Reviews

Reviewed by Tom Beckers, 15 Feb 2023 10:32

I have read this stage 1 proposal with great interest. It addresses an interesting research question of basic and translational interest with overall sophisticated and appropriate methods. Whatever the study's results, they should be of interest to the field. However, whereas the overall rationale for the study is sound and convincing, quite a number of specific aspects of the design, procedures, hypotheses and analyses of the proposed study are less well justified or elaborated. As such, I think that there are a number of issues throughout the manuscript that merit addressing prior to the start of data collection. I list them below in order of appearance in the manuscript:

We would like to thank the reviewer for his constructive evaluation and the helpful comments which have clearly improved the manuscript and the planned study.

1. The description of the prior study from which the current proposal takes inspiration is rather confusing to me (p. 3). First, on line 12 and beyond, I think CS and US have been switched; the increase and decrease in pain are USs, not CSs. More importantly, the description of the results of that study (line 17-21) on the one hand suggests that changes in differential CS valence ratings over the course of extinction training were similar for appetitive and aversive CSs (lines 17-19) but at the same time states that extinction of aversive CS valence ratings was incomplete (with an unstated implication that is was complete for appetitive CS valence; lines 19-21). It would be good to clarify this.

In response to the reviewer's comment, we revised the respective sections in the manuscript. While there were indeed no differences between differential ratings of appetitive and aversive CS during extinction training (i.e., in the slope), CS valence ratings at the end of the extinction phase differed significantly from those during habituation but only for the aversive CS. The revised sentences on page 3 now reads:

"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)".

2. Further down the introduction (p. 4), it is stated that Sevenster et al. (2012) showed that instructed extinction immediately abolished differential US expectancies but left SCR to the CS+ unaffected (lines 5-7). While factually correct, this is a bit misleading, given that differential SCR was completely absent from the first extinction trial in the instructed extinction group (see Sevenster et al., 2012, Figure 4). More broadly, I think there is little evidence in the literature to support the claim that instructed extinction is less effective for SCR than for US expectancies (in fact, even the claim by Sevenster et al., 2012, that it is less effective for fear-potentiated startle has been disputed).

We agree that physiological (SCR/startle) and cognitive-evaluative responses should behave in a similar way as outcomes assessing instruction effects, since also SCR has shown clear susceptibility to instructions in previous instructed reversal studies (Mertens et al., 2018; Mertens & De Houwer, 2016). We have now thoroughly revised and restructured our introduction accordingly. The addition to our manuscript reads as follows (see page 4):

"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)."

3. I don't fully understand what the rationale is for predicting a stronger effect of instructions on extinction of the appetitive than the aversive CS (e.g., p. 5, lines 18-19, but also elsewhere). Given that without instructions, extinction is expected to be weaker/slower for the aversive CS, one would think that there is more room for instructions to facilitate extinction for that cue. The authors seem to be ambivalent about this as well, because further down in the manuscript, they make different predictions in this regard for US expectancies (see H4c on p. 24) than for CS valence (see H4d on p. 25), without further justification or discussion. I think this needs straightening out, given that testing this specific interaction between US type and instruction is the core raison d'etre for the proposed study's factorial design. In that sense, it is also a bit strange to formulate this analysis as being exploratory; the authors are clearly intending to do confirmatory analyses to test the presence of an interaction and undoubtedly hope to draw directional conclusions from the results regarding the presence or absence of a difference in the effect of instructed extinction on appetitive versus aversive learning.

We agree that one of the reasons for our complex study design is indeed the investigation of the three-way interaction (i.e., whether there is a more pronounced effect of instruction on extinction slopes of appetitive compared to aversive CS). We have now changed the manuscript accordingly and describe the hypothesis as confirmatory. Although the aversive CS leaves more leeway for a boosting effect of instructions on CS valence as a measure of extinction, we assume that negative outcome expectations/predictions will be updated more carefully in accordance with a 'better safe than sorry strategy'. Based on this assumption, we expect the instruction effects on aversive US expectancy to be less pronounced. We thus expect the instruction to have stronger effects on expectancy ratings of appetitive than aversive CS which is expected to manifest as steeper extinction slope for the $\mathrm{CS}_{\mathrm{decrease}}$. CS valence ratings will only be obtained between phases. They are therefore less fine-grained but are nevertheless expected to show the same overall effects as expectancy ratings. We have adapted this in the revised manuscript. H4 on page 7 now reads:

"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".

We also adapted Table 1 on pages 12-13 accordingly.

4. I think the introduction in its present form does not provide sufficient justification for the inclusion of the various dependent variables that will be measured. The authors do state that the US expectancy ratings will be the primary measure of interest, but what are the valence ratings, SCRs and pupil dilation responses each supposed to add? Why are they included? And how will the authors handle divergence in results between these measures? Justification for the inclusion of SCR and pupil dilation in particular is not trivial. Is there good evidence that SCR and pupil dilation are appropriate measures here, in a design that involves a salient appetitive as well as a salient aversive US? Particularly given the nature of the proposed procedure, where trials start with a pre-US situation of moderate pain: The fact that the appetitive CS no longer signals a reduction in pain during extinction might be perceived as an aversive outcome in the appetitive group, which could support an increase rather than a decrease in SCR to the appetitive CS during extinction, which would not actually reflect a lack of learning. The same may be true for pupil dilation. This may all hinder interpretation of possible differences between the appetitive and the aversive CS during extinction. At a minimum, this warrants some justification/consideration. None of these issues would seem to plague the US expectancy ratings, for which a direct comparison of responding to the appetitive and the aversive CS seems much more straightforward.

