ratings of the painfulness of the thermal stimulation. The USmedium will be calibrated to a moderate pain intensity level (Visual Analogue Scale (VAS); VAS 40), the USincrease to a high intensity level (VAS 80), and the USdecrease to VAS 0.

* + 1. ***Outcome measures***
       1. **Behavioral measures**

The experiment will be controlled by a script using Psychophysics Toolbox Version 3 (PTB-3) in Matlab R2022a (The MathWorks, Inc., Natick, MA) and run in Ubuntu XX.XX on a XX PC (will be specified at Stage 2). All behavioral outcome measures will be obtained using VAS with questions presented in German. One-dimensional VAS ratings (e.g., of US painfulness), will be coded as 0 to 100; two-dimensional ratings, such as US (un-)pleasantness (very pleasant vs. very unpleasant) or CS valence (very pleasant vs. very unpleasant) will be coded with scales ranging from -100 to +100. Questions and verbal anchors for each scale are given below. For each rating, the initial cursor position will be randomly set between 0 and 75, or -75 and +75, respectively. Participants will provide their ratings using the buttons of a USB mouse. Pressing the left button will move the cursor to the left and pressing the right button will move it to the right. If ratings are not confirmed by pressing the middle button within the allotted time, the last cursor position will be logged, but the rating will be classified as invalid.

***US expectancy rating.*** Participants will rate their expectancy regarding the upcoming temperature development (“Which temperature development do you expect with which probability?”) based on the assumed probability of the expected event upon presentation of either of the CS. Ratings will be provided using a VAS ranging from *most probably cooling* to *most probably heating*, with *no change* (0) in the middle. Participants will be allowed a maximum of 7 s for their ratings.

***CS valence rating.*** CS valence ratings (“How do you perceive this geometric figure?”) will be obtained using a VAS ranging from *very pleasant* to *very unpleasant*, with *neutral* (0) in the middle. Participants may take up to 20 s to rate CS valence.

***CS-US contingency rating.*** Participants will rate how often a CS was followed by a temperature change (“How often was this geometric figure followed by a temperature change?”) on a scale ranging from *100% cooling* to *100% heating*, with *no change* (0) in the middle. Ratings will have to be provided within 20 s.

All abovementioned ratings will be converted into numerical scores between -100 and 100.

***US painfulness rating.*** The VAS for US painfulness (“How painful was this thermal stimulus?”) will range from *not painful at all* to *unbearably painful*. Participants will be allowed a maximum of 15 s to provide their ratings.

***US (un-)pleasantness rating.*** (Un-)pleasantness ratings of the US temperature stimuli (“How pleasant/unpleasant was this thermal stimulus?”) will be obtained on a VAS ranging from *very pleasant* to *very unpleasant*, with *neutral* in the middle. Participants will be allowed a maximum of 15 s to provide their ratings.

US related outcome measures will be converted to numerical ratings between 0 and 100 (painfulness) and -100 – 0 – 100 ((un-)pleasantness).

***Covariates.*** As pain-related cognitions and personality traits have been shown to influence learning (Nees & Becker, 2018), we will collect the following measures along with gender and age as potential covariates of interest.Immediately before the start of the actual experiment, participants will rate their arousal (“How tense do you feel?”) as well as their fear regarding the upcoming pain stimuli (“How anxious are you regarding the upcoming pain stimuli?”) on a VAS ranging from *not at all* *tense/anxious* to *very tense/anxious*, respectively. Ratings will be converted to numerical scores between 0 – 100. Furthermore, questionnaires will be used to explore the relationship between pain-related learning and pain-related cognitions as well as personality traits. Questionnaires include the State Trait Anxiety Depression Inventory (STADI; Laux et al., 2013; Renner et al., 2018), the short version of the Pain Sensitivity Questionnaire (PSQ-20; Fliege et al., 2005; Levenstein et al., 1993), the Anxiety Sensitivity Index-3 (ASI-3; Kemper et al., 2009, 2011; Taylor et al., 2007), the German Brief Experiential Avoidance Questionnaire (BEAQ; Gámez et al., 2011; Schaeuffele et al., 2022), the Pain Catastrophizing Scale (PCS; Meyer et al., 2008; Sullivan et al., 1995), and the Pain Anxiety Symptom Scale (PASS-20; Kreddig et al., 2015; McCracken & Dhingra, 2002).Each questionnaire will be analyzed according to its respective manual. Participants will be excluded from analyses based on the respective cut-off values of clinically relevant questionnaires.

* + - 1. **Autonomic physiological measures**

Skin conductance and pupil size will be acquired as physiological indices of cognitive and emotional processing.

***Pupillometry.*** The experiment will be conducted in a room with constant artificial light only to ensure that pupil constrictions and dilations solely occur in response to the experimental stimuli. The eye-tracking camera will be positioned 70 cm from the participant. Visual stimuli will be presented on a monitor positioned about 97 cm away from the participants. This distance will be maintained through the use of a chin rest, which will also reduce head movements.

An EyeLink 1000 Plus eye-tracking system in the desktop mount configuration (SR Research Ltd., Ottawa, Canada), will be used to collect pupillometry data on every trial. Data will be acquired from the right eye at a sampling rate of 500 Hz. The eye tracker will be calibrated to the right eye with a standard 9-point calibration and validation (displayed at 58% of the screen horizontally, and 53% of the screen vertically) to ensure adequate tracking in each participant with a maximum mean deviance of < .5° visual angle and a maximum value of < 1° visual angle. The eye-tracker will be recalibrated and validated before each experimental phase.

***Skin conductance responses.*** A BIOPAC MP150 system will be used to collect skin conductance continuously throughout the experiment with two disposable Ag/AgCl electrodes (0.8 cm diameter) and conductive electrode cream (SYNAPSE®; Kustomer Kinetics). Electrodes will be attached to the thenar and hypothenar eminence of the left hand. The sampling rate of the system will be set to 2kHz.

* + - 1. **MRI data acquisition**

Anatomical scans, resting-state fMRI, and diffusion tensor imaging (DTI) measurements will be acquired with a 3T MR system (MAGNETOM Vida, Siemens Healthineers AG, Erlangen, Germany) and a 64-channel head coil.

***Imaging parameters.*** Structural images will be acquired with a 3D-MRPage T1-weighted sequence (repetition time (TR): 2530 ms, echo time (TE): 2.34 ms, flip angle: 7°). Blood oxygen level-dependent (BOLD) contrast images will be acquired using echo-planar imaging (EPI) with transversal slices encompassing the whole brain (56 slices, resolution 2.5 × 2.5 × 2.5 mm, TE 32.0 ms, TR 616 ms, Simultaneous Multi-Slice (SMS) 8x, 1168 volumes, 11 min 59 s). Multi-shell diffusion-weighted imaging (DWI) will be performed with a resolution of 2 x 2 x 2 mm (133 directions, b-values: 0, 1000, 1800, 2500 s/mm2, TE 103 ms, TR 3500 ms, flip angle: 52°, 64 slices, full brain coverage).

* 1. **Design**

The study employs a mixed three-factorial design with one between-subjects factor, *instruction group* (instructed or uninstructed), and two within-subjects factors, *CS type* (CSincrease, CSdecrease, CSmedium) and *time* (see below for details). The dependent variables are US expectancy, CS valence, CS-US contingency, US painfulness and US (un‑)pleasantness ratings, SCR, and pupil size.

Group allocation and further experimental parameters will be (pseudo-)randomized, taking instruction group, allocation of the geometric figures to the CS types, order of US presentation during the temperature calibration, and order of US presentation during acquisition training and reinstatement manipulation into account.

* 1. **Procedure**

The experiment will comprise two sessions. On the first day, participants will complete the brain imaging session. On the second day, they will take part in the behavioral experiment including autonomic measurements.

* + 1. ***Imaging session***

The imaging session will include anatomical, rsfMRI and DTI scans. After providing written informed consent, participants will be screened for MRI exclusion criteria and eligible participants will be asked for buccal swabs which will – together with the DTI scans – be analyzed as part of a larger research initiative (SFB 1280). Participants will subsequently be positioned in the scanner. For the acquisition of rsfMRI data, participants will be instructed to lie still in the scanner with their eyes open for 12 minutes, without any further task. For the anatomical (6 minutes) and DTI (8 minutes) scan, participants will again lie still in the scanner, but may close their eyes.

