Point-to-point Reply

We would like to thank the recommender and the reviewers for their useful suggestions and provide below a detailed response to each comment. Please note that the reviewers' comments are in black with our reply underneath in blue and passages of the manuscript further in italics and underlined.

Reviewer 1 (Yoann Stussi):

The stage-1 registered report by Bruntsch et al. details the rationale and a protocol for conducting a meta-analysis of the associations between anxiety-related traits and threat acquisition and extinction during Pavlovian conditioning. The relevant studies will be identified with a systematic literature search and these associations will be assessed by means of a series of nested random-effects models using self-report anxiety-related questionnaires and psychophysiological and self-report measures of threat acquisition and extinction. Expected results are that there is a positive association between anxiety-related traits and threat acquisition and extinction. Moderator analyses will be also conducted to explore the effects of different anxiety-related questionnaires and methodological characteristics (e.g., reinforcement rate, clinical vs. non-clinical groups, type of outcome measures, etc.). Finally, an item-based content analysis will be performed to examine the overlap in content between anxiety-related questionnaires, with a limited content overlap being expected.

The registered report is very clear and well documented. The additional files and documents openly available on the associated Open Science Framework repository are clear and comprehensive. The main research question and secondary research questions are scientifically valid and well justified. Based on existing meta-analyses (e.g., Duits et al., 2015; Morriss et al., 2021), the hypothesis of a positive—albeit relatively modest—association between anxiety-related trait and threat acquisition and extinction is logic, well-motivated, and plausible. The minimal number of studies required to be included in the meta-analysis is clearly justified with an a priori power analysis based on related but more specific meta-analyses (Morriss et al., 2021; Sep et al., 2019). The methodology and analytic plan appear sound and feasible. The literature search and meta-analysis components are rigorously described with a high level of detail. The analytic pipeline additionally includes meta-analytic techniques to control for publication bias and quality checks are also considered, which contributes to increasing the chances of the meta-analysis to provide informative findings and calibrated interpretations. Overall, my evaluation is that this work has great potential to offer a highly informative and worthy contribution to the study of the links between individual differences in anxiety-related symptomatology and threat conditioning processes. I have a few comments that may be worth considering to strengthen this registered report even further. Specifically, I wonder about (a) the comprehensiveness of the meta-analytic methods used and decisions made to control for publication bias and (b) whether the nature of the conditioned (CS) and unconditioned (US) stimuli should be considered as another moderator in the moderator analyses. I describe these comments below along with some other minor points. I hope they will be constructive and helpful.

Primary comments

1) It may be particularly beneficial to consider further statistical approaches—such as trim-and-fill, precision-effect test (PET), precision-effect estimate with standard error (PEESE), PET-PEESE, p-uniform—that aim to correct for potential publication bias in addition to the three-parameter selection model and the p-curve analysis. Because these various meta-analytic methods perform differently under various conditions (i.e., presence/absence of questionable research practices, publication bias, heterogeneity; see, e.g., the simulation work by Carter et al., 2019), the comprehensive inclusion of these techniques may contribute to more finely establishing the sensitivity of the associations between anxiety-related traits and threat acquisition and extinction. This could in turn allow for more calibrated and nuanced interpretations of the findings. If the authors prefer to exclusively use the three-parameter selection model and p-curve analysis, it would be important to provide a thorough justification of why these methods were selected over other available methods.

Reply 1.1: Thank you for these valuable suggestions. In our revised Stage 1 RR, we now also include the PET, PEESE and PET-PEESE approach into our analysis, as suggested. The paragraph we added for these methods can be found below. However, we are not planning to include the p-uniform approach as the only difference between this and the p-curve approach is the use of different implementations of the estimation algorithm and both seem to perform equally well based on the literature (Carter et al., 2019). However, we are going to include the z-curve as an additional approach as it is suggested to perform better under conditions of effect size heterogeneity (Brunner & Schimmack, 2020). Further, we are hesitant to include the trim-and-fill method as it has been reported to perform quite poorly (Carter et al., 2019) and because it has been recommended to not use it at all for meta-analysis (http://datacolada.org/30). We modified the text accordingly:

Page 22: "Moreover, to adjust for small-study effects, the meta-regression techniques precision-effect test (PET), precision-effect estimate with standard error (PEESE), and the combined PET-PEESE will be employed (Carter et al., 2019). While PET outperforms PEESE when the true effect is zero, PEESE outperforms PET when a true effect is present. The PET-PEESE method aims to balance out these opposite biases by combining them (Stanley & Doucouliagos, 2014). All three methods will be applied to models generated using the function rma() of the metafor package (Viechtbauer, 2010)."

Page 22: "In addition, a z-curve analysis will be conducted. This method, described as an improvement and extension of the p-curve analysis, is suggested to provide more accurate estimates under conditions of effect size heterogeneity (Brunner & Schimmack, 2020). The approach converts p-values into Z-scores to integrate results from different studies and then analyzes the distribution of these z-values. The calculations will be performed using the R package zcurve (Bartoš & Schimmack, 2020)."

2) Relatedly, I was unsure whether the benefits of the decision to only include published and peer-reviewed studies in the meta-analysis outweigh its costs. Whereas I understand that incorporating unpublished data (e.g., via calls to relevant societies and mailing lists as well as dissertations and theses databases such as ProQuest; https://www.proquest.com) would require a significant additional amount of work, this could potentially help provide a more

accurate effect-size estimate of the associations between anxiety-related traits and threat learning and extinction that is less likely to be overinflated by publication bias. The inclusion of unpublished data was notably done in Duits et al.'s (2015) meta-analysis. If the authors still prefer to only include published and peer-reviewed studies, it would be extremely beneficial to provide a more elaborated cost-benefit analysis for this decision that would carefully balance the benefits associated with it (e.g., less time-consuming, no potential fluctuations of effect sizes) and the risks (e.g., risk of higher publication bias and overestimated effect sizes).