The decision to include these additional outcome measures was based on the following considerations: CS valence ratings and SCR measurements will be included to test whether previous findings using these outcomes can be replicated (e.g., van der Schaaf & Schmidt et al., 2022; https://drks.de/search/de/trial/DRKS00027448). In our view, CS valence ratings – although only obtained at the end of each phase – provide interesting and complementary information as they reflect an emotional component, which is not captured by US expectancy, and which has been shown to be more resistant to extinction than expectancy (e.g., Dirikx et al., 2004; Zbozinek, Hermans, et al., 2015; Zbozinek, Holmes, et al., 2015). Second, we want to complement these subjective and cognitive-emotional measures with SCR and pupil dilation as two physiological measures. SCR amplitudes have been found to reflect both appetitive and aversive acquisition and extinction learning (Andreatta & Pauli, 2015). A recent meta-analysis showed that pupil dilation is an appropriate measure for both appetitive and aversive conditioning (Finke et al., 2021), which was also less prone to habituation during the experiment, thus potentially making pupil dilations more sensitive to extinction effects than SCR (Leuchs et al., 2019). By including both methods, we intend to capture the relevant processes.

The interpretation of our results will be based on the primary outcome of US expectancy ratings. In case of divergent results for primary and secondary outcomes, all results will be considered in the overall interpretation of the findings, but the main conclusions will be based on the primary outcome.

Regarding the reviewer's thoughts on valence ratings and SCR of the $\text{CS}_{\text{decrease}}$ during extinction training, we would like to mention that in our previous study, $\text{CS}_{\text{decrease}}$ did not become unpleasant during extinction training (van der Schaaf & Schmidt et al., 2022). We therefore do not expect aversive learning for the $\text{CS}_{\text{decrease}}$ during extinction and vice versa, i.e., appetitive learning for the $\text{CS}_{\text{increase}}$.

We also added the rationale for the selection of outcome measures to our introduction (see page 4), which now reads as follows:

"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)".

5. Relatedly, no clear justification is provided for including only US expectancy as a measure of conditioning in the analyses for the second manuscript.

We apologize that a justification was indeed missing. We now clarify that we focus on US expectancy ratings which is also the main outcome for the behavioral analyses (Manuscript 1) as it has shown instruction effects on conditioning in previous studies (e.g., Duits et al., 2017; Mertens et al., 2016; Scheveneels et al., 2019; Sevenster et al., 2012). The revised part of the introduction (p. 7) reads as follows:

"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)".

6. Regarding the sample size and stopping rule, two issues warrant elaboration. First, I think it would be more appropriate to stop data collection if for all hypotheses a BF10 of either 6 or 1/6 is reached, rather than a BF > 6 for all hypotheses, so as to not bias data collection towards positive results. Second, it isn't clear how the authors, starting from the effect size in Sevenster et al. (2012), arrived at their intended sample size of 150 (p. 14). This deserves elaboration.

We agree with the reviewer's point on the stopping rule and have now adjusted the criterion accordingly in our manuscript (p. 16).

"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 (BF_{10}) in favor of our (main) hypotheses (i.e., H1-4a) reach $BF_{10}>6$ or $BF_{10}<1/6$ (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."

The intended sample size of *N* = 150 was chosen to enable us to find medium effect sizes for the effects of interests specified in all hypotheses based on power calculations. The calculation was approximated using the chi-square test from the pwr package in R, as we will use the Wald chi-square tests of model results of the winning model to assess the fixed effects. Since to our knowledge, there is no tool available to calculate the sample size needed given the smallest effect of interest in LMM for our analyses which include fixed-effects interactions (such as http://jakewestfall.org/power/), and as it is not recommended to base sample size estimations solely on pilot data, we based our decision on the reported approximation method and on studies using a similar design reporting small to medium effect sizes (Schlitt et al., 2022; van der Schaaf et al., 2022; https://drks.de/search/de/trial/DRKS00027448). Due to the more complex design involving three CS, and the planned comparison of the two CS^+, we expect a smaller effect of the instruction on extinction efficacy than reported by Sevenster et al. (2012), and therefore aimed for a higher sample size. In addition to considering previous experiments, we also followed the recommendations for investigating unknown effect sizes, i.e., to have a sample size of $n \geq 50$ participants per group in a between-subjects design (Simmons et al., 2013).

7. Reinstatement of conditioned responding (or recovery of extinguished responding in general) isn't really covered in the introduction; as a result, the inclusion of a reinstatement phase in the experiment feels like it is lacking a clear rationale.

We would like to thank the reviewer for this comment. Reinstatement effects have previously been shown for pain-related learning (Icenhour et al., 2015; Meulders et al., 2015; Schmidt et al., 2020) but not for pain relief learning. Furthermore, such aversive and appetitive effects have not been directly compared during reinstatement. Reinstatement will therefore be used to test for extinction efficacy in both pain and pain relief learning in the same sensory modality. We now clarified this in the introduction of our manuscript (pages 3 and 6).

"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."

"Extinction efficacy will also be tested using a reinstatement manipulation and test phase."

8. The authors list an appropriately ordered difference in US painfulness ratings as manipulation check. Fair as this may be, I think a more relevant positive control would be the observation of differential acquisition for all measures and for both types of US by the end of acquisition.

Following the reviewer's suggestion, we have included an additional manipulation check based on CS-US contingency ratings to assess whether, overall, participants learned the respective contingencies as specified in Table 1. Please note that in our previous study contingency ratings indicated successful learning for all CS types (van der Schaaf & Schmidt et al., 2022). We are therefore confident that a similar level of learning can be achieved in the planned study. Furthermore, we decided to add CS-US contingencies as a potential covariate to our analyses (compare section 2.5.2.1 Behavioral measures p. 27 *"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.").*

9. The US is variably described as an increase/decrease/constant temperature, or as a pain exacerbation / pain decrease / no change in pain. One is a way to achieve the second, obviously, but it would help clarity if the authors used a consistent terminology for what the USs are throughout.

We now consistently use the terms "CS/US increase/decrease/medium" and "temperature increase/decrease" along with "no temperature change" (i.e., in reference to the temperature levels used to induce different pain levels) in the methods and results sections of the manuscript (please see tracked changes). Only in an interpretative context (e.g., in the abstract and discussion), we also use terms such as "medium pain level" and "pain exacerbation/relief" for better readability.