* + 1. ***Behavioral session***

On the day of the main experiment, the experimenter will apply capsaicin cream (1%, 8-methyl-N-vanillyl-trans-6-nonenamide, 98%, Sigma, diluted in 5% ethanol-KY jelly) to a 3 x 3 cm site at the left volar forearm. Capsaicin sensitizes the skin and thus allows the induction of a medium-level tonic baseline pain with safe low-level heat application (Leknes et al., 2008; Petersen & Rowbotham, 1999; van der Schaaf and Schmidt et al., 2022). While the cream is taking effect, the site will be covered with a patch and participants will complete questionnaires on psychological processing (see above, in 2.2.3.1. covariates), demographic information, pain, and handedness (Edinburgh Handedness Inventory; Oldfield, 2013). After 45 min, the remaining cream will be removed and the thermode will be attached to this site.

A temperature calibration procedure consisting of three phases will be performed to adjust the temperature levels of the thermode to the individual pain sensitivity (previously described in van der Schaaf and Schmidt et al., 2022). In the first phase, participants will be gradually familiarized with a constant medium level painful stimulation. This phase will also be used to determine the range of temperature levels in the second phase, which allows a regression model-based estimation of the final three temperature levels used in the main experiment (i.e., the temperatures used to induce i) a constant, medium level painful stimulation, ii) pain exacerbation, iii) pain relief). In a third and final phase, these stimuli will be validated in a procedure that mimics the thermal stimulation during the main experiment.

In the first phase, a procedure using gradually increasing temperature levels will be carried out. Participants will first rate the painfulness of a constant 28°C stimulus (most probably not painful) as a baseline measure. Two seconds after the rating, the temperature will be increased by one step. The step size is defined as follows: from 28-42°C in steps of 2°C and from 42-47°C in steps of 1°C. After stimulus presentation for 9 s, the next rating will follow. If this rating is between VAS 40-60, a further rating of the same temperature stimulus will be obtained after 5 s, before the next trial commences. The procedure ends once the maximum temperature (i.e., 47°C) has been reached or the participant’s ratings reach a pain intensity of VAS 60. This procedure will be carried out twice, and the resulting temperature level rated as ~VAS 50 (x) in the second round will be used to choose an adequate temperature range for the following calibration phase.

In the second phase, 20 separate thermal stimuli will be applied. Specifically, temperature levels ranging from x -1.5 to x + 3.0°C in steps of 0.5°C around the temperature (x) from the first phase will be used. Each temperature level will be presented twice, resulting in 10 different temperature levels applied in pseudorandomized order and with a stimulus duration of 8 s (time before temperature onset 5 s, baseline temperature between stimuli = 26°C). Pain intensity ratings will follow 12 s after the onset of the heat stimulus, before the next trial starts. Based on the ratings for each of these heat stimuli, three temperature levels will be chosen as the USmedium, USincrease, and USdecrease using a linear regression analysis.

In the third calibration phase, these temperature levels will be validated, first only using the baseline temperature (VAS 40) to assess habituation by obtaining four ratings of US painfulness and (un‑)pleasantness, and second including USincrease (i.e., VAS 80) and USdecrease stimuli (i.e., VAS 0, calculated as USmedium minus 10°C, minimum 20°C). The US will be presented for 8 s, as in the main experiment. One of six pre-defined stimulation orders, including the presentation of five USmedium, and three USincrease and USdecrease each, will be used for the second validation test. Here, the temperature level will change after 4-7 s if required by the respective pre-defined order. Another 12 s later, participants will rate US painfulness and (un‑)pleasantness before the next trial starts. If participants report an insufficient pain intensity level (mean US painfulness rating of the USmedium < 10) within the predefined safety limits (44°C for the USmedium, 46°C for the USincrease), or rate stimuli inconsistently (criterion: mean USincrease > USmedium > USdecrease), the calibration procedure will be repeated. If repetition also does not lead to the fulfillment of these criteria, they will be excluded from further participation.

Following this calibration procedure, participants will complete the differential conditioning experiment. In preparation, electrodes to measure skin conductance will be attached, participants will provide their arousal and pain-related fear ratings, and a calibration and validation of the eye-tracker will be carried out.

***2.4.3. Differential conditioning paradigm***

**2.4.3.1. Trial structure**

An example trial including US expectancy is shown in Fig. 1 (right panel). After the presentation of a fixation cross for 4-7 s, which will serve as the intertrial interval (ITI) and as a baseline period for subsequent changes in pupil size, the trial will start. First, the CS is presented for 9 s. If participants are asked to rate US expectancy in the respective trial, the VAS is presented upon CS presentation onset and remains visible until a response has been provided, or the time limit has been reached (7 s). After 7.5 s of CS presentation, the USincrease/decrease is presented in reinforced trials, while the temperature level remains unaltered in unreinforced trials. The US is applied for 8 s and 1.5 s after US onset, the CS is replaced by the fixation cross which remains on the screen until the end of the trial. Each trial ends with a 4 s delay before the next trial starts, or the rating scales for US painfulness and (un‑)pleasantness are displayed on the screen.

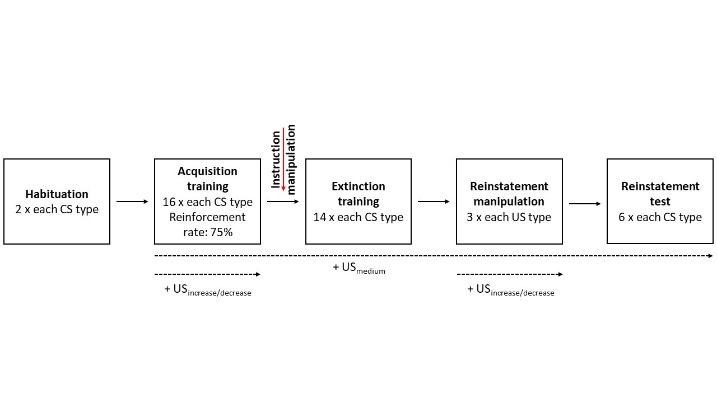
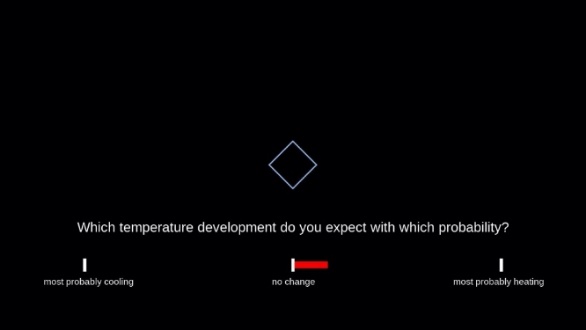
 

Fig. 1. **Left panel:** Differential conditioning paradigm. A pain intensity of ~ VAS 40 will be induced by applying thermal heat stimuli to capsaicin-pretreated skin (ongoing pain, USmedium). An increase and decrease in temperature level will lead to pain exacerbation (~ VAS 80, USincrease) and pain relief (~ VAS 0, USdecrease), respectively. The design consists of habituation, acquisition, extinction training, and reinstatement manipulation and test. The instruction manipulation depending on group allocation will be implemented before the extinction training. **Right panel:** US expectancy rating. The trial structure including US expectancy and US painfulness and (un‑)pleasantness ratings is as follows: After 4-7 s, where a fixation cross is presented, the CS will be shown alongside the US expectancy question which will be displayed until a rating is provided, or when the maximum time of 7 s has been reached. 7.5 s after CS+ onset, a temperature change (USincrease or USdecrease) will occur for 8 s in 75% of the trials. The CS will be presented for 9 s, after which the fixation cross will reappear on the screen. 12 s after US onset, the US painfulness question will be presented for up to 15 s or until a rating has been provided. After 0.5 s, the US (un-)pleasantness question will follow.

**2.4.3.2. Structure of the main experiment**

The experiment will consist of a habituation, acquisition, extinction, and reinstatement phase (Fig. 1, left panel). In the habituation phase, participants will only provide CS valence ratings, whereas both CS valence and CS-US contingency ratings will be obtained following the acquisition, extinction, and reinstatement phase.

***Habituation phase.*** In the habituation phase, each CS is presented twice. At the first presentation, participants will rate CS valence. This rating will serve as a baseline measure. CS are never reinforced during the habituation phase.

***Acquisition training.*** Before this phase, participants will be instructed that geometric stimuli could be followed by either no change in temperature, a temperature increase or a temperature decrease. They will, however, not be informed about the exact contingencies, directions of change in temperature, or CS-US associations. The temperature level will be increased to the individual baseline temperature (USmedium). The acquisition training consists of 16 presentations of each CS, resulting in 48 trials. During this phase, one CS is paired with a temperature increase (CSincrease), a second is paired with a temperature decrease (CSdecrease) and a third is paired with no change in temperature (CSmedium). The CSincrease/decrease will be reinforced in 75% of the trials. The trial order is pseudorandomized using six predefined orders. In each of them, the first and last CSincrease/decrease presentations will always be reinforced. Furthermore, each CS is presented a maximum number of three times in a row. The number of CS presentations of each type is the same within both halves of the acquisition training. Participants will rate their US expectancy during the first and every fourth presentation of each CS type, respectively. Furthermore, they will rate US painfulness and (un-)pleasantness after every fourth reinforced trial of the same type. Ratings are also obtained from CSmedium trials following the same rationale.