Reply 1.2: This has been discussed a lot in our group and while we see the advantage of including grey literature in general, we would like to restrain our literature search to only peer-reviewed articles and preprints, but are not planning to include other kinds of grey literature (e.g., dissertations). We are aware that this might result in a higher publication bias and potentially inflated effect sizes. Still, we think that this is the most doable approach for our project due to several reasons. (1) We are not aware of any systematic way to assess all kinds of grey literature and therefore fear missing important studies and presenting an incomplete picture of the grey literature and therefore bias the results in a different way. Further, (2) including all sorts of grey literature would lead to significantly more time and work that is needed to conduct this study as a lot of studies will probably not report effect sizes. In turn, these would then be calculated by us and this can just be realized if we have all the necessary information. If this should not be the case, we would need to contact the authors which would result in even more time and maybe no response at the end - in particular when it comes to dissertations and student projects. Lastly, (3) adding all kinds of grey literature might lead to a higher number of small sample sizes and would therefore also bias the meta analytic results.

To also monitor if any changes appeared in the included preprints, we decided that we will code if the studies are a preprint or a published article and check before conducting the meta-analysis if there occurred any changes in the analysis or reported effects of the preprints.

This modified procedure has also been explained in the manuscript:

Page 12: "Published or peer-reviewed studies as well as preprints will be included. Any other kind of grey literature (e.g., dissertations or conference abstracts) will not be included, while being aware that this comes with the risk of a potential overestimation of the effect."

3) The moderator analyses do not consider potential effects related to the CS and US nature. Nonetheless, these factors exert a powerful influence on threat acquisition and extinction (see, e.g., Rescorla & Wagner, 1972). For instance, threat-relevant or affective relevant CS have been shown to induce faster threat acquisition and enhanced resistance to extinction during Pavlovian threat conditioning (e.g., Öhman & Mineka, 2001; Stussi et al., 2018, 2021), while electro-tactile USs have been reported to elicit stronger physiological conditioned responses than loud screams (e.g., Ney et al., 2023). Although I'm not aware of any robust evidence that the CS and US nature moderate the associations between anxiety-related traits and threat acquisition and/or extinction, addressing this question may provide valuable information (provided there are enough studies investigating these aspects that can be included in the meta-analysis). Based on the proposal that threat-relevance (or

"preparedness") is a key factor in the development of phobias (Öhman & Mineka, 2001; Seligman, 1971), this may have especially relevant translational implications for clinical research and applications using threat conditioning as a model for the etiology and maintenance of anxiety-related conditions. Please note, however, that this is merely a suggestion and that I'm not requesting the the CS and the US nature to be included as moderators in the meta-analysis.

Reply 1.3: We thank the reviewer for this valuable suggestion and added fear relevance of the CS and different types of US as moderators. Therefore, we will extract if the CS has been threat-relevant, of emotional valence or neutral in the selected studies based on the suggestions of Öhman & Mineka (2001), and Stussi et al. (2018, 2021). Furthermore, we will assess what type of US has been used (electrotactile vs. "other" as an umbrella category). The information will be coded in column AH and AN of the coding template. We also added these additions in the manuscript as follows:

Page 23: The following categorical variables will be included as potential moderators in separate subgroup analyses (provided that sufficient data for the analyses can be extracted): type of questionnaire assessing anxiety-related traits (e.g., Intolerance of Uncertainty scale vs. STAI-T vs. Big Five Neuroticism scale), group specifics (non-clinical vs. clinical vs. at risk vs. mixed), sampling processes (median-splits vs. extreme groups), outcome measure type (ratings vs. SCR vs. FPS vs. HR vs. PD), ***relevance of the CS (threat relevance vs. emotional valence vs. neutral), US type (electrotactile vs. "other"),*** *reinforcement rate, continuous vs. discrete operationalization of the anxiety-related trait, effect size type and study quality.*

Secondary comments

4) In the introduction (p. 4), I believe it could be worth referring to more recent work linking threat conditioning processes to the etiology and maintenance of anxiety-related conditions (e.g., Beckers et al., 2023; Zinbarg et al., 2022). This may contribute to further highlighting the translational relevance and timeliness of threat conditioning as a laboratory model for anxiety-related disorders, which might not necessarily be obvious to audiences outside of the field.

Reply 1.4: Thank you for this useful suggestion. We have now included these references according to the reviewer's suggestions.

5) I'm not sure I fully understood the decision to use the mean of reported effect size from two prior meta-analyses as an effect size estimate for the power analysis. Whereas I understand that meta-analyses often provide a good basis for effect-size estimates, it has been suggested that the lowest available or meaningful effect size be used to adjust for publication bias (see https://rr.peercommunityin.org/PCIRegisteredReports/about/full_policies#h_6720026472751613309075757, point 3.3). A more conservative approach could thus be to use the minimal effect size (g = 0.22) reported in Morriss et al. (2021) as an effect-size estimate for the power analysis. Based on the power curve plot (which is a great addition) and Duits et al.'s (2015) meta-analysis, I'm confident that the minimal number of studies (n = 23-24) that need be

included to achieve sufficient statistical power to observe an effect size of g = 0.22 for the primary analysis will be reached.

Reply 1.5: Thank you for this valuable suggestion. Based on this comment, we changed our estimated effect size from the mean of all the reported significant effect sizes to the smallest effect size in Morriss et al. (2021) of 0.28 - which was obtained for the association between the IU-12 and SCR difference scores during the whole extinction phase.

With the estimated effect size of 0.28, 13 studies are needed to achieve sufficient statistical power. We have also marked in our plot the lower bound of our estimated effect size that we expect to include the 'true' effect size and is therefore of special interest to us. Accordingly, we changed the passages in the text (see below), updated the plot and modified the R script on OSF (https://osf.io/huq69).

Please note that the smallest effect size in Morriss et al. (2021) is 0.22, which is, however, derived from an effect that refers to a subscale of the Intolerance of Uncertainty questionnaire. We decided to not consider this as the minimal effect size because of the more narrow focus of the underlying analysis.