10. Participants are excluded if they have recently participated in pharmacological studies (p. 14), but can they have taken part in conditioning/extinction experiments before? It seems like that might affect their speed of acquisition and extinction learning rather substantially (cf. the literature on re-acquisition and re-extinction).

We apologize that we forgot to mention this exclusion criterion. This is now specified on pages 16-17 of the methods section:

"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. […]".

11. It is a bit odd that the label for the US expectancy scale reads 'most probably cooling decrease' on the one side and 'most probably heating' on the other side (p. 16, line 11-12). Why not just 'most probably cooling' and 'most probably heating' (or, alternatively, 'most probably cooling decrease' and 'most probably heating increase')?

We thank the reviewer for spotting this error. The original German label reads "sehr wahrscheinlich Abkühlung", which indeed translates into "most probably cooling". This has now been corrected.

12. On p. 17, covariates are introduced that haven't been mentioned previously. Will these questionnaire scores be used for screening out participants only, or will they also be used for additional analyses on included participants?

Questionnaire scores will be used for both – to screen out participants but also as covariates to explore associations between pain-related associative learning and individual pain-related cognition (Nees & Becker, 2018). This is now explained in the section on covariates on pages 18-19 (*"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."*), and it is further described in the analysis plan (pp. 25-29, exemplarily shown for US expectancy ratings, p. 26):

"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."

13. I found the description of the calibration procedure on p. 20 difficult to follow. In particular the follow sentence I failed to make sense of (lines 2-3): "This second step will consist of the application of the selected temperature level within a range of -1.5 to and +3.0°C in steps of 0.5°C".

We agree that the description of the calibration processes was quite dense and in part difficult to follow. We have therefore thoroughly revised this section on pages 21-23 of the manuscript as shown below. We also cite a reference to the description of the calibration procedure in our previous study (van der Schaaf & Schmidt et al., 2022).

"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 US_{medium}*,* $US_{increase}$*, and* $US_{decrease}$ *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 $US_{increase}$ *(i.e., VAS 80) and* $US_{decrease}$ *stimuli (i.e., VAS 0, calculated as* US_{medium} *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* US_{medium}*, and three* $US_{increase}$ *and* $US_{decrease}$ *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* $US_{medium} < 10$*) within the predefined safety limits (44°C for the* US_{medium}*, 46°C for the* $US_{increase}$*), or rate stimuli inconsistently (criterion: mean* $US_{increase} > US_{medium} > US_{decrease}$*), 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."*

14. Regarding the trial structure: the intertrial interval of 4-7 s seems unlikely to be sufficient for unconditioned skin conductance responses to the US to return to baseline. A longer interval seems indicated.

In response to the reviewer's point, we would like to clarify that each trial ends with a 4s delay following US offset. Together with the ITI of 4-7s the total time between US offset and onset of the subsequent CS therefore amounts to 8-11 s, which exceeds the minimum recommended duration for SCR to return to baseline (Lonsdorf et al., 2017). This information has been added to page 23:

"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 $US_{increase/decrease}$ 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".

15. Why does the extinction phase involve a slightly smaller number of trials per CS than the acquisition phase? This also translates in different numbers of trials in between US expectancy ratings, for instance, and more generally makes the analyses a bit difficult to compare between acquisition and extinction. Why not simply have 16 trials per CS in each phase?

We planned our study design based on previous experiments that used 16 trials per CS during acquisition training, and 12 trials during extinction training (van der Schaaf & Schmidt et al., 2022; https://drks.de/search/de/trial/DRKS00027448). The slightly higher number of trials in the extinction phase of the present experiment is due to the fact that the behavioral ratings provided throughout this phase were presented less frequently in previous experiments (ratings were provided every 4th trial in previous studies vs every 3rd trial in the current experiment). We chose to increase the frequency of assessing these ratings as we expect a fast effect of instruction (as for instance in Sevenster et al., 2012) and wanted to be able to capture this effect appropriately. To ensure that we can also capture changes towards the end of the extinction training, we decided to add two trials per CS type. Given that we will use a tonic heat pain model (i.e., ongoing pain) which is clearly uncomfortable, we aim to limit the duration of the experiment (and therefore the number of trials) to a minimum. We thus decided to only include 14 trials per CS in the extinction training.

16. Regarding the analyses, I found the use of 'time' for the independent variable of trial a little unfortunate, given that also 'time bin' is used as an IV in some analyses, and time in the latter case refers to within-trial time, whereas in the former case it refers to a much larger scale. 'Trial' would probably be a clearer descriptor than 'time'.

We thank the reviewer for pointing this out and for his suggestion. We would, however, like to suggest renaming the within-trial time from "time bin" to "bin" instead and keeping "time" for the larger scale factor as ratings are not provided on every trial and we use the time factor as a continuous variable in our analyses, which does not correspond to trial numbers. We also adapted the wording in our manuscript accordingly.

17. SCR will be divided in FIR and SIR. It would be good to provide a rationale for this (or, alternatively, to not distinguish between FIR and SIR).

We thank the reviewer for giving us the opportunity to clarify this part of the analysis. The FIR has been associated with an orienting response, while the SIR is thought to reflect US expectancy (Öhman, 1972, 1974; Wolter & Lachnit, 1993). We decided to distinguish between FIR and SIR based on a study by Jentsch et al. (2020), who observed more pronounced FIR to CS in the early acquisition phase, while SIR was increased in the later part of acquisition training. They also reported that only FIR decreased during extinction training. Considering the different processes thought to underly FIR and SIR and their differential temporal dynamics during acquisition and extinction we will analyze these responses separately which is now explained in more detail in the method section (2.5.2.3. SCR data, p. 29).

"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."

18. Regarding the second-level connectivity analyses (p. 28), I would have expected that connectivity would be used as a predictor and acquisition and extinction indices as outcomes (lines 23-25).