***Instruction.*** Before the extinction training, participants will be told that the position of the electrodes for SCR recordings and the thermode has to be reassessed. The thermode will be removed and the electrodes will be inspected visually. The thermode, and if necessary, the electrode will be reattached. Subsequently, half of the participants (i.e., the instructed group) will be informed that the geometric figures will no longer be accompanied by temperature changes. The uninstructed group will receive no instruction. Both groups will be told that, when the experiment continues, they will again be asked to rate their expectancy of temperature changes, as well as their painfulness and (un-)pleasantness. Furthermore, they will be reminded to provide their ratings as swiftly as possible. The thermode will then be reapplied and the experiment will continue.

***Extinction training.*** The extinction training consists of 14 presentations of each CS (42 trials in total). The trial order will be randomized with the same number of presentations of each CS within each half of the phase. CS will never be reinforced. Expectancy ratings will be obtained during the first presentation of each CS type, and subsequently on every third presentation of the respective type. US painfulness and US (un-)pleasantness will be rated five times, evenly distributed across the phase. These ratings will follow each CS type at least once.

***Reinstatement phase.*** The reinstatement phase will commence with the unannounced presentation of the US (reinstatement manipulation). Each US is presented three times in a pseudorandomized order, again according to six pre-defined orders. Each presentation is followed by ratings of US painfulness and US (un-)pleasantness.

In the subsequent reinstatement test, each CS is presented six times in a randomized order, resulting in a total of 18 trials. CS will never be reinforced. Participants will rate US expectancy during the first, third, and sixth presentation of each CS. US painfulness and US (un-)pleasantness will be rated four times in total, distributed evenly across the phase. These ratings will follow each CS type at least once.

* 1. **Analysis plan**
     1. **Exclusion criteria**

US painfulness ratings of the acquisition training phase will be assessed for each participant. All participants with ratings that deviate from the envisaged order (i.e., acquisition mean USincrease > USmedium > USdecrease) will be excluded from analyses. Furthermore, participants will be excluded based on the respective cut-off values of clinically relevant questionnaires as specified in section 2.2.3.1. covariates. Exclusion criteria based on specific outcome measures are specified in the respective sections.

* + 1. **Frequentist analyses**

Analyses will be carried out using R X.X.X (will be specified at Stage 2; R Core Team, 2021). The R package lme4 will be used to calculate linear mixed models (LMM; Bates et al., 2014). First, the adequate model type will be chosen based on the effect of interindividual variability. This will be tested by comparing a fixed effects linear regression model to a linear mixed model including random intercepts for participants. Models will be estimated according to the restricted maximum likelihood (REML) approach. If the individual factor significantly improves model fit according to likelihood ratio tests and the Akaike information criterion (AIC) for model comparison, the LMM will be selected. The significance of individual fixed effects of the factors of interest will be reported as results of a Wald chi-square tests of model results of the winning model. Contributions of the different levels of the included terms will be assessed using post-hoc tests. Thresholds for hypothesis tests will be set to α ≤ .02.

* + - 1. ***Behavioral measures***

***US expectancy ratings.*** US expectancy ratings, ranging from -100 to indicate the expectation that the temperature level will most probably decrease, to 0 (no change), and +100 (most probably temperature increase), will be recoded for CSdecrease trials (value x (-1)). Only ratings of CSincrease/decrease trials will be included in the analyses. Positive values therefore indicate correct learning of the respective associations.

Analyses will be performed to assess changes in ratings within acquisition training, extinction training, and reinstatement test, respectively. The last rating during acquisition training is included as a baseline rating in the analysis of the extinction training. *CS type*, *instruction group* (except in the analysis of the acquisition training), and *time*, as well as their interaction terms, will be added as fixed effects explanatory variables to the analyses. Time, i.e., the rating number of the respective CS type, will be defined as a continuous variable to allow for the identification of general effects over time. US expectancy ratings will be obtained as the dependent variable. Differences between the instruction groups, resulting in a three-way interaction of *instruction group* × *CS type* × *time*, are of particular interest.

Furthermore, US painfulness ratings, gender, age, arousal, and pain-related fear as well as anxiety, depression, and pain catastrophizing as assessed by the respective questionnaires will be added as potential covariates in an exploratory manner and tested for model improvement using likelihood ratio tests and the AIC for model comparison. In case of a significant effect of covariates, both the model with and without the respective covariates will be presented. Hypotheses will be tested based on the model without covariates and the validity of the interpretation will be discussed based on the model including covariates.

H1a: To test the first hypothesis, we will assess the statistical significance of the interaction *CS type* × *time* during acquisition training, as described in 2.4.1. We expect a steeper slope for the CSincrease than the CSdecrease.

H2a: We will assess the model with factors *CS type* and *time*, as well as their interaction, in the uninstructed group during extinction. A statistically significant interaction of *CS type* × *time*, and a steeper extinction slope for the CSdecrease compared to the CSincrease would confirm the hypothesis.

H3a: The third hypothesis concerns the instruction effect and will be examined with the interaction *instruction group* × *time*. We expect a steeper slope in the instructed than the uninstructed group.

H4a: We will test whether the instruction differentially affects extinction slopes. We expect an *instruction group* × *CS type* × *time* interaction, with steeper slopes particularly for the CSdecrease compared to the CSincrease in the instructed group.

***CS valence ratings.*** LMM analyses will be performed to assess the effect of conditioning on changes in differential CS valence ratings over the course of the experiment. *CS type*, *instruction group*, and *phase*, as well as their interactions, will be added as fixed effects, differential CS valence ratings (CSincrease – CSmedium, CSmedium – CSdecrease) as the dependent variable. *Phase* will be entered as a categorical variable. We will explore potential effects of the covariates as specified in the analysis section of US expectancy. Additionally, CS-US contingency ratings will be tested as a potential covariate.

H1d: A statistically significant interaction between *CS type*and *time* will support our assumption of a steeper slope for the CSincrease than the CSdecrease in the acquisition training.

H2d: Analyzing the uninstructed group, a statistically significant interaction between *CS type*and *time* will support our assumption of a steeper extinction slope for the CSdecrease compared to the CSincrease.

H3d: We will test this hypothesis using the interaction *instruction group* × *time*. We expect steeper slopes in the instructed group.

H4d: We will assess this hypothesis by testing the statistical significance of the *instruction group* × *CS type* × *time* interaction. We expect the instruction to facilitate the extinction particularly for the CSdecrease.

***CS-US contingency ratings.*** Analyses will assess differences in learning of the contingencies over the course of the experiment and between CS types and instruction groups. CS-US contingency ratings will be analyzed as described for CS valence ratings. CS-US contingency ratings, with ratings of the CSdecrease multiplied by minus one as explained for US expectancy, will constitute the dependent variable. Thus, positive values will represent correct learning of CS-US contingency.

***US painfulness and (un-)pleasantness ratings.*** The analyses will describe potential differences in experienced pain intensity between both instruction groups throughout the experiment and in association with the CS types, to ensure that the temperature calibration was successful, and no group differences occurred. US painfulness, or US (un-)pleasantness ratings, respectively, will be the dependent variable. First, the differential perception of the three temperatures (USincrease, USmedium, USdecrease) will be compared. To assess stability over time, the model will include all ratings from the acquisition training. Since there is no difference between instruction groups in this phase, *group* will not be included as a factor. *US type*, and *time* are included as fixed effects. Second, the perception of USmedium will be compared between the phases and between instruction groups. Thus, the model will include *time* and *instruction group* as fixed effects. For both models, effects of the described covariates will be assessed, except for US painfulness when it is the dependent variable.