Page 16: "To determine how many studies are necessary for sufficiently powered meta-analytic models (ß >= 80%), an a priori power analysis solely for the primary meta-analysis of associations between anxiety-related traits and fear acquisition/ extinction has been conducted based on current recommendations (see Power Analysis Section in Harrer et al., 2022). ***For our target effect size, a Hedges' g of 0.28 (small-to-medium) was chosen based on the smallest effect size of relevance reported in the two prior conducted meta-analyses (Sep et al., 2019; Morriss et al., 2021).*** *This chosen effect size represents the association of the whole self-reported IU-12 questionnaire (i.e., not a select subscale) and SCR differences scores during the entire extinction phase (Morriss et al., 2021) .*

A primary power analysis was performed using this estimated Hedge's g as the target effect size and α = 0.05. Further, an expected mean sample size of 33 individuals per group (under the assumption that "high" and "low" anxiety-trait groups will have equal sample sizes) has been assumed based on a review reporting a mean number of participants of 66.2 for studies on individual differences in fear conditioning (Lonsdorf & Merz, 2017). This analysis revealed a minimum number of ***13 studies*** *is needed to achieve ß >= .8 for the to-be-conducted meta-analysis for acquisition and extinction."*

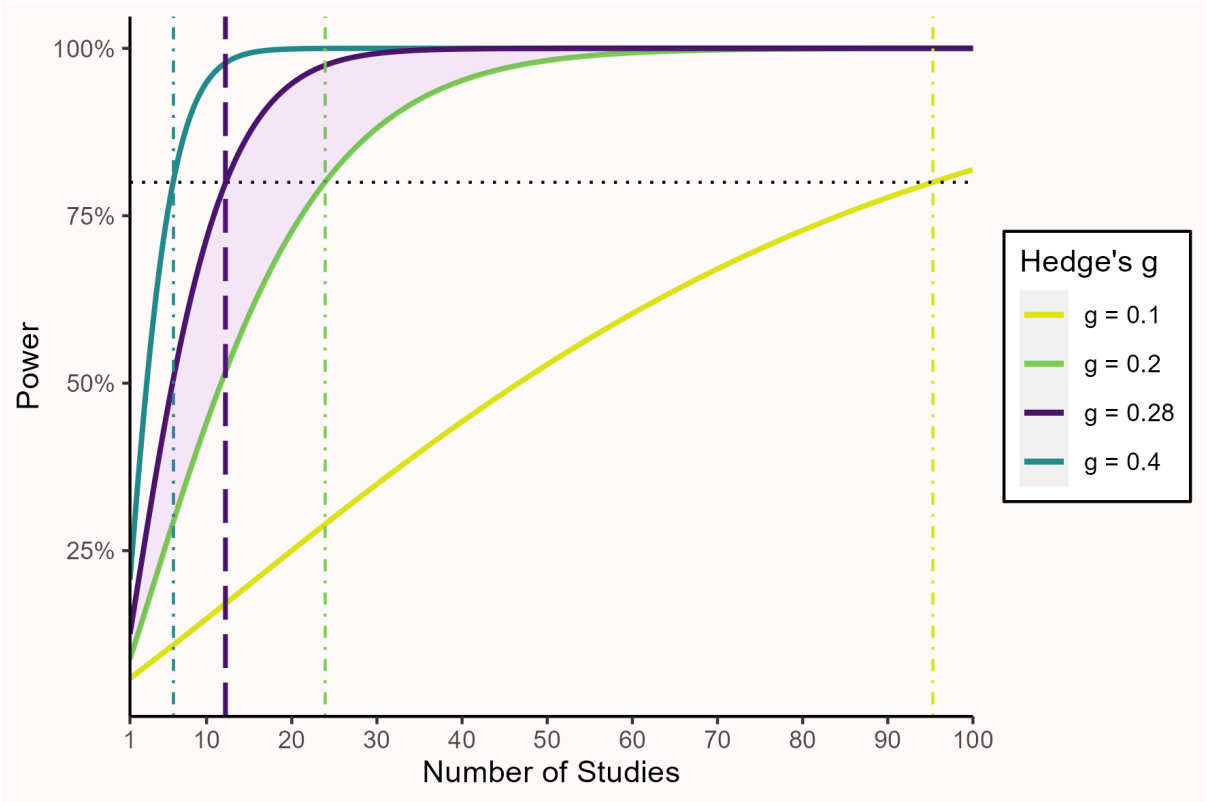

Figure 1

The Figure illustrates the power of the random-effects meta-analyses with different numbers of studies (n=66 in each study). The x-axis demonstrates the number of studies included in the meta-analysis scaling from 1 to 100. The y-axis shows the expected power (0-100%). The four colored lines represent power curves given different assumed Hedges' g values: 0.1, 0.2, 0.316 (the estimated effect size) or 0.4. These effect sizes correspond to the following approximate Pearson r values: 0.05, 0.1, ***0.139*** *and 0.196 (calculated with the d_to_r function of the effectsize R package (Ben-Shachar et al., 2020)).* ***The purple area between the graph for our estimated effect size and g = 0.2 shows the lower bound estimate which we are focusing on.*** *The horizontal dotted line marks a power of 80% while the vertical dotted line indicates the number of studies required to achieve ß >= .8.*

6) On page 16, "in association to anxiety-related traits" should be "in association with anxiety-related traits".

Reply 1.6: Thank you. We have now corrected this.

7) In the Study design table, under the column "Theory that could be shown wrong by the outcomes" and for the question "Do effects differ if different questionnaires have been used to assess anxiety-related traits?", it would be important to condition the proposed interpretation on (a) the finding of an overall positive association between anxiety-related traits and threat acquisition and/or extinction, and (b) the confidence interval around the effect size for the moderator effect of the questionnaires allowing to determine whether the observed effects are consistent with an absence of effect (if the confidence interval is

narrow) or a larger range of effects (if the confidence interval is wide). Without these conditions, I do not think it would be possible to unequivocally interpret the absence of statistically significant moderator effect of questionnaires as "support for an overall anxiety-related trait as potentially more explanatory of fear acquisition and extinction learning than any single questionnaire."