We would like to explain that we plan to use the methods implemented in the CONN toolbox, which are similar to SPM. The common approach here is that brain activity (or connectivity) is explained by behavior, not the other way around. However, we still intend to perform advanced predictive modelling approaches, which would indeed allow to predict behavior from connectivity, but we did not include this in our registered report due to a) considerations of the scope, extent and complexity of the current manuscript and b) the more data-driven and thus explorative nature of such analyses.

19. Regarding the same analyses, it isn't clear whether also hypotheses regarding acquisition will be tested – the title on line 16 does suggest so, but on line 31-32, only hypotheses regarding extinction are mentioned.

We apologize for not making this clearer. We indeed want to test these hypotheses for both acquisition and extinction training. This has been clarified in the manuscript (p. 31, *"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 $CS_{increase}$ during acquisition and extinction, and for H2 with the respective index for the $CS_{decrease}$"*).

20. On the next page, to test the direct effect of instruction, US expectancy immediately before and after the instruction will be compared for the instructed group only (lines 5-7). This seems a bit odd. Wouldn't it make more sense to compare the difference in US expectancy from Acq5 to Ext1 between the instructed and the non-instructed group? Likewise, to evaluate the effect of instruction on learning through experience, it would seem more sensible to compare the difference in US expectancy from the end of acquisition to the end of extinction between both groups. Confusingly, further down (lines 19-23), the proposed test of the hypothesis does involve an interaction with instruction group, in contradiction to the preceding section.

We apologize for the confusion. Indeed, we will compare the associations of both mentioned behavioral indices of the instruction effect with connectivity data between groups, as stated in our hypotheses.

The immediate effect of instruction will be assessed by comparing the difference in US expectancy from Acq5 to Ext1 between the instructed and the uninstructed group. The effect of instruction on learning through experience will be assessed by comparing US expectancy slopes from the end of acquisition to the end of extinction between the instructed and uninstructed group.

We will calculate the described indices for both groups but expect interindividual differences in this index to be positively correlated with connectivity only in the instructed group. Thus, we will test for group differences in the slope of the correlation between behavioral indices and connectivity. We have now adapted the description of the proposed analyses accordingly. The section now reads (p.31):

"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 ($CS_{instructed.Ext1}$ – $CS_{instructed.Acq5}$) for the $CS_{increase}$ and $CS_{decrease}$ 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., $CS_{decrease}$ value x (-1))".

21. Regarding the pilot study (p. 30), the authors indicate that one participant was excluded due to poor data quality. It would be good to know how poor data quality is defined, given that this may happen for the proposed study is well and should perhaps be mentioned as ground for exclusion.

Eye tracking data of this one participant of our pilot study were excluded due to the occurrence of repeated sudden shifts in pupil size unrelated to the experimental paradigm (see Fig. 1, in red). Such anomalies were likely due to errors in the estimation of the pupil size (i.e., due to partially closed eyes, e.g., before or after blinks, drift in pupil position due to tiredness, or misidentification of non-pupil areas belonging to the pupil). We had already instructed participants to refrain from the use of eye makeup, which can be misspecified as belonging to the pupil due to black color. To avoid similar problems in the proposed study, we will instruct participants to keep their eyes completely open between regular blinks. Furthermore, we have now implemented a criterion to exclude trials with such sudden shifts. Using a sliding window approach, we will calculate the standard deviation (SD) of the preprocessed pupil size in time windows of 100 ms from -1 s before CS onset until 7.5 s after CS onset. If any SD exceeds 0.2, the respective trial will be excluded. The exclusion criterion for the phases (*"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."*) will also take these exclusions into account. We have now revised our manuscript accordingly. The changes in the manuscript on page 28 read as follows.

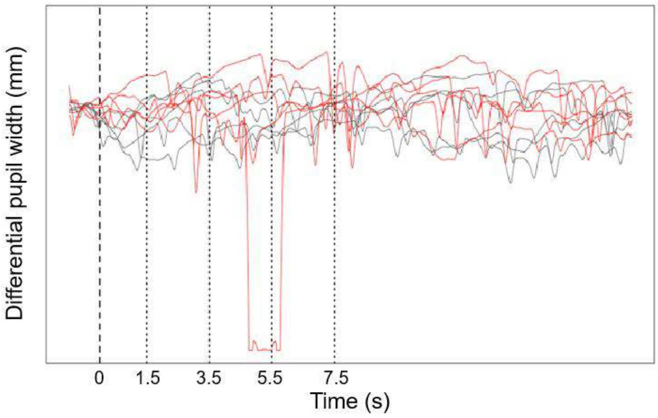

Fig. 1. Differential pupil width relative to a pre-stimulus baseline per trial is shown over time. The dashed line represents the CS onset, the dotted lines 1.5, 3.5, 5.5, and 7.5 s after CS onset, respectively. Trials excluded according to the criterion classifying sudden shifts in pupil size are shown in red, regular trials in gray.

"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."

22. While discussing the results of the pilot study, the authors indicate on p. 31 that, whereas the mean expectancy for the CSdecrease returned to zero the mean expectancy for the CSincrease remained elevated at the end of extinction. I'm not certain that that is clear from Figure 2, in particular when considered relative to responding to CSmedium.

We agree that the difference of US expectancy between CS types is more evident in Figure 4 and thus revised the discussion of the figures and respective interpretations accordingly. Please see the manuscript for highlighted changes on pages 33 and 35. We also agree that when considering the differential expectancy relative to the CS$_{\text{medium}}$, the conclusion would be different. Since the development of US expectancy ratings of the presented CS states individual learning processes for each CS type and thus, in contrast to CS valence, constitutes a more independent measure when comparing experimental conditions, we will analyze ratings individually by comparing correct expectancy of the CS$_{\text{increase}}$ and CS$_{\text{decrease}}$ directly.