* + - 1. **Pupillometry data**

Pre-processing and analysis of pupillometry data will be performed using R X.X.X (will be specified at Stage 2). Recorded data will be down-sampled to 100 Hz and smoothed with a low-pass filter at 5 Hz. Pupil size will be converted from arbitrary units to millimeters using the method reported by Hayes and Petrov (2016). Missing data (e.g., due to blinks) will be interpolated using a linear approximation. Pupil size will be normalized by subtracting pupil sizes after CS onset from a pre-stimulus baseline (i.e., the mean pupil size in the 1000 ms prior to CS onset), to ensure that random fluctuations in pupil size over time do not affect results (Mathôt et al., 2018). Furthermore, this approach accounts for a decrease in tonic baseline pupil size, which may occur over the course of an experiment (Leuchs et al., 2017). Trials with more than 50% missing data during the relevant time frame (1000 ms pre-stimulus baseline, and CS presentation before US onset) will be excluded. Furthermore, trials with sudden shifts in pupil size will be excluded. We will identify these using a sliding window approach, in which we calculate the standard deviation of the pupil size in windows of 100 ms ranging from 1 s before CS onset until 7.5 s after. If any SD exceeds 0.2, the respective trial will be discarded. If more than 50% of a participant’s trials for either the acquisition, extinction training, or reinstatement test has to be discarded, the participant’s data of the entire respective phase will be excluded from the pupillometry analyses.

We will calculate the average pupil width from the normalized pupil size for three intervals between the end of the initial dip due to CS presentation and the onset of the US (1.5-3.5, 3.5-5.5, 5.5-7.5 s). Because we have no *a priori* hypothesis as to when the conditioned pupil response will occur, we decided to divide the time window from the initial light response to the CS presentation to the onset of the US into three segments to allow for the analysis of differences in the development of the pupil responses over time.

All trials with ratings of US expectancy will be excluded from the analysis to avoid effects induced by rating-related movements, cognitive processes, and different lighting conditions. Since the ratings are obtained on every fourth or third trial of each CS type in the acquisition and extinction training, respectively, values of the remaining pupil size reactions to trials between two rating trials will be averaged. Ratings are obtained in the first, third, and sixth trial of the reinstatement test, therefore, the fourth and fifth trial will be averaged.

Analyses will be performed separately for acquisition training, extinction training and reinstatement test. As for US expectancy, the factors *CS type*, *instruction group* (except in the acquisition training analysis), *time*, and *time bin*, as well as their interaction terms, will be included in the models as fixed effects. For the acquisition model, also the habituation phase will be included, and for the extinction model, the last set of averaged trials from the acquisition training will be included. Average pupil width from the respective time frames will be recorded as outcome variable. Furthermore, we will report results from a model focusing on only the bin indicating the largest differentiation. We will apply Bonferroni-Holm correction to account for multiple comparisons within the respective experimental phases. Hypotheses will be tested using the model with one selected bin as described for US expectancy ratings. Finally, effects of the stated covariates will be assessed.

* + - 1. **SCR data**

Data will be down-sampled to 20 Hz and low-pass filtered at 2 Hz. Local maxima and preceding minima will be identified within the data. Amplitudes will then be calculated by subtracting the minimum value from the subsequent maximum value. The response windows, from which the maximal amplitude will be extracted, will be set to 1 to 4 s after CS onset for the first interval response (FIR), and to 4 to 7.5 s for the second interval response (SIR). A more pronounced FIR to CS during early acquisition, and increased SIR later within acquisition training has been observed by Jentsch et al. (2020). In previous studies, the FIR has been associated with orienting behavior (i.e., responding to new circumstances and habituation over time), while the SIR is understood to reflect US expectancy (Öhman, 1972, 1974; Wolter & Lachnit, 1993). Therefore, the data analysis will distinguish between FIR and SIR. For trials with a resulting amplitude <0.01 μS or without a local minimum, amplitude will be set to 0. Finally, SCRs will be log transformed to reduce skew in the data (ln(x+1)).

As with pupillometry data, all trials containing a rating of US expectancy will be excluded from further analyses, and the remaining two or three trials between successive ratings will be averaged. Furthermore, inferential analyses will be carried out as described for pupillometry data. Hypotheses will be tested as described for US expectancy ratings, separately for the outcomes FIR and SIR, and effects of the stated covariates will be assessed.

* + 1. **Bayesian inference**

In addition to LMM, we will use Bayesian hypothesis testing to quantify the evidence for or against the stated hypotheses (BF of the alternative hypothesis over the null hypothesis, BF10). Bayes factors will be estimated based on the models reported in the frequentist analyses with an approximation for BF using the Bayesian Information Criterion (BIC) (Wagenmakers, 2007). To interpret the strength of evidence given by BF10 for each hypothesis, we will use Jeffrey’s Scale of Evidence (Jeffreys, 1998), which considers a BF10 > 100 to be extreme evidence for H1, BF10 = 30-100 very strong evidence for H1, BF10 = 10-30 strong evidence for H1, BF10 = 3-10 moderate evidence for H1, BF10 = 1-3 anecdotal evidence for H1, BF10 = 1/3-1 anecdotal evidence for H0, BF10 = 1/10-1/3 moderate evidence for H0, BF10 = 1/30-1/10 strong evidence for H0, BF10 = 1/100-1/30 very strong evidence for H0, and BF10 < 1/100 as extreme evidence for H0.

**2.5.4. Neural measures and correlation with behavior**

**2.5.4.1. Preprocessing and first-level analyses**

We will perform SBFC analyses to test our hypotheses. For preprocessing, we will use the fMRIPrep preprocessing pipeline X.X.X (version will be specified at Stage 2; Esteban et al., 2019, 2020). The resulting preprocessed functional and anatomical volumes in Montreal Neurological Institute (MNI) space, segmentation masks, potential confounder variables, and realignment, scrubbing, and QC\_timeseries covariates will be imported into the CONN toolbox (Whitfield-Gabrieli & Nieto-Castanon, 2012). In CONN, a Gaussian kernel of 8 mm full-width half-maximum (FWHM) will be applied for spatial smoothing. In a nuisance regression for denoising, we will perform an anatomical component-based noise correction method correction (aCompCor) (Behzadi et al., 2007), which provides a principal component analysis estimating the individual differentiation of physiological noise, white matter and cerebrospinal fluid (CSF), extracting five components each, motion regression based on the preprocessing estimates provided from fMRIPrep, and bandpass filtering threshold at 0.1 Hz and 0.01 Hz. We will censor time frames with high motion (“scrubbing”, framewise displacement > 0.5 mm) and exclude participants if more than 25% timeframes were censored. Participants who report having fallen asleep during data acquisition will be excluded from further analyses.

SBFC analyses will use the left and right dlPFC and vmPFC, amygdala, and striatum as seeds for first-level analyses in the CONN toolbox. Amygdala and striatum masks will be derived from the Harvard-Oxford Atlas (Desikan et al., 2006), with the striatal masks constructed by combining the caudate, putamen, and nucleus accumbens. Separate masks for the dorsal and ventral striatum will be created by a split at MNI coordinate *z* = 0, according to Cooper et al. (2012). Subsequent analyses will be performed for the ventral and dorsal striatum separately. Masks for the vmPFC and dlPFC will be obtained from a whole brain functional parcellation (https://identifiers.org/neurovault.collection:2099) created with Neurosynth’s meta-analysis tool (Yarkoni et al., 2011). We will use the rsfMRI scan to derive correlation maps between the respective seeds and all other brain voxels using Pearson’s correlations and no weighting. Fisher’s z-transformation will be used to normalize individual *r* statistics (resampled to 1 mm3 voxels, original: 2.5 × 2.5 × 2.5 mm).

**2.5.4.2. Second-level analyses: Acquisition and extinction training**

In the second-level analyses, extracted connectivity of vmPFC, amygdala, and striatum will be correlated with the respective observed index of acquisition. This index will be defined as the change in US expectancy of the CSincrease (aversive) from the beginning to the end of acquisition training, represented by the slope of ratings in this phase (Hab-Acq5), and accordingly for the CSdecrease (appetitive). Further, data from extinction learning from only the uninstructed group will be analyzed accordingly, i.e., from the end of the acquisition training to the end of extinction training, represented by the slope of ratings in these phases (Acq5-Ext5). The respective indices will be used as the predictor in a general linear model (GLM) framework, with connectivity as the outcome. Resulting statistical maps will be converted to Z-scores and processed with probabilistic threshold-free cluster enhancement (pTFCE, v2.2.1, Spisák et al., 2019), a topology-based method to enhance the detectability of neuroimaging signal to boost statistical belief in spatially extended clusters of connectivity differences. Enhanced pseudo z-score images will be thresholded to maintain a false discovery rate (FDR) of *q* < .02.

Hypotheses: We will assess the statistical significance of the correlation between the respective connectivity and behavioral indices, i.e., for H1 with the change in US expectancy for the CSincrease during acquisition and extinction, and for H2 with the respective index for the CSdecrease.