Reply 1.7: Thank you for this comment. We agree that the claim might have been too strong which we stated in the study design table. Therefore, we deleted this claim from the column "Theory that could be shown wrong by the outcomes" as we can only conclude that among the moderators we tested in our analysis none of them shows a moderating effect and does not directly speak in favor of an overall anxiety-related trait as a better explanation than any single questionnaire. We added this weaker interpretation under the column "Interpretation given different outcomes". You can find the modified study design table either at the end of the manuscript or on OSF (https://osf.io/snqgr).

REFERENCES

- Beckers, T., Hermans, D., Lange, I., Luyten, L., Scheveneels, S., & Vervliet, B. (2023). Understanding clinical fear and anxiety through the lens of human fear conditioning. Nature Reviews Psychology, 2, 233-245. https://doi.org/10.1038/s44159-023-00156-1

- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. Advances in Methods and Practices in Psychological Science, 2(2), 115-144. https://doi.org/10.1177/2515245919847196

- Duits, P., Cath, D. C., Lissek, S., Hox, J. J., Hamm, A. O., Engelhard, I. M., van den Hout, M. A., & Baas, J. M. P. (2015). Updated meta-analysis of classical fear conditioning in the anxiety disorders. Depression and Anxiety, 32(4), 239-253. https://doi.org/10.1002/da.22353

- Morriss, J., Wake, S., Elizabeth, C., & van Reekum, C. M. (2021). I doubt it is safe: A meta-analysis of self-reported intolerance of uncertainty and threat extinction training. Biological Psychiatry Global Open Science, 1(3), 171-179. https://doi.org/10.1016/j.bpsgos.2021.0.011

- Ney, L. J., Nichols, D. S., & Lipp, O. V. (2023). Fear conditioning depends on the nature of the unconditional stimulus and may be related to hair levels of endocannabinoids. Psychophysiology, 60(8), Article e14297. https://doi.org/10.1111/psyp.14297

- Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. Psychological Review, 108(3), 483-522. https://doi.org/10.1037/0033-295X.108.3.483

- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prosky (Eds.), Classical conditioning II: Current research and theory (pp. 64-99). Appleton-Century-Crofts.

- Seligman, M. E. P. (1971). Phobias and preparedness. Behavior Therapy, 2(3), 307-320. https://doi.org/10.1016/S0005-7894(71)80064-3

- Sep, M. S. C., Steenmeijer, A., & Kennis, M. (2019). The relation between anxious personality traits and fear generalization in healthy subjects: A systematic review and meta-analysis. Neuroscience and Biobehavioral Reviews, 107, 320-328. https://doi.org/10.1016/j.neubiorev.2019.09.029

- Stussi, Y., Pourtois, G., Olsson, A., & Sander, D. (2021). Learning biases to angry and happy faces during Pavlovian aversive conditioning. Emotion, 21(4), 742-756. https://doi.org/10.1037/emo0000733

- Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. Journal of Experimental Psychology: General, 147(6), 905-923. https://doi.org/10.1037/xge0000424

- Zinbarg, R. E., Williams, A. L., & Mineka, S. (2022). A current learning theory approach to the etiology and course of anxiety and related disorders. Annual Review of Clinical Psychology, 18, 233-258. https://doi.org/10.1146/annurev-clinpsy-072220-021010

Reviewer 2 (anonymous):

This registered report describes a meta-analysis of fear conditioning (acquisition, extinction) and anxiety-based individual differences. I think that the justification is sound, the proposed methods are rigorous and the topic worthwhile. I am particularly impressed by the proposed rigour of the study. I have only a few minor comments below to make on this draft report:

1) The first sentence of the introduction states “Anxiety-related conditions such as generalized anxiety, obsessive-compulsive, or post traumatic stress disorders”, but (1) PTSD is no longer considered to be an anxiety disorder under the DSM and (2) PTSD symptomology is not considered in this meta-analysis. Suggest removing PTSD from the first sentence as it wrongly sets expectations for what the meta-analysis will include

Reply 2.1: We thank the reviewer for this comment. Fear conditioning has been considered an experimental model of both anxiety- and stress-related (e.g., PTSD) pathologies and hence we think it is meaningful to include PTSD as an example. We, however, agree that PTSD is considered a stress-related disorder and not an anxiety disorder and have revised this sentence accordingly :

*Page 5: “**Anxiety- and stress-related conditions** such as generalized anxiety, obsessive-compulsive, or post traumatic stress disorders cause considerable individual suffering, but also impair daily functioning (e.g., work, home, relationships), and therefore generate substantial costs for the society (Bandelow et al., 2017).”*

2) Page 4 – “Reducing these individual and societal costs requires understanding why some individuals, but not others, develop anxiety-related conditions - for instance, following adverse life events”. The ending of this sentence feels incomplete – the example “following adverse life events” doesn’t well describe the issue introduced in the first half of the sentence.

Reply 2.2: We have revised this sentence and deleted the second part of the sentence.

Page 5: "Reducing these individual and societal costs requires understanding why some individuals, but not others, develop anxiety-related conditions."

3) Page 5 – HiTOP abbreviation used for the first time – should be spelled in full on first use

Reply 2.3: Thank you, we have now defined the acronym.

Page 6: "(e.g., Hierarchical Taxonomy Of Psychopathology (HiTOP); Kotov et al., 2017; Ringwald et al., 2023; Watson et al., 2022)"

4) pupil dilation (PD)) – double closed bracket should be corrected

Reply 2.4: We thank the reviewer and have followed the reviewers suggestions.