Reviewed by Gaëtan Mertens, 06 Feb 2023 15:03

I've read the RR1 manuscript titled "Modulatory effects of instructions on extinction efficacy in appetitive and aversive learning: A registered report" by dr. Busch and colleagues carefully. In the RR proposal, the authors want to investigate the effects of verbal instructions on aversive and appetitive conditioning, using heat pain (or pain relief) stimulation. For several outcome measures (i.e., CS expectancy ratings, CS valence ratings, pupil dilation and skin conductance responses) the effects of instructed extinction and aversive vs. appetitive conditioning will be assessed.

All in all, I think this is a good research proposal on a topic that, due to the methods involved (e.g., psychophysiological measures), usually suffers from restricted sample sizes. Therefore, I believe it is a good thing that a well-powered RR will be conducted on this topic. Furthermore, the report is well-written and the authors seem experts on the involved methods. As such, I do not have many things to add.

We thank the reviewer for this positive evaluation of our proposed study and manuscript and the helpful comments.

My only more major comment is related to the different dependent variables (i.e., CS expectancy ratings, CS valence ratings, pupil dilation and skin conductance responses) and how differences in results between these DVs should be interpreted. What if the CS type X time interaction during acquisition or extinction is significant for one DV, but not for the authors. How should the hypothesis then be interpreted? I think that, in principle, the different DVs test the same hypothesis (e.g., steeper extinction slopes for appetitive than for aversive CSs). Therefore, I believe that correction for multiple testing should probably be applied. In that case, if a significant effect (using an adjusted alpha-level) is observed for any of the DVs, this effect can be interpreted as supporting the hypothesis.

We thank the reviewer for this valuable point which was also raised by the first reviewer (please see our reply to his point 4). We intend to focus on US expectancy ratings as the main outcome but added valence ratings and autonomic physiological data for a more comprehensive picture of the underlying effects of appetitive and aversive conditioning. Please note that the interpretation of our results will be based on the primary outcome of US expectancy ratings. In case of divergent results between the primary and secondary outcomes, we will base our main study conclusions on the results of the primary outcome variable. Please also note that correction for multiple comparisons is not deemed necessary if there is only one primary outcome measure (see Feise (2002) and Lonsdorf et al. (2017); in our case US expectancy).

The selection of outcome measures is now explained as follows on page 4 of the main manuscript:

"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)".

Smaller comments:

- Title: I struggle a bit with the terminology, because typically when considering appetitive conditioning, I think of things like pairing CSs with chocolate or erotic pictures (van den Akker et al., 2017). What the authors do in their paradigm seems more akin to relief learning (i.e., relief from a painful stimulus). I am not entirely sure whether this is the same thing as appetitive conditioning. However, I do not have good recommendations for the authors to change their terminology (except for maybe "pain" and "pain relief" learning).

We thank the reviewer for this comment and agree that, at first glance, the termination of an unpleasant event does seem to differ from an inherently appetitive event. However, in a previous study, we were able to show that pain relief can acquire positive valence (van der Schaaf & Schmidt et al., 2022), indicating that it functions as an appetitive event. We also decided on this terminology for consistency reasons as the same terminology has been used in previous studies using similar experimental designs (e.g., Mohr et al., 2008; Seymour et al., 2005; van der Schaaf & Schmidt et al., 2022).

- ***P. 4: I believe that Sevenster et al. (2012) observed a lack of effects of instructions on the startle response, rather than skin conductance responses. And even this interpretation is somewhat dubious, because in their follow-up tests, Sevenster et al. (2012) observed facilitated extinction in the instructed extinction group with startle as well. Indeed, the literature indicates mostly ubiquitous effects of verbal instructions with different DVs (Atlas & Phelps, 2018; Costa et al., 2015; Mertens et al., 2018; Mertens & De Houwer, 2016).***

We thank the reviewer for pointing this out. We have thoroughly rewritten and restructured the introduction and included the suggested references in the section that motivates the choice of primary and secondary outcomes.

The addition reads *"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)"*. Furthermore, as stated in our response to comment 2 from Tom Beckers, we agree that it cannot be concluded from the literature that instructed extinction is less effective on SCR than US expectancy.

- ***Fig. 1: Perhaps this figure can be a bit reorganized (particularly for any eventual publications) to make the size of the left panel figure larger (e.g., by putting the two images below one another, rather than next to each other).***

We revised the figure as suggested.

- ***P. 23: Regarding the covariate analyses including US painfulness ratings, gender, age, etc. Are these really needed? Perhaps in an exploratory sense, they could be interesting. However, adding them to the main analyses based on model improvement seems not needed and could complicated the interpretation of the results in my view. Particularly, due to randomization, any systematic effects of these covariates should be nullified. Furthermore, when effects are reported including the covariates, it may be hard to gauge for readers whether effects crucially depend on the inclusion of the covariates. Hence, for ease of interpretation, I would recommend to simply not include these covariates in the main analyses.***

We agree and would like to clarify that we intend to include these analyses for exploratory purposes only (compare van der Schaaf & Schmidt et al., 2022). For hypothesis testing we plan to rely on the basic model (i.e., without added potential covariates). In subsequent, exploratory steps we will test whether including the aforementioned variables may improve model fit. We do indeed not expect the respective variables to be relevant in terms of group differences but still test for the potential influence of several covariates of interest on learning in general. For example, pain-associated cognitions and personality traits have been linked to learning (Nees & Becker, 2018). Furthermore, gender differences have been discussed (e.g., Dalla & Shors, 2009; Jackson et al., 2006). Most importantly, we want to make sure that US painfulness (covariate of no interest) is not systematically related to changes in the outcome variables over time. Please see the sections covariates (in 2.2.3.1, pp. 18-19) and 2.5.2.1. (p. 26-29) for highlighted changes.

"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."

"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."

I am not an expert in fMRI analyses, so unfortunately, I could not really evaluate the appropriateness of the analyses. However, at first sight, I think the analyses seem appropriate and I believe that the preregistration of the analyses pipeline in this RR is very valuable, given the many degrees of freedom in analyzing fMRI datasets (Botvinik-Nezer et al., 2020).