**2.5.4.3. Second-level analyses: Instruction effect on extinction learning**

To assess the association between the connectivity of pain- and relief-networks and individual effects of instructions on extinction efficacy, scores representing the individual instruction effect on extinction in the instructed group are calculated. We will use two indices, which will both be calculated for both groups to assess differences in the associations between the groups. The first index will focus on the direct effect of the instruction by comparing US expectancy immediately before and after the instruction (CSinstructed.Ext1 – CSinstructed.Acq5) for the CSincrease and CSdecrease separately. The second will be used to study the interaction between instruction and experience. To this end, extinction slopes of all individuals are obtained from expectancy ratings from the final rating during acquisition training until the final rating during extinction (i.e., CSdecrease value x (-1)). Finally, individual effect scores will be entered alongside instruction group as factors into one-way analysis of covariance (ANCOVA) covariate interaction models with connectivity strength of dlPFC, vmPFC, and striatum as derived from first-level SBFC analyses as the outcome. Resulting statistical maps will be converted to Z-scores and processed with probabilistic threshold-free cluster enhancement (pTFCE, v2.2.1, Spisák et al., 2019). Enhanced pseudo z-score images will be thresholded to maintain a false discovery rate (FDR) of q < .02.

Hypotheses. We will assess the statistical significance of the interaction *instruction group* × *behavioral index* for both CS types, and both behavioral indices, i.e., for H3 immediate changes in US expectancy, H4 the index for the interactive effect of the instruction with experience. The expected effect would be represented by the interaction of group and behavioral index, indicating differences in the outcome for only the instructed group.

1. **Pilot study**

To inform the strategic planning of this study, we conducted a pilot study to compare the effect of our differential conditioning procedure and the instruction on the outcomes CS valence and US expectancy ratings. While CS valence ratings were used in the original design by van der Schaaf and Schmidt et al. (2022), instructed extinction had no significant effect on valence ratings (Luck & Lipp, 2016) and expectancy might be the more susceptible outcome to investigate instruction effects. Furthermore, we intended to refine our analysis plan for the pupillometry data and develop clear exclusion criteria based on the temperature calibration procedure, as well as the perception of US painfulness during the experiment. Since these criteria were only developed based on the pilot data, exclusion criteria were not yet fully applied as described.

* 1. ***Method***

*N* = 42 participants (*n* = 12 male, *n* = 29 female, *n* = 1 no entry, age = 24.24 [SD = 3.93]) were recruited for the pilot study. We excluded *n* = 4 participants during the calibration procedure due to insufficiently high levels of pain, *n*= 5 participants were excluded due to uncharacteristic perception of the US during the experiment (very low US painfulness ratings for the USmedium, or very inconsistent ratings of US painfulness of the 3 US types), *n*= 2 due to an uncharacteristic perception of the US as indexed in CS ratings (i.e., extreme outliers), *n*= 2 participants due to technical issues, and *n* = 1 participant due to an interruption during the procedure of the main experiment. Pupillometry data from *n* = 2 participants was missing due to technical issues, as well as all data from the reinstatement phase of *n* = 1 participant. We thus excluded data from *n* = 1 participant from pupillometry analyses due to missing data in the reinstatement phase.

The pilot study was carried out using the methods described for the main study, with a few adaptations. We included four groups and two between-subjects factors (instruction group: instructed vs uninstructed; rating group: valence vs expectancy) in a fully factorial design. In the valence groups, CS valence was rated during experimental phases, and US expectancy was rated before and between phases. The expectancy groups were run as described in the methods section of the planned study. The pilot study only included behavioral and autonomic physiological measures and therefore required only one session per participant. Since no MRI data was acquired, MRI-related exclusion criteria were not applied.

Due to the rather small sample size (*n*= 6-8 per group), we only report results graphically. Furthermore, we focus on the outcomes the pilot study was conducted for, i.e., the main behavioral outcome and the methodology for the pupillometry analyses.

* 1. ***Results***

**3.2.1. Behavioral ratings.**

Results from the expectancy groups are shown in Fig. 2. As expected, most participants learned to predict USincrease/decrease based on CS presentation. Overall, learned US expectancy seemed to decrease faster at the start of the extinction training in the instructed than the uninstructed group, indicating successful expectancy manipulation.

Results from the valence groups are depicted in Fig. 3. Unfortunately, participants did not consistently rate the CSdecrease as pleasant. For the CSincrease, no clear differences in ratings during the extinction training can be observed between instruction groups, although the instructed group seemed to have provided slightly lower mean ratings. It has to be pointed out that due to the low sample size, outliers might have significantly biased these observations.

Graphical user interface, chart

Description automatically generated

Fig. 2. US expectancy ratings (raw value) during the acquisition (Acq1–Acq5) (left panel), extinction (Ext1–Ext5), and reinstatement phases (Rein1–Rein3) (right panel). Extinction and reinstatement phase ratings are provided separately for the instructed (top) and the uninstructed (bottom) group. Ratings are given as means and 95% confidence interval. Single data points in dot shape. Dashed line separates the phases. Data are provided for N = 15 participants.

Chart

Description automatically generated

Fig. 3. CS valence ratings (raw value) during the acquisition (Acq1–Acq5) (left panel), extinction (Ext1–Ext5), and reinstatement phases (Rein1–Rein3) (right panel). Extinction and reinstatement phase ratings are provided separately for the instructed (top) and the uninstructed (bottom) group. Ratings are given as means and 95% confidence interval. Single data points in dot shape. Dashed line separates the phases. Data are provided for N = 13 participants.

Based on these observations, we decided to use US expectancy as the primary outcome rating in the main study, while CS valence ratings will be acquired during habituation and after acquisition and extinction training, and the reinstatement phase. Values of the main outcome during extinction training, the phase on which our main hypotheses are focused, are shown in Fig. 4. Of note, the mean expectancy for the CSdecrease returned to zero, whereas the mean expectancy for the CSincrease remained elevated.

Chart, line chart

Description automatically generated

Fig. 4. US expectancy ratings for the CSincrease and CSdecrease (CSdecrease x (-1)) show correct expectations for both events with positive values in the last trial of the acquisition (Acq5) and during extinction training (Ext1–Ext5). Ratings are given as means and 95% confidence interval. Single data points in dot shape. Data are provided for N = 15 participants. Data is presented in linear regression lines for the instructed (full) and uninstructed (dotted) group, respectively.

**Pupillometry.** The mean differential pupil response of all participants over all trials of the experiment is shown in Fig. 5. After the initial light response, a pupil dilation compared to baseline started 1.5 s after CS onset. Based on these findings, we chose to analyze data from 1.5 to 7.5 s after CS onset, split into three bins, in the main experiment.