5) "Additional aspects such as reinforcement rate or the specific type of instructions (see e.g., Lonsdorf et al., 2017)". Some instances of the referencing are biased towards the Lonsdorf review. I don't have a problem with that review being cited in cases like this but I think it is fairer to cite empirical work (especially first demonstrations) that describe the exact topics being described, rather than a broad review of the work. For example, for generalization there is a meta-analysis on the topic recently published (Cooper et al 2022), and there are empirical studies/reviews that have examined reinforcement rate and specific instructions (e.g., Luck & Lipp, 2016)

Reply 2.5: Thank you. We have added additional references as suggested to provide a more comprehensive background.

Page 6: "Additional aspects such as reinforcement rate or the specific type of instructions ***(see e.g., Lonsdorf et al., 2017; Luck & Lipp, 2016)*** *shape the predictability and hence the strength of the experimental situation by introducing different levels of ambiguity."*

Page 7: " During fear generalization, participants are presented with the CS+, CS-, as well as with stimuli resembling both conditioned stimuli (generalized stimuli; GSs) to different degrees ***(see Cooper et al., 2022)****. The strength of CRs to the different GSs allows to infer to what degree defensive responding generalizes to stimuli similar but not identical to the conditioned stimuli."*

6) Page 6 "limited extent, (Lissek et al., 2006)while weak" – spacing should be corrected

Reply 2.6: We thank the reviewer and have followed the reviewers suggestions.

7) Page 12 "Studies employing avoidance or eye blink conditioning will be excluded as well as studies implementing only context conditioning while cue conditioning experiments implementing additional context manipulations will be included" needs some punctuation to break up the phrases in this sentence

Reply 2.7: We thank the reviewer and have rephrased this section in the manuscript as follows:

Page 13: "Studies employing avoidance or eye blink conditioning will be excluded from the literature search. Similarly, while cue conditioning experiments implementing additional

context manipulations will be included, studies implementing only context conditioning will be excluded."

8) Page 15 "package compute.es (Del Re, 2010), version will be added at stage 2)" – correct parentheses here

Reply 2.8: We thank the reviewer and have followed the reviewers suggestions.

9) Page 16 "Based on a priori power analysis, at least 10 effect sizes from the literature have to be extracted for associations between acquisition or extinction training and anxiety-related traits to justify fitting the described models" – are the authors suggesting that if fewer than 10 effect sizes are obtainable then no meta-analysis will be attempted?

Reply 2.9: Yes, if we should have less than 13 effect sizes for a primary meta analysis (i.e., effect sizes for acquisition or extinction for either CS discrimination, CS+ or CS- values), we will not conduct a meta-analysis for this effect of interest. Of note, the minimal number of needed effect sizes changed due to a recalculation based on another reviewer comment (see comment 5 of reviewer 1). We also rephrased this in the text to clarify it (see reply 2.10).

10) Please also confirm here that the difference measurements will be analysed separately

Reply 2.10: We confirm that we will conduct a separate analysis for each phase and for each key effect of interest including difference measurements (i.e., CS discrimination, CS+, CS-). To clarify this, we also rephrase the paragraph:

Page 18: Based on a priori power analysis, at least 13 effect sizes from the literature have to be extracted to conduct a primary analysis of the associations between acquisition or extinction training and anxiety-related traits for any key effect of interest (i.e., CS discrimination, CS+ or CS- values). All the analyses will be separately conducted and if the required number of effect sizes for one of these analyses cannot be extracted, we will refrain from conducting it.

Reviewer 3 (Luigi Degni):

The current registered report "Associations between anxiety-related traits and fear acquisition and extinction - an item-based content and meta-analysis" by authors Bruntsch et al. aims to investigate, through a meta-analysis of the existing literature, the association between fear acquisition and extinction, and anxiety related traits. Studies that used at least one among several different measures of fear conditioning (e.g., SCR, fear-potentiated startle) and one among different questionnaires to assess anxiety-related traits (e.g., STAI-T) are considered for the analysis. Several moderator and additional analyses are planned to be included in the manuscript.

The study is extremely well written, the hypotheses are clear and the planned methodology is explained very well and in detail. Importantly, the authors are careful in following the PRISMA guidelines for meta-analyses. Moreover, the research question is scientifically valid, and I believe the meta-analysis can interest a broad audience in several fields, including

neuroscience, psychology and psychiatry, possibly become an important tool for future studies on fear conditioning and anxiety.

However, I have some minor concern that I believe should be addressed prior to support its progress to stage 2.

1) Despite a previous meta-analysis on the relationship between fear extinction and anxiety-related traits was already published (Morriss et al., 2021), the authors want to focus part of this meta-analysis on such fear conditioning phase to give a more comprehensive understanding of the phenomenon, because the previous meta-analysis considered a limited number of variables. I appreciate this idea, but I believe authors should clarify which different results they expect from the new analysis, compared to Morriss and colleagues, and how such expected results may add novel insight into the relationship between extinction and anxiety related traits, in order to justify this new analysis.

Reply 3.1: We thank the reviewer for pointing out that it was not sufficiently clear from our writing what differentiates our planned analysis from the previously conducted meta-analysis by Morriss et al. (2021). Morriss et al. (2021) examined the association between self-reported Intolerance of Uncertainty scales (IU-27, IU-12, and their subscales P-IU and I-IU), trait anxiety (TA) and skin conductance response during threat extinction. The proposed study aims to extend the results of Morriss et al. (2021) in several aspects.

(1) First of all, we do not focus on specific questionnaires, like IU, STAI-T and STICSA, but try to take a multitude of different questionnaires into account that are related to various anxiety-related traits.

(2) Also, in regards to the selected outcome measure, we do not exclusively focus on SCR as a measure of interest but also take into account other types of outcome measures, such as subjective ratings (e.g., fear, anxiety or US expectancy), fear-potentiated startle, heart rate and pupil dilation.

(3) Further, we investigate the associations between anxiety-related traits and various outcome measures not only in extinction but also acquisition. As outlined in the introduction, we are convinced that it is important to learn more about fear acquisition as this phase is inherently present in studies using the fear conditioning paradigm and takes place before any other phase (hence results on any subsequent phase may merely reflect continued effects from this learning phase). Thus, getting a deeper understanding of the associations between anxiety-related traits and fear acquisition processes is expected to provide valuable insights also for the interpretation of results in following phases (e.g., extinction specificity of effects vs. broader effects).