We thank the reviewer for their positive evaluation.

References

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

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. Nature, 582(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

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

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

van den Akker, K., Schyns, G., & Jansen, A. (2017). Altered appetitive conditioning in overweight and obese women. Behaviour Research and Therapy, 99, 78–88. https://doi.org/10.1016/j.brat.2017.09.006

Reviewed by anonymous reviewer, 15 Feb 2023 15:09

This registered report submission outlines a study to answer the research question of whether verbal instruction may make extinction more efficient during appetitive (pain relief) than aversive (pain exacerbation) learning, additionally investigating associations of learning and extinction indices with pre-task resting-state fMRI connectivity between regions-of-interest and the rest of the brain. Majority of the submission is very clearly written and extremely thorough. In most aspects, it is an excellent registered report for a well-planned study. I am satisfied with the behavioral and psychophysiological section of the submission and would accept the submission almost as-is regarding the parts intended for Manuscript 1, with only minor points to address. However, I have some central criticisms pertaining to the research questions, theoretical background and hypotheses posed in the resting-state fMRI section intended for Manuscript 2, which I think should be resolved before acceptance.

We thank the reviewer for their positive and constructive evaluation and the comments raised which have helped to further improve our manuscript.

1A. The scientific validity of the research question(s)

- ***The scientific validity of the stated research questions for the non-MRI part of the submission intended for Manuscript 1 is good and I have no comments on the research questions themselves. There are only a couple minor things I would like to point out about the background provided:***
    - ***Introduction, p. 4, paragraph on common neural systems for aversive and appetitive learning mechanisms: all literature used to support claim "Together, these studies suggest a common neural system for appetitive and aversive learning mechanisms" used only secondary (monetary) reinforcer and not primary appetitive reinforcer such as the pain relief in this study plan. Especially with talk of "biological relevance", I think this is a rather important distinction and should be a qualifier (and otherwise, what is this paragraph trying to say?). There is also a rather extensive animal literature on the neural basis and/or commonalities of aversive and appetitive learning, which is not cited at all.***

We thank the reviewer for pointing this out. We revised the introduction accordingly and have also included the suggested literature on animal research, as well as on pain relief as the appetitive US. The paragraph now includes (p. 5):

"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".

We furthermore agree with the reviewer that similar to the behavioral manuscript, the novelty of our design is that it allows the direct within-modality comparison of appetitive and aversive effects (compare van der Schaaf & Schmidt et al., 2022).

- ***The research question for resting-state fMRI component intended for Manuscript 2 is written as follows (p. 7, Introduction): "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, and the effect of the instruction". This research question is not proposed well in the submission in relation to existing theory and how relevant the question may be for the field.***
    - ***Firstly, the theoretical background for the research question is not presented in a convincing manner. For example, p. 4 of Introduction, the sentence "Changes in resting-state functional connectivity of the amygdala after acquisition learning have been reported (Schultz et al., 2012)" does not specify what kind of changes were found. Why would it be relevant that there is any change in functional connectivity? Moreover, the sentence "Connectivity of the amygdala has been shown to be clinically relevant for the prediction of treatment outcome (Klumpp et al., 2014)" does not specify what kind of connectivity of the amygdala and with what other brain region/network, and what clinical condition and treatment were involved, and why this is relevant for the research question at hand. Finally, "Individual aversive acquisition learning (Kincses et al., 2023), and extinction learning (Belleau et al., 2018) were associated with brain connectivity, and connectivity changes" is extremely unspecific. Associated how to what indices of acquisition of extinction learning (the strength or speed of learning, or something else?), what brain connectivity, and what connectivity changes?***
    - ***Secondly, it is obvious that any kind of learning is associated with changes in the brain so it would be important to try to device studies that can precisely answer questions such as "associated how", "what kind of changes", "where exactly do the changes occur" and "are these changes actually important for the learning". It is not clear from the submission how answering the above research question would advance our understanding of the main phenomenon under study, i.e. influence of instructions on extinction learning in the context of conditioned expectation of pain exacerbation and relief, and their neural mechanisms.***

We thank the reviewer for their comments. We have now thoroughly rewritten the introduction on resting state connectivity to provide the theoretical background and rationale for the proposed analyses and to clarify the points raised. Furthermore, we have now clarified our rationale and the theoretic impact such a finding might have. We would like to point out that we do not intend to investigate changes in connectivity (i.e., from pre- to post-learning) but are interested in the relationship between resting state connectivity assessed prior to the behavioral experiment and behavioral markers of pain-related learning (for acquisition and extinction) and instruction effects.

To date, only a few studies have investigated rsfMRI in relation to fear conditioning including acquisition and extinction. Based on the finding that regions which coactivate during a task tend to also coactivate during rest (Smith et al., 2009), we selected ROIs such as the amygdala and PFC for the connectivity analyses that have shown task-based engagement related to conditioning, instructions, and pain (e.g., Fullana et al., 2016). We believe that our findings could be used to develop methods to estimate the individual benefit from exposure or instruction-based therapeutic approaches. Please see the introduction of our manuscript for highlighted changes (pp. 5-6, and 7-8).

"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."

"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. [...] 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.*"*

1B. The logic, rationale, and plausibility of the proposed hypotheses (where a submission proposes hypotheses)

- ***The hypotheses for Manuscript 1 are specified well and linked to a sound theoretical background.***
  - ***Minor point. In Table 1, last row, column "Theory that could be shown wrong by the outcomes" (p. 9-10), it is stated that "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), which would be in line with single-process accounts of fear learning (e.g., Brewer, 1974; Mitchell et al., 2009), which suggest a common basis for affective and expectancy learning". Even if the outcome may contrast with the evidence from previous studies reviewed in Luck and Lipp (2016), I am confused as to why the authors think that CS valence being affected by instructions would support the single-process account (and therefore, provide evidence against the dual-process account) of fear learning. The arbitration between single- and dual-process accounts hinges on whether fear learning in humans relies on forming conscious expectations of aversive outcomes (measured via US expectancy ratings), or whether it can manifest in two independent learning processes, where in addition to conscious contingency learning there is lower-level learning (usually considered to be reflected in physiological CR) that may be outside conscious awareness and possibly not influenced by verbal instructions. Unless the authors argue that CS valence ratings can be taken as an example of the latter type of learning even if rating CS valence involves explicit reporting and therefore consciously accessing valence representations, which I would contend with, the outcome of this analysis does not in fact offer any unambiguous support for the single-process account. Of course, I may have misunderstood the argument but in that case, I would ask the authors to clarify it in the text.***

We thank the reviewer for this very helpful comment which prompted us to reconsider our interpretation of the dual-process theory. We agree that the susceptibility of CS valence ratings to instructions would not be sufficient to reject the dual-process account. We therefore adapted the point accordingly in Table 1.