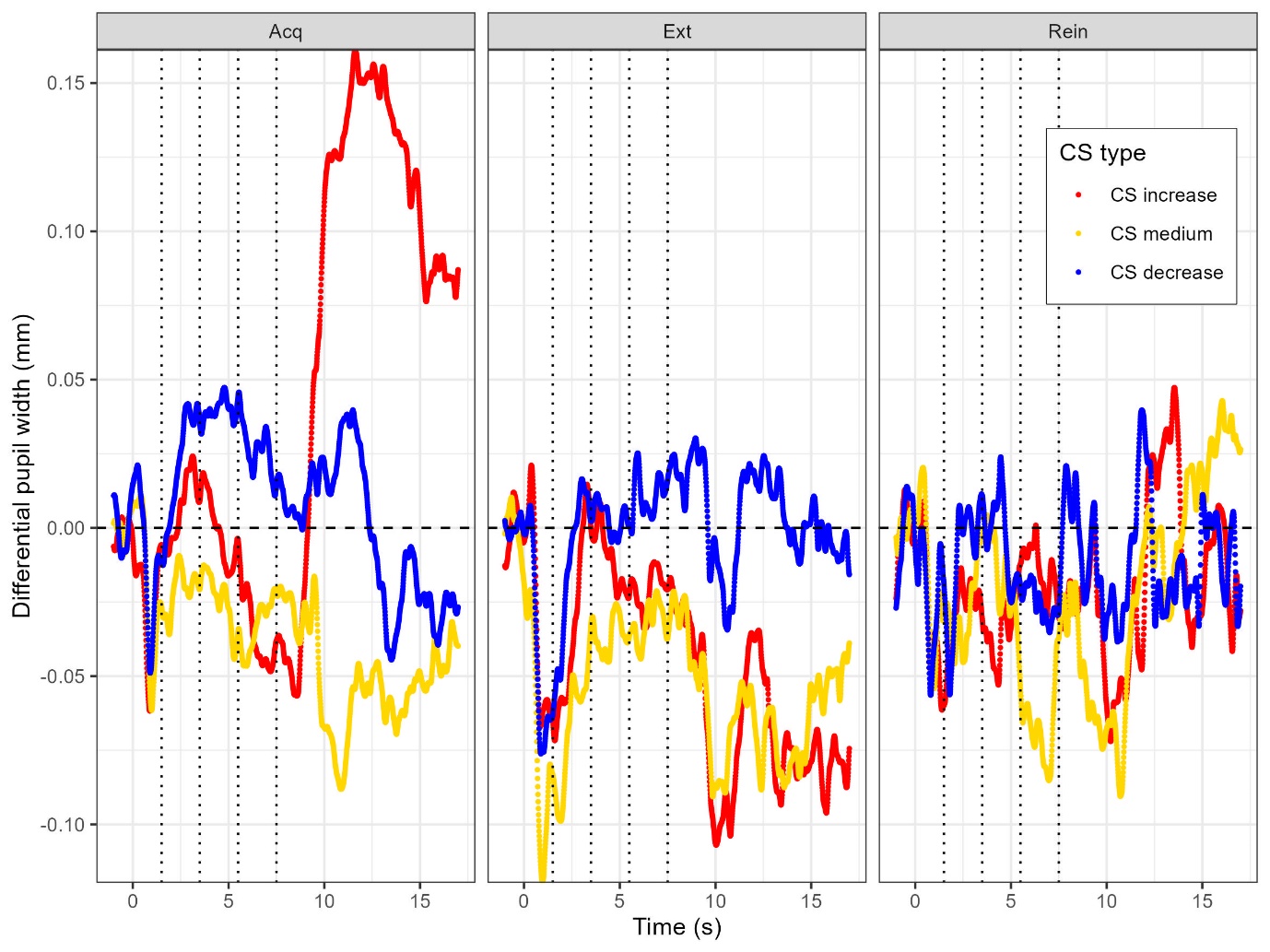


Fig. 5. Mean changes in pupil width (mm) relative to the pre-stimulus baseline over the time course of a trial, averaged across all participants and within the respective phases (acquisition training, extinction training, reinstatement phase). Dotted lines indicate analysis bins for the main study (following the initial light reaction to the CS, 1.5 after CS onset, until the onset of the US, 7.5 s after CS onset). Data are provided for N = 25 participants.

**Data availability**

We commit to sharing our raw data and materials on acceptance of the Stage 2 manuscript. The data will be shared on OSF in BIDS.

**Code availability**

We commit to sharing our code scripts and materials on acceptance of the Stage 2 manuscript.

**References**

Andreatta, M., & Pauli, P. (2015). Appetitive vs. Aversive conditioning in humans. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00128

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

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

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The Role of the Dorsal Striatum in Reward and Decision-Making: Figure 1. *The Journal of Neuroscience*, *27*(31), 8161–8165. https://doi.org/10.1523/JNEUROSCI.1554-07.2007

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4*. https://doi.org/10.48550/ARXIV.1406.5823

Battaglia, S., Harrison, B. J., & Fullana, M. A. (2022). Does the human ventromedial prefrontal cortex support fear learning, fear extinction or both? A commentary on subregional contributions. *Molecular Psychiatry*, *27*(2), 784–786. https://doi.org/10.1038/s41380-021-01326-4

Becerra, L., Navratilova, E., Porreca, F., & Borsook, D. (2013). Analogous responses in the nucleus accumbens and cingulate cortex to pain onset (aversion) and offset (relief) in rats and humans. *Journal of Neurophysiology*, *110*(5), 1221–1226. https://doi.org/10.1152/jn.00284.2013

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Belleau, E. L., Pedersen, W. S., Miskovich, T. A., Helmstetter, F. J., & Larson, C. L. (2018). Cortico-limbic connectivity changes following fear extinction and relationships with trait anxiety. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsy073

Belova, M. A., Paton, J. J., & Salzman, C. D. (2008). Moment-to-Moment Tracking of State Value in the Amygdala. *Journal of Neuroscience*, *28*(40), 10023–10030. https://doi.org/10.1523/JNEUROSCI.1400-08.2008

Cha, J., Greenberg, T., Carlson, J. M., DeDora, D. J., Hajcak, G., & Mujica-Parodi, L. R. (2014). Circuit-Wide Structural and Functional Measures Predict Ventromedial Prefrontal Cortex Fear Generalization: Implications for Generalized Anxiety Disorder. *Journal of Neuroscience*, *34*(11), 4043–4053. https://doi.org/10.1523/JNEUROSCI.3372-13.2014

Cooper, J. C., Dunne, S., Furey, T., & O’Doherty, J. P. (2012). Human Dorsal Striatum Encodes Prediction Errors during Observational Learning of Instrumental Actions. *Journal of Cognitive Neuroscience*, *24*(1), 106–118. https://doi.org/10.1162/jocn\_a\_00114

Costa, V. D., Bradley, M. M., & Lang, P. J. (2015). From threat to safety: Instructed reversal of defensive reactions: Reversing defensive reactions. *Psychophysiology*, *52*(3), 325–332. https://doi.org/10.1111/psyp.12359

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Dirikx, T., Hermans, D., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2004). Reinstatement of Extinguished Conditioned Responses and Negative Stimulus Valence as a Pathway to Return of Fear in Humans. *Learning & Memory*, *11*(5), 549–554. https://doi.org/10.1101/lm.78004

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

Duits, P., Richter, J., Baas, J. M. P., Engelhard, I. M., Limberg-Thiesen, A., Heitland, I., Hamm, A. O., & Cath, D. C. (2017). Enhancing effects of contingency instructions on fear acquisition and extinction in anxiety disorders. *Journal of Abnormal Psychology*, *126*(4), 378–391. https://doi.org/10.1037/abn0000266

Esteban, O., Ciric, R., Finc, K., Blair, R. W., Markiewicz, C. J., Moodie, C. A., Kent, J. D., Goncalves, M., DuPre, E., Gomez, D. E. P., Ye, Z., Salo, T., Valabregue, R., Amlien, I. K., Liem, F., Jacoby, N., Stojić, H., Cieslak, M., Urchs, S., … Gorgolewski, K. J. (2020). Analysis of task-based functional MRI data preprocessed with fMRIPrep. *Nature Protocols*, *15*(7), 2186–2202. https://doi.org/10.1038/s41596-020-0327-3

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Feng, P., Zheng, Y., & Feng, T. (2016). Resting-state functional connectivity between amygdala and the ventromedial prefrontal cortex following fear reminder predicts fear extinction. *Social Cognitive and Affective Neuroscience*, *11*(6), 991–1001. https://doi.org/10.1093/scan/nsw031

Finke, J. B., Roesmann, K., Stalder, T., & Klucken, T. (2021). Pupil dilation as an index of Pavlovian conditioning. A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *130*, 351–368. https://doi.org/10.1016/j.neubiorev.2021.09.005

Fliege, H., Rose, M., Arck, P., Walter, O. B., Kocalevent, R.-D., Weber, C., & Klapp, B. F. (2005). The Perceived Stress Questionnaire (PSQ) Reconsidered: Validation and Reference Values From Different Clinical and Healthy Adult Samples: *Psychosomatic Medicine*, *67*(1), 78–88. https://doi.org/10.1097/01.psy.0000151491.80178.78

Flor, H. (2012). New developments in the understanding and management of persistent pain. *Current Opinion in Psychiatry*, *25*(2), 109–113. https://doi.org/10.1097/YCO.0b013e3283503510

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry*, *21*(4), 500–508. https://doi.org/10.1038/mp.2015.88

Gámez, W., Chmielewski, M., Kotov, R., Ruggero, C., & Watson, D. (2011). Development of a measure of experiential avoidance: The Multidimensional Experiential Avoidance Questionnaire. *Psychological Assessment*, *23*(3), 692–713. https://doi.org/10.1037/a0023242

Hashmi, J. A., Kong, J., Spaeth, R., Khan, S., Kaptchuk, T. J., & Gollub, R. L. (2014). Functional Network Architecture Predicts Psychologically Mediated Analgesia Related to Treatment in Chronic Knee Pain Patients. *The Journal of Neuroscience*, *34*(11), 3924–3936. https://doi.org/10.1523/JNEUROSCI.3155-13.2014

Hayes, D. J., Duncan, N. W., Xu, J., & Northoff, G. (2014). A comparison of neural responses to appetitive and aversive stimuli in humans and other mammals. *Neuroscience & Biobehavioral Reviews*, *45*, 350–368. https://doi.org/10.1016/j.neubiorev.2014.06.018

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, *48*(2), 510–527. https://doi.org/10.3758/s13428-015-0588-x

Icenhour, A., Langhorst, J., Benson, S., Schlamann, M., Hampel, S., Engler, H., Forsting, M., & Elsenbruch, S. (2015). Neural circuitry of abdominal pain-related fear learning and reinstatement in irritable bowel syndrome. *Neurogastroenterology & Motility*, *27*(1), 114–127. https://doi.org/10.1111/nmo.12489

Javanbakht, A., Duval, E. R., Cisneros, M. E., Taylor, S. F., Kessler, D., & Liberzon, I. (2017). Instructed fear learning, extinction, and recall: Additive effects of cognitive information on emotional learning of fear. *Cognition and Emotion*, *31*(5), 980–987. https://doi.org/10.1080/02699931.2016.1169997

Jeffreys, H. (1998). *The theory of probability*. OUP.