Hence, we think that our analyses will give us a deeper insight into the association between anxiety-related traits and fear conditioning and extinction processes and add substantial information over and above the analyses of Morriss et al. if their findings were unique to the intolerance of uncertainty scale or might apply to a broader category such as anxiety-related traits as well. We have revised this section to flesh this point out more clearly:

Page 9: "Therefore, the aim is to extend the knowledge of the association between anxiety-related traits and fear acquisition and extinction processes by including various

outcome measures and different questionnaires. This will allow us to investigate if previously reported effects are generalizable to different outcome measures and experimental phases or if they are specific to a particular experimental phase, i.e., generalization (Sep et al., 2019), and specific outcome measures and constructs, i.e. SCR and IUS (Morriss et al., 2021), as reported in previous work."

Page 9: "Hence, specificity of these previously reported associations between anxiety-related traits and conditioned responding during extinction training (Morriss et al., 2021) and fear generalization (Sep et al., 2019) ***cannot be inferred in absence of knowledge of potential associations during fear acquisition training****."*

2) The authors are going to conduct the literature search on Web of Science and Pubmed. I suggest the authors to add Psycinfo to the other databases. I think such database can most fit with the query strings proposed by the authors. Moreover, this database overlaps only partially with the results found through Pubmed, whose articles are often selectively related to clinical aspects.

Reply 3.2: Thank you for these valuable suggestions. We will included Psycinfo as another database and also added this information in the manuscript:

Page 10: "The literature search will be conducted by using Web of Science, Pubmed and PsycInfo."

3) The test of outlier will be conducted using two different methods, respectively based on the extreme effect size and on the "leave-one-out method". I have some concern about this last method. Specifically, which are the criteria that the function InfluenceAnalysis compute to select the outliers? Please specify them, if possible.

Reply 3.3: The *InfluenceAnalysis* function of the *dmetar* (Harrer et al., 2019) package reports several influence characteristics and further a Baujat Plot and two forest plots based on the leave-one-out method and sorted by effect size or I^2 for the evaluation of influential studies. The calculation of the influence characteristics is based on the *influence.rma.uni* function of the *metafor* R package (Viechtbauer, 2010). The following criteria will be therefore calculated: externally standardized residual, DFFITS value, Cook's distance, covariance ratio, the leave-one-out amount of (residual) heterogeneity, the leave-one-out test statistic of the test for (residual) heterogeneity and DFBETAS value(s). Studies are seen as influential, based on suggestions by Viechtbauer et al. (2010), if one of the following criteria apply: (1) The absolute DFFITS value is larger than $3\times\sqrt{(p/(k-p))}$, where p is the number of model coefficients and k the number of cases, (2) The lower tail area of a chi-square distribution with p degrees of freedom cut off by the Cook's distance is larger than 50%, (3) The hat value is larger than $3\times(p/k)$ or (4) Any DFBETAS value is larger than 1 (Viechtbauer et al., 2010). We are aware that these criteria are arbitrary cut-off values suggested by Viechtbbauer et al. (2010).
Therefore and based on your comment, we decided to extend our approach following the methodology of Theriault et al. (2024). They introduce a composite score method for identifying statistical outliers, averaging binary (0 or 1) classifications from multiple methods. This score indicates the probability of an observation being classified as an outlier by any method, with scores of 0.5 or higher classified as outliers. We will adapt this by calculating a composite score based on all influence characteristics and apply the mentioned criterion for

identifying and excluding outliers. Baujat plots and forest plots will be used as visual aids to verify the composite score. In case there are outlier studies, we will perform iterations of our primary models, excluding studies identified as outliers or influential.If you still have concerns about this method, we would be eager to hear your arguments to further improve our analysis. We have reformulated the corresponding paragraph in the manuscript:

Page 19: "In addition, the aim is to also assess studies that will significantly influence our results based on other aspects than extreme effect sizes (e.g., larger sample sizes relative to median sample size). For this, the function InfluenceAnalysis of the dmetar package (Harrer et al., 2019) will be used. The function calculates several influence characteristics (i.e., externally standardized residual, DFFITS value, Cook's distance, covariance ratio, the leave-one-out amount of (residual) heterogeneity, the leave-one-out test statistic of the test for (residual) heterogeneity and DFBETAS value(s)) and further reports a Baujat plot and two forest plots based on the leave-one-out method, sorted by effect size or I2 for the evaluation of influential studies. If a potentially unduly influential study is to be excluded, it will be based on the approach of Theriault et al. (2024) for detecting statistical outliers in single- and multi-level models. A composite score will be calculated from all influence metrics. In this composite, each influence metric will be classified using a binary (0 or 1) scale where 1 indicates that a metric identifies the study as unduly influential, and these individual binary scores will then be averaged. This composite score reflects the probability of an observation being classified as influential, with scores of 0.5 or higher indicating unduly influential. Baujat and forest plots will be used to verify the composite score. For any influential studies identified, the meta-analytic results will be reported both including and excluding the respective study or studies."

4) Another doubt is about the quality assessment of included studies. Authors state that they will assess it by quantifying some quality indicators: for each criterion, studies can receive a score of two points (lower numbers would correspond to higher quality, and higher number would indicate lower quality). The total score will be included between 0 and 14. Moreover, authors want to assess other descriptive variables, but they affirm that "these additional descriptive information will not be included in the quality assessment score". So, it is not clear to me how they want to assess these information and eventually how they can be visualized with the Risk of bias tool if they are not quantitatively assessed.

Reply 3.4: Thank you for pointing out that this was not sufficiently clear. The descriptive information will be extracted in the coding stage (see extraction sheet columns: DG-DI). They will be later on reported if relevant but not visualized in the Risk of bias tool, as this tool will only be used for variables that are going to be scored. We refrained from quantitatively evaluating every detail of the quality assessment due to several reasons:

(1) It can be difficult to quantify some aspects between scores of 0 and 2 which might influence the quality of a study.