- ***Since the research question posed in the resting-state fMRI part of the submission is not well-specified and/or grounded in theory, the associated hypotheses are also vague. The hypotheses are written in the form: "significant associations of an effect of interest (e.g., acquisition index, effect of instructions on extinction efficacy) with the resting-state connectivity between seed ROIs (left and right dlPFC and vmPFC, amygdala, and striatum), and the rest of the voxels in the brain".***
    - ***Significant association as per which metric precisely, and in which direction? E.g., an entirely hypothetical example for a directional, more precise hypothesis: higher functional connectivity defined as Pearson's correlation (or other measure) between amygdala and vmPFC (or a specific resting-state network, e.g., the salience network) measured before task is expected to be associated with higher extinction efficiency during the task.***
    - ***If specific hypotheses are not justified by previous literature/existing theory, this part of the submission should be introduced as highly exploratory, and the final form of Manuscript 2 should also reflect this.***
    - ***Additional minor point for H1+2 of Manuscript 2 (p. 6, from line 24): Each individual's acquisition and extinction of CR quantified only as the slope of US expectancy, excluding the other measured (e.g. SCR, pupil size). Since this is presumably has to do with the intent of studying the impact of verbal instruction, it would be good to state clearly that US expectancy is used here to quantify the CR as it is likely the measure most influenced by verbal instruction.***

We thank the reviewer for raising these points. We based our hypotheses and the selection of ROIs for the analysis of rsfMRI data on task-related findings since Smith et al. (2009) had found very similar brain regions associated with cognitive processes during rest and task-based fMRI. Furthermore, we applied the reasoning by Hashmi et al. (2014), who assumed preparatory activity in the respective regions. In the scope of this registered report, we rely on this "simplification" to restrict model space. At a later stage, a whole brain predictive modelling approach could also be performed to systematically assess resting state correlates. This approach was not included in the current registered report due to the exploratory nature of this method, and to keep the extent of the work program within reasonable limits.

We have revised the wording of our hypotheses according to the reviewer's suggestions. We will use Pearson's correlations (bivariate correlation is the default in CONN's first-level analysis) and expect positive associations between behavioral indices and connectivity (for the instruction hypotheses particularly in the instructed group). We focused on US expectancy as the index of learning since it is our main outcome in the behavioral experiment. Furthermore, we agree that at least in studies on instructed extinction but not instructed reversal, US expectancy has been the most robust outcome. Please also note our response to question 5 of reviewer Tom Beckers.

"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). 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."

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable)

- ***The methodology and analysis pipeline for the behavioral and psychophysiological analyses for the analyses intended for Manuscript 1 seem sound and feasible.***
    - ***Minor point for Table 1, first row (manipulation check US type), column "Interpretation given different outcomes": "Relevant effect: A statistically significant main effect of the factor US type, indicating that ratings are higher for the USincrease than the USmedium, and higher ratings for the USmedium than the USdecrease …" – To be precise, a main effect does not in fact indicate that USincrease > USmedium > USdecrease. It only indicates that there is a significant mean difference overall between at least some of the levels of this factor. I am sure the authors know this as post-hoc comparisons are mentioned elsewhere but the statement here should be corrected.***

Table 1 (p. 9) has been revised accordingly. Please see highlighted changes.

"Relevant effect: A statistically significant main effect of the factor US type, with post-hoc tests indicating that ratings are higher for the US$_{increase}$ than the US$_{medium}$, and higher for the US$_{medium}$ than the US$_{decrease}$ during acquisition would suggest successful manipulation."

- ***I cannot comment precisely on the details of the resting-state analyses intended for Manuscript 2 as I am not an expert in resting-state fMRI. The outlined preprocessing steps seem sound for fMRI data analysis in general.***

We thank the reviewer for this positive assessment.

- ***Since the sampling is based on Bayes Factor plus maximal sample size as stopping criterion, either the minimum sample size to be collected should also be defined (and be substantial enough to mitigate the issue of possible false positive evidence when reaching the evidence threshold after only very few participants since "most misleading evidence happens at early terminations of a sequential design") or a very high evidence threshold used (e.g. BF10 ≥ 30; see Schönbrodt & Wagenmakers, 2018, section "Sequential Bayes factor with maximal n: SBF+maxN"). Note that if the authors want to be able to claim absence of evidence, i.e. support for null hypothesis, a Bayes Factor stopping criterion for H0 should also be set (as the sample size needed to reach strong enough evidence for H1 and H0 can be different). Moreover, explaining the protocol of data checking for whether stopping criterion is fulfilled should be included in the submission.***

We thank the reviewer for pointing this out. As suggested, we incorporated a stopping criterion for H0 (please see also our reply to question 6 from Tom Beckers). We furthermore specified a minimum sample size (*N* = 80) and state that we decided to assess whether BF10 has been met after every 10th participant. Once the stopping criterion has been reached, all participants who have already undergone scanning will complete the study and will be included in the final analyses in accordance with the inclusion criteria. Please see the Methods section 2.1. Participants on page 16 for changes.

"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 (BF_{10}) in favor of our (main) hypotheses (i.e., H1-4a) reach BF_{10}>6 or BF_{10}<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."