Jentsch, V. L., Wolf, O. T., & Merz, C. J. (2020). Temporal dynamics of conditioned skin conductance and pupillary responses during fear acquisition and extinction. *International Journal of Psychophysiology*, *147*, 93–99. https://doi.org/10.1016/j.ijpsycho.2019.11.006

Kemper, C. J., Ziegler, M., & Taylor, S. (2009). Überprüfung der psychometrischen Qualität der deutschen Version des Angstsensitivitätsindex-3. *Diagnostica*, *55*(4), 223–233. https://doi.org/10.1026/0012-1924.55.4.223

Kemper, C. J., Ziegler, M., & Taylor, S. (2011). *ASI-3—Angstsensitivitätsindex-3*. https://doi.org/10.23668/PSYCHARCHIVES.4526

Klein, S., Kruse, O., Tapia León, I., Van Oudenhove, L., van ’t Hof, S. R., Klucken, T., Wager, T. D., & Stark, R. (2022). Cross-paradigm integration shows a common neural basis for aversive and appetitive conditioning. *NeuroImage*, *263*, 119594. https://doi.org/10.1016/j.neuroimage.2022.119594

Klumpp, H., Keutmann, M. K., Fitzgerald, D. A., Shankman, S. A., & Phan, K. L. (2014). Resting state amygdala-prefrontal connectivity predicts symptom change after cognitive behavioral therapy in generalized social anxiety disorder. *Biology of Mood & Anxiety Disorders*, *4*(1), 14. https://doi.org/10.1186/s13587-014-0014-5

Kreddig, N., Rusu, A. C., Burkhardt, K., & Hasenbring, M. I. (2015). The German PASS-20 in Patients with Low Back Pain: New Aspects of Convergent, Divergent, and Criterion-Related Validity. *International Journal of Behavioral Medicine*, *22*(2), 197–205. https://doi.org/10.1007/s12529-014-9426-2

Laux, L., Hock, M., Bergner-Köther, R., Hodapp, V., & Renner, K.-H. (2013). *Das State-Trait-Angst-Depressions-Inventar*. Hogrefe.

Leeuw, M., Goossens, M. E. J. B., Linton, S. J., Crombez, G., Boersma, K., & Vlaeyen, J. W. S. (2007). The Fear-Avoidance Model of Musculoskeletal Pain: Current State of Scientific Evidence. *Journal of Behavioral Medicine*, *30*(1), 77–94. https://doi.org/10.1007/s10865-006-9085-0

Leknes, S., Brooks, J. C. W., Wiech, K., & Tracey, I. (2008). Pain relief as an opponent process: A psychophysical investigation. *European Journal of Neuroscience*, *28*(4), 794–801. https://doi.org/10.1111/j.1460-9568.2008.06380.x

Leknes, S., Lee, M., Berna, C., Andersson, J., & Tracey, I. (2011). Relief as a Reward: Hedonic and Neural Responses to Safety from Pain. *PLoS ONE*, *6*(4), e17870. https://doi.org/10.1371/journal.pone.0017870

Leknes, S., & Tracey, I. (2008). A common neurobiology for pain and pleasure. *Nature Reviews Neuroscience*, *9*(4), 314–320. https://doi.org/10.1038/nrn2333

Leuchs, L., Schneider, M., Czisch, M., & Spoormaker, V. I. (2017). Neural correlates of pupil dilation during human fear learning. *NeuroImage*, *147*, 186–197. https://doi.org/10.1016/j.neuroimage.2016.11.072

Leuchs, L., Schneider, M., & Spoormaker, V. I. (2019). Measuring the conditioned response: A comparison of pupillometry, skin conductance, and startle electromyography. *Psychophysiology*, *56*(1), e13283. https://doi.org/10.1111/psyp.13283

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C., & Andreoli, A. (1993). Development of the perceived stress questionnaire: A new tool for psychosomatic research. *Journal of Psychosomatic Research*, *37*(1), 19–32. https://doi.org/10.1016/0022-3999(93)90120-5

Li, J., Delgado, M. R., & Phelps, E. A. (2011). How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences*, *108*(1), 55–60. https://doi.org/10.1073/pnas.1014938108

Liljeholm, M., & O’Doherty, J. P. (2012). Contributions of the striatum to learning, motivation, and performance: An associative account. *Trends in Cognitive Sciences*, *16*(9), 467–475. https://doi.org/10.1016/j.tics.2012.07.007

Linton, S. J. (2000). A review of psychological risk factors in back and neck pain. *Spine*, *25*(9), 1148–1156. https://doi.org/10.1097/00007632-200005010-00017

Lissek, S., & Tegenthoff, M. (2021). Higher functional connectivity between prefrontal regions and the dorsal attention network predicts absence of renewal. *Behavioural Brain Research*, *412*, 113413. https://doi.org/10.1016/j.bbr.2021.113413

Luck, C. C., & Lipp, O. V. (2016). Instructed extinction in human fear conditioning: History, recent developments, and future directions. *Australian Journal of Psychology*, *68*(3), 209–227. https://doi.org/10.1111/ajpy.12135

Martynova, O., Tetereva, A., Balaev, V., Portnova, G., Ushakov, V., & Ivanitsky, A. (2020). Longitudinal changes of resting-state functional connectivity of amygdala following fear learning and extinction. *International Journal of Psychophysiology*, *149*, 15–24. https://doi.org/10.1016/j.ijpsycho.2020.01.002

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

McCracken, L. M., & Dhingra, L. (2002). A Short Version of the Pain Anxiety Symptoms Scale (PASS-20): Preliminary Development and Validity. *Pain Research and Management*, *7*(1), 45–50. https://doi.org/10.1155/2002/517163

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

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

Mertens, G., Kuhn, M., Raes, A. K., Kalisch, R., De Houwer, J., & Lonsdorf, T. B. (2016). Fear expression and return of fear following threat instruction with or without direct contingency experience. *Cognition and Emotion*, *30*(5), 968–984. https://doi.org/10.1080/02699931.2015.1038219

Meulders, A., Rousseau, A., & Vlaeyen, J. W. S. (2015). Motor Intention as a Trigger for Fear of Movement-related Pain: An Experimental Cross-US Reinstatement Study. *Journal of Experimental Psychopathology*, *6*(3), 206–228. https://doi.org/10.5127/jep.043614

Meulders, A., & Vlaeyen, J. W. S. (2013). Mere Intention to Perform Painful Movements Elicits Fear of Movement-Related Pain: An Experimental Study on Fear Acquisition Beyond Actual Movements. *The Journal of Pain*, *14*(4), 412–423. https://doi.org/10.1016/j.jpain.2012.12.014

Meyer, K., Sprott, H., & Mannion, A. F. (2008). Cross-cultural adaptation, reliability, and validity of the German version of the Pain Catastrophizing Scale. *Journal of Psychosomatic Research*, *64*(5), 469–478. https://doi.org/10.1016/j.jpsychores.2007.12.004

Milad, M. R., & Quirk, G. J. (2012). Fear Extinction as a Model for Translational Neuroscience: Ten Years of Progress. *Annual Review of Psychology*, *63*(1), 129–151. https://doi.org/10.1146/annurev.psych.121208.131631

Nees, F., & Becker, S. (2018). Psychological Processes in Chronic Pain: Influences of Reward and Fear Learning as Key Mechanisms – Behavioral Evidence, Neural Circuits, and Maladaptive Changes. *Neuroscience*, *387*, 72–84. https://doi.org/10.1016/j.neuroscience.2017.08.051

Öhman, A. (1972). Factor Analytically Derived Components of Orienting, Defensive, and Conditioned Behavior in Electrodermal Conditioning. *Psychophysiology*, *9*(2), 199–209. https://doi.org/10.1111/j.1469-8986.1972.tb00754.x

Öhman, A. (1974). Orienting reactions, expectancy learning, and conditioned responses in electrodermal conditioning with different interstimulus intervals. *Biological Psychology*, *1*(3), 189–200. https://doi.org/10.1016/0301-0511(74)90011-8

Oldfield, R. C. (2013). *Edinburgh Handedness Inventory* [Data set]. American Psychological Association. https://doi.org/10.1037/t23111-000