(2) Further, to include some of these descriptive variables as quantitative criteria could systematically penalize older studies, in which for example open science practices were not commonly used. Still, we would like to assess such open-science related aspects. Therefore, such additional pieces of information will be extracted from the individual studies and only descriptively reported and not visualized.

(3) Another reason to include these descriptive variables is to check for procedures that may distort the results and would therefore exclude its effect size from the analysis. An example would be response quantification approaches which differ for different stimulus types (as reported in some studies of a recent review in the field of fear conditioning research, Ruge et al. 2023) most likely leading to systematic differences between the averaged responses to the CS+ and the CS- or questionable data participant exclusion strategies (see Lonsdorf et. al., 2019 elife) or inappropriate operationalization of constructs (see eg. Lonsdorf et al., 2019 Biological psychiatry). If a study is excluded due to flawed methods, we will run the primary models with and without the corresponding study and report the results.

We also added more information about this in the manuscript:

Page 20: "However, this additional descriptive information will not be included in the quality assessment score, as it is either highly individual and therefore difficult to objectively quantify some variables between scores of 0 and 2, or would systematically penalize older studies, in which for example open science practices were not commonly used."

Page 21: "Still, we may exclude effect sizes of studies based on the use of flawed methods or participant exclusions that are highly likely to distort the final results (e.g. SCR quantification approaches that confound CS and US responses due to inappropriate scoring windows. If this should be the case, the reasoning for the exclusion will be outlined in detail in the Stage 2 RR."

5) Authors mention the possibility to perform an additional moderator analysis on "the implementation of the experiment in an MRI scanner". What does "additional" mean? In other words, how does this moderator differ from the others proposed?

Reply 3.5: It does not differ from the other moderator analyses. To avoid confusion we rephrase this sentence as follows:

Page 23: "Also, the implementation of the experiment in the MRI scanner (as opposed to the behavioral lab) will be assessed as a possible moderator, as it has been shown to introduce sampling bias regarding individual differences such as trait anxiety (see Charpentier et al., 2021; and Sjouwerman et al., 2020)."

Reviewer 4 (Marco Badioli):

The authors of the manuscript "Associations between anxiety-related traits and fear acquisition and extinction – an item-based content and meta-analysis" aim to investigate the relationship between implicit and explicit anxiety-related trait measures and classical fear conditioning acquisition and extinction. Moreover, the authors aim to assess the modulating effect of explicit anxiety-related traits at both questionnaires and questionnaire item levels.

This registered report demonstrates the potential to fill a gap in the literature concerning the relationship between anxiety-related traits and classical fear conditioning features. The manuscript appears to be very well-written and comprehensive in both theoretical and methodological aspects. However, some minor concerns should be addressed to help improve the quality of the manuscript:

1) The search strategy initially focused on standardized procedure to reduce possible biases; however, at the end of the search process, the authors decided to add “fearfulness” and “harm avoidance” as search terms “to cover more aspects of anxiety-related traits”. This decision could risk weakening the carefully executed standardized search strategy. Could the authors provide a more detailed explanation of why new search terms were added? Additionally, “fearfulness” and “harm avoidance” didn’t appear in the query strings on OSF. Please add these two search terms.

Reply 4.1: Thank you for this important comment. Indeed, the mentioned query strings were added after the naive search has already been entered into litsearchr. Two co-authors joined the project at a later stage, when the naive search had already been done and pointed out that we should include these terms to cover even more potential anxiety-related traits, which we agreed on.
As the reviewer stated these two terms should also be included in the naive search, we conducted the litsearchr search again including the additional terms. In this second “naive” search no new words came up that would be relevant for the refinement of our search terms. This is documented in the updated file “add_searchterms_MJK_MB.csv” (https://osf.io/qc7v6). We also added the naive search terms and the nbib files of our second search used for litsearchr on OSF (https://osf.io/45qea/) in the “litsearchr” folder but did not update the list of included final search terms based on the litsearchr results, as no new and relevant terms came up in this new search. We also updated the list of search terms (page 10) and the corresponding paragraph in the manuscript:

Page 11: “More precisely, the first developed search terms (see Table 1, ‘Naive search terms’) were used as the basis for a naive literature search in Web of Science on May 6, 2024. From the resulting literature, new terms were extracted based on the keywords and titles that appeared in at least 3 articles and contained at least 2 words. From the resulting terms, excluded predefined terms (i.e., ‘stopwords’) which were implemented by litsearchr (commonly used words, such as ‘is’, ‘not’, or ‘but’) and expanded by MKJ (words not in the scope of our study, such as ‘autism’, ‘covid-19’, or ‘opioid’) were excluded. Then, the most frequent 80% of the terms were selected to obtain a feasible number of terms to check for possible inclusion (i.e., 292 terms). MBr and MKJ independently assessed whether or not each term should be included in the final literature search. Subsequently, choices were reevaluated and disagreements resolved between MBr and MKJ. Reasons for exclusion of terms discussed were recorded and can be found on OSF. The final search terms are provided in Table 1 in a simplified format and the exact query strings for the specific databases can be found on OSF.”

2) Although eligibility criteria were well explained; reporting them in list form could enhance readability.

Reply 4.2: We thank the reviewer for this suggestion. A list is rather uncommon for a scientific paper and takes up a lot of space, and therefore we report the criteria in the current form. However, we uploaded a file on OSF (https://osf.io/dk4xp) containing all eligibility criteria and linked it in the text:

Page 14: “All eligibility criteria can also be found on the OSF.”

3) The authors reported that they will use τ2 parameter as one of the measures of between-study heterogeneity, which represents “the standard deviation of the true effect size”. Actually, τ2 represents the variance and not the standard deviation of the true effect size (Harrer et al., 2022).