  - ***Possible minor mistake in the sampling plan in Table 2, first row, p. 11: Is it correct here that 75 participants per group mentioned, or is full N = 150 used for these analyses without group separation?***

We clarified this in the table (*"Acquisition max. N = 150, extinction max. n = 75"*). For the acquisition training, we will use the full sample, while for the extinction training, we will only use the uninstructed group since we focus on the extinction effect independent of instruction in this analysis.

1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses

- ***The methodological detail is largely excellent. The exceptions to this are:***
  - ***Section 2.4.3.2. Acquisition training (p. 20, line 20): It would be good to refine the description here a bit more to unambiguously state that the participants received only instruction about the existence of contingencies but not of the actual contingencies themselves (and therefore, participants had to learn the contingencies during acquisition through experience), to avoid***

> ***misunderstanding since this was an instructed conditioning study but (as far as I understood) the actual contingency instruction was only given for the extinction phase.***

In response to the reviewer's point, we clarified our description (p. 24), which now reads *"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 actual contingency instruction will indeed only be provided for the extinction training (i.e., only the instructed group). No group-specific instruction or detailed instructions on contingencies will be given prior to the acquisition training. Instruction will therefore only differ between groups prior to the extinction training.

- ***Section 2.5.2.2. Pupillometry data (p. 25, line 10-11): "We will apply a correction to account for multiple comparisons". What multiple comparisons correction will be used and to what tests will it be applied? This should be mentioned for all analyses.***

We would like to clarify that we intend to use Bonferroni-Holm correction for multiple testing in pupillometry analyses of the respective experimental phases (i.e., correction will be applied for analyses including bin as a time factor within the trial and analyses using only the selected bin, i.e., the bin indicating the largest differentiation). This has been added to the manuscript on page 29:

"We will apply Bonferroni-Holm correction to account for multiple comparisons within the respective experimental phases".

Please also note question 1 from Gaëtan Mertens regarding multiple comparisons in the further analyses. Specifically, a correction for multiple comparisons is not required if one primary measure (here US expectancy) is specified (Feise, 2002; Lonsdorf et al., 2017).

- ***The resting-state fMRI analysis section does not detail how the correlation maps are obtained with CONN toolbox: "… first-level analyses in the CONN toolbox. We will use the rsfMRI scan to derive correlation maps between the respective seeds and all other brain voxels". For those readers who are not experts in using the specific toolbox, it is not evident at all how the analysis is done, what options might be used, etc. Presumably, there are degrees of freedom to how these analyses can be conducted.***

We agree that this part of our data analysis indeed requires more detailed information. We specify that the preprocessing will be performed using the standardized fMRIPrep processing pipeline (Esteban et al., 2019, 2020) which offers very limited degrees of freedom. Final steps prior to the first-level analysis will be carried out in the CONN toolbox, i.e., smoothing and a nuisance regression for denoising. For the first-level analysis, we will select "seed-to-voxel analyses only", with the default options "correlation (bivariate)" and "no weighting". Please see section 2.5.4.1. on pages 30-31 for highlighted changes.

"SBFC analyses will use the left and right dlPFC and vmPFC, amygdala, and striatum as seeds, with masks derived from the FSL Harvard-Oxford Atlas (Desikan et al., 2006), for first-level analyses in the CONN toolbox. 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 mm^3 voxels, original: 2.5 × 2.5 × 2.5 mm)."

- ***It is not clear why the authors chose the specific threshold for probabilistic threshold-free cluster enhancement z-score image false discovery rate (q < .02).***

In response to the reviewer's point, we would like to clarify that probabilistic threshold-free cluster enhancement (pTFCE) is a method to enhance the detectability of neuroimaging signal by performing topology-based belief boosting, i.e., it integrates cluster size information into voxel-wise statistical inference. This method provides enhanced p-values that are then used for thresholding using FDR to control for multiple comparisons. The threshold (q < .02) for FDR, i.e., the percentage of features called significant that are truly null, was chosen to meet the eligibility criteria for a submission to Cortex (compare https://rr.peercommunityin.org/about/pci_rr_friendly_journals#h_4920688494031618419330727). Please see sections 2.5.4.2. and 2.5.4.3. on pages 31-32 for highlighted changes.

1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).

- ***Yes. Very minor point: I am sure the authors intend to do this as part of quality checks even if it was not mentioned, but in addition to checking the US type effect for US painfulness and US unpleasantness ratings, it should also be checked that there is a reliable US response for SCR and pupil size. The manipulation check would be USincrease > USmedium for the aversive side, while USdecrease > USmedium is possible in the appetitive case (in contrast to USmedium > USdecrease for the rating measures) due to the valence-independence of SCR and pupil size.***

We thank the reviewer for their valuable suggestion, which we included in the design table. We agree that this is a critical test for the validity of data collection and the painfulness manipulation regarding the US.

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

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

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

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

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

Dalla, C., & Shors, T. J. (2009). Sex differences in learning processes of classical and operant conditioning. *Physiology & Behavior*, *97*(2), 229–238. https://doi.org/10.1016/j.physbeh.2009.02.035

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

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*, *2*(1), 8. https://doi.org/10.1186/1471-2288-2-8

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

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

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

Jackson, E. D., Payne, J. D., Nadel, L., & Jacobs, W. J. (2006). Stress Differentially Modulates Fear Conditioning in Healthy Men and Women. *Biological Psychiatry*, *59*(6), 516–522. https://doi.org/10.1016/j.biopsych.2005.08.002

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

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

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

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

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

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

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

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

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

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

Mohr, C., Leyendecker, S., Mangels, I., Machner, B., Sander, T., & Helmchen, C. (2008). Central representation of cold-evoked pain relief in capsaicin induced pain: An event-related fMRI study. *Pain*, *139*(2), 416–430. https://doi.org/10.1016/j.pain.2008.05.020

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

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

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

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after P-Hacking. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2205186

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

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

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

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

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