Oldham, S., Murawski, C., Fornito, A., Youssef, G., Yücel, M., & Lorenzetti, V. (2018). The anticipation and outcome phases of reward and loss processing: A neuroimaging meta‐analysis of the monetary incentive delay task. *Human Brain Mapping*, *39*(8), 3398–3418. https://doi.org/10.1002/hbm.24184

Pauli, W. M., Gentile, G., Collette, S., Tyszka, J. M., & O’Doherty, J. P. (2019). Evidence for model-based encoding of Pavlovian contingencies in the human brain. *Nature Communications*, *10*(1), 1099. https://doi.org/10.1038/s41467-019-08922-7

Petersch, B., & Dierkes, K. (2022). Gaze-angle dependency of pupil-size measurements in head-mounted eye tracking. *Behavior Research Methods*, *54*(2), 763–779. https://doi.org/10.3758/s13428-021-01657-8

Petersen, K. L., & Rowbotham, M. C. (1999). A new human experimental pain model: The heat/capsaicin sensitization model. *NeuroReport*, *10*(7), 1511–1516. https://doi.org/10.1097/00001756-199905140-00022

Pfannmöller, J., & Lotze, M. (2019). Review on biomarkers in the resting-state networks of chronic pain patients. *Brain and Cognition*, *131*, 4–9. https://doi.org/10.1016/j.bandc.2018.06.005

Renner, K.-H., Hock, M., Bergner-Köther, R., & Laux, L. (2018). Differentiating anxiety and depression: The State-Trait Anxiety-Depression Inventory. *Cognition and Emotion*, *32*(7), 1409–1423. https://doi.org/10.1080/02699931.2016.1266306

Schaeuffele, C., Knaevelsrud, C., Renneberg, B., & Boettcher, J. (2022). Psychometric Properties of the German Brief Experiential Avoidance Questionnaire (BEAQ). *Assessment*, *29*(7), 1406–1421. https://doi.org/10.1177/10731911211010955

Scheveneels, S., Boddez, Y., De Ceulaer, T., & Hermans, D. (2019). Ruining the surprise: The effect of safety information before extinction on return of fear. *Journal of Behavior Therapy and Experimental Psychiatry*, *63*, 73–78. https://doi.org/10.1016/j.jbtep.2018.11.001

Schlitt, F., Schmidt, K., Merz, C. J., Wolf, O. T., Kleine-Borgmann, J., Elsenbruch, S., Wiech, K., Forkmann, K., & Bingel, U. (2022). Impaired pain-related threat and safety learning in patients with chronic back pain. *Pain*, *163*(8), 1560–1570. https://doi.org/10.1097/j.pain.0000000000002544

Schmidt, K., Forkmann, K., Elsenbruch, S., & Bingel, U. (2020). Enhanced pain-related conditioning for face compared to hand pain. *PLOS ONE*, *15*(6), e0234160. https://doi.org/10.1371/journal.pone.0234160

Schultz, D. H., Balderston, N. L., & Helmstetter, F. J. (2012). Resting-state connectivity of the amygdala is altered following Pavlovian fear conditioning. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00242

Sescousse, G., Caldú, X., Segura, B., & Dreher, J.-C. (2013). Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *37*(4), 681–696. https://doi.org/10.1016/j.neubiorev.2013.02.002

Sevenster, D., Beckers, T., & Kindt, M. (2012). Instructed extinction differentially affects the emotional and cognitive expression of associative fear memory: Instructed extinction of startle response and SCR. *Psychophysiology*, *49*(10), 1426–1435. https://doi.org/10.1111/j.1469-8986.2012.01450.x

Seymour, B., O’Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., & Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, *8*(9), 1234–1240. https://doi.org/10.1038/nn1527

Sharpe, M. J., & Schoenbaum, G. (2016). Back to basics: Making predictions in the orbitofrontal–amygdala circuit. *Neurobiology of Learning and Memory*, *131*, 201–206. https://doi.org/10.1016/j.nlm.2016.04.009

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain’s functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*(31), 13040–13045. https://doi.org/10.1073/pnas.0905267106

Solomon, R. L., & Wynne, L. C. (1954). Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review*, *61*(6), 353–385. https://doi.org/10.1037/h0054540

Spisak, T., Kincses, B., Schlitt, F., Zunhammer, M., Schmidt-Wilcke, T., Kincses, Z. T., & Bingel, U. (2020). Pain-free resting-state functional brain connectivity predicts individual pain sensitivity. *Nature Communications*, *11*(1), 187. https://doi.org/10.1038/s41467-019-13785-z

Spisák, T., Spisák, Z., Zunhammer, M., Bingel, U., Smith, S., Nichols, T., & Kincses, T. (2019). Probabilistic TFCE: A generalized combination of cluster size and voxel intensity to increase statistical power. *NeuroImage*, *185*, 12–26. https://doi.org/10.1016/j.neuroimage.2018.09.078

Stanley, A. T., Lippiello, P., Sulzer, D., & Miniaci, M. C. (2021). Roles for the Dorsal Striatum in Aversive Behavior. *Frontiers in Cellular Neuroscience*, *15*, 634493. https://doi.org/10.3389/fncel.2021.634493

Sullivan, M. J. L., Bishop, S. R., & Pivik, J. (1995). The Pain Catastrophizing Scale: Development and validation. *Psychological Assessment*, *7*(4), 524–532. https://doi.org/10.1037/1040-3590.7.4.524

Taylor, S., Zvolensky, M. J., Cox, B. J., Deacon, B., Heimberg, R. G., Ledley, D. R., Abramowitz, J. S., Holaway, R. M., Sandin, B., Stewart, S. H., Coles, M., Eng, W., Daly, E. S., Arrindell, W. A., Bouvard, M., & Cardenas, S. J. (2007). Robust dimensions of anxiety sensitivity: Development and initial validation of the Anxiety Sensitivity Index-3. *Psychological Assessment*, *19*(2), 176–188. https://doi.org/10.1037/1040-3590.19.2.176

van der Schaaf, M. E., Schmidt, K., Kaur, J., Gamer, M., Wiech, K., Forkmann, K., & Bingel, U. (2022). Acquisition learning is stronger for aversive than appetitive events. *Communications Biology*, *5*(1), 302. https://doi.org/10.1038/s42003-022-03234-x

Vlaeyen, J. W. S., Kole-Snijders, A. M. J., Boeren, R. G. B., & van Eek, H. (1995). Fear of movement/(re)injury in chronic low back pain and its relation to behavioral performance. *Pain*, *62*(3), 363–372. https://doi.org/10.1016/0304-3959(94)00279-N

Vlaeyen, J. W. S., & Linton, S. J. (2012). Fear-avoidance model of chronic musculoskeletal pain: 12 years on. *Pain*, *153*(6), 1144–1147. https://doi.org/10.1016/j.pain.2011.12.009

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wendt, J., Hufenbach, M. C., König, J., & Hamm, A. O. (2020). Effects of verbal instructions and physical threat removal prior to extinction training on the return of conditioned fear. *Scientific Reports*, *10*(1), 1202. https://doi.org/10.1038/s41598-020-57934-7

Wendt, J., & Morriss, J. (2022). An examination of Intolerance of Uncertainty and contingency instruction on multiple indices during threat acquisition and extinction training. *International Journal of Psychophysiology*, *177*, 171–178. https://doi.org/10.1016/j.ijpsycho.2022.05.005

Whitfield-Gabrieli, S., & Nieto-Castanon, A. (2012). *Conn*: A Functional Connectivity Toolbox for Correlated and Anticorrelated Brain Networks. *Brain Connectivity*, *2*(3), 125–141. https://doi.org/10.1089/brain.2012.0073

Wolter, J., & Lachnit, H. (1993). Are anticipatory first and second interval skin conductance responses indicators of predicted aversiveness? *Integrative Physiological and Behavioral Science*, *28*(2), 163–166. https://doi.org/10.1007/BF02691221

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635

Zbozinek, T. D., Hermans, D., Prenoveau, J. M., Liao, B., & Craske, M. G. (2015). Post-extinction conditional stimulus valence predicts reinstatement fear: Relevance for long-term outcomes of exposure therapy. *Cognition and Emotion*, *29*(4), 654–667. https://doi.org/10.1080/02699931.2014.930421

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

**Acknowledgements**

The work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project A11, project number 316803389 - SFB 1280 (Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – Projekt A11, Projektnummer316803389 - SFB 1280).

**Author Contributions**

Study concept and design: Bingel, Schmidt. Data acquisition: Busch. Analysis: all authors. Drafting of manuscript: all authors. Critical revision of manuscript: all authors. Funding: Bingel, Schmidt.

**Competing Interests**

We declare that we have no competing interests.