Reply 4.3: We thank the reviewer and have corrected this error:

Page 18: “Between-study heterogeneity of the meta-analysis will be parameterized via ***τ2, the variance of the true effect sizes****, I2, the percentage of variability (not caused by the sampling error), H2, the ratio of observed variation, and the Q-test that is used to check if the variation of the meta-analysis significantly surpasses what would be expected under the null hypothesis (Harrer et al., 2022).”*

4) τ2, I2, H2, and Q-test parameters will be used to assess the between-study heterogeneity. While τ2, I2, and H2 could provide different types of information regarding between-study heterogeneity, Q-test should be used carefully as it could produce biased results based on the statistical power of the meta-analysis (Harrer et al., 2022).

Reply 4.4: The reviewer states an important point which we are aware of. As the calculation for the I2 and H2 are based on the Q-test parameter we will still include it but any claims regarding the between-study heterogeneity will be mostly based on the other parameters. Moreover, we wanted to report the Q-parameter to also enable the comparison to other meta-analyses that reported it as well. We rephrase this section in the manuscript in the following way:

Page 18: “Any claim about between-study heterogeneity will be based on τ2, I2 and H2 but not on the Q-parameter as this could bias, based on the statistical power, the results of the meta-analysis (Harrer et al., 2022). The Q-test will still be included in the analysis to enable the comparison with other meta-analyses that used it.”

5) Quality assessment of the included studies will be conducted by using a well-established protocol. However, it is not clear what will be the inclusion/exclusion criteria of the studies based on the protocol. Authors should clarify if there will be a cut-off in the rating or if other parameters will be used.

Reply 4.5: We thank the reviewer for bringing this point to our attention. So far, we did not plan to exclude any study based on their study quality rating and only include this information in one of the subgroup analyses. As our protocol has been designed for the evaluation of the study quality and not for exclusion and it is based on protocols of other researchers but has also been modified for our purposes, we are hesitant to apply an arbitrary cut-off value for excluding a study based on subjective quality criteria, as there is no consensus or guideline for that.
However, we leave the option open to exclude effect sizes of studies based on flawed methods or participant exclusions which may distort the results. Reasons could be inappropriate response quantification approaches that are not tailored to the used outcome measure (see Ruge et al., 2023), questionable data participant exclusion strategies (see Lonsdorf, Klingelhöfer-Jens, et al., 2019) or inappropriate operationalization of constructs (see Lonsdorf, Merz, et al., 2019).
Given the potential for unforeseen issues to arise in the selected studies, we refrain from

providing specific criteria for this. However, we will explain our reasoning for each effect size/study in the manuscript should this be necessary.

Page 21: "The quality of studies included in the meta-analysis will be visualized with the 'Risk of bias' (robvis) tool (McGuinness & Higgins, 2021) and the scores will be used to evaluate if poor study quality is inflating effect sizes. Therefore, they will be analyzed in a moderator analysis (see section Moderator analyses).

No study will be excluded based on a poor study quality score as there are no guidelines for an appropriate cut-off value. Still, we may exclude effect sizes of studies based on the use of flawed methods or participant exclusions that are highly likely to distort the final results (e.g. SCR quantification approaches that confound CS and US responses due to inappropriate scoring windows. If this should be the case, the reasoning for the exclusion will be outlined in detail in the Stage 2 RR"

References:

Bandelow, B., Michaelis, S., & Wedekind, D. (2017). Treatment of anxiety disorders. *Dialogues in Clinical Neuroscience*, *19*(2), 93–107. https://doi.org/10.31887/DCNS.2017.19.2/bbandelow

Brunner, J., & Schimmack, U. (2020). Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2018.874

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *dmetar: Companion R Package For The Guide 'Doing Meta-Analysis in R* (0.1.0) [Software]. http://dmetar.protectlab.org/

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., … Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology

(HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, *126*(4), 454–477. https://doi.org/10.1037/abn0000258

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

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

Lonsdorf, T. B., Merz, C. J., & Fullana, M. A. (2019). Fear Extinction Retention: Is It What We Think It Is? *Biological Psychiatry*, *85*(12), 1074–1082. https://doi.org/10.1016/j.biopsych.2019.02.011

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

Morriss, J., Wake, S., Elizabeth, C., & van Reekum, C. M. (2021). I Doubt It Is Safe: A Meta-analysis of Self-reported Intolerance of Uncertainty and Threat Extinction Training. *Biological Psychiatry Global Open Science*, *1*(3), 171–179. https://doi.org/10.1016/j.bpsgos.2021.05.011

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522. https://doi.org/10.1037/0033-295X.108.3.483

Ringwald, W. R., Forbes, M. K., & Wright, A. G. C. (2023). Meta-analysis of structural evidence for the Hierarchical Taxonomy of Psychopathology (HiTOP) model.

Psychological Medicine, *53*(2), 533–546. https://doi.org/10.1017/S0033291721001902

Ruge, J., Ehlers, M., Kastrinogiannis, A., Klingelhöfer-Jens, M., Koppold, A., & Lonsdorf, T. (2023). *How adverse childhood experiences get under the skin: A systematic review, integration and methodological discussion on threat and reward learning mechanisms*. PsyArXiv. https://doi.org/10.31234/osf.io/nfpbj

Sep, M. S. C., Steenmeijer, A., & Kennis, M. (2019). The relation between anxious personality traits and fear generalization in healthy subjects: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, *107*, 320–328. https://doi.org/10.1016/j.neubiorev.2019.09.029

Stussi, Y., Pourtois, G., Olsson, A., & Sander, D. (2021). Learning biases to angry and happy faces during Pavlovian aversive conditioning. *Emotion*, *21*(4), 742–756. https://doi.org/10.1037/emo0000733

Stussi, Y., Pourtois, G., & Sander, D. (2018). Enhanced Pavlovian aversive conditioning to positive emotional stimuli. *Journal of Experimental Psychology: General*, *147*(6), 905–923. https://doi.org/10.1037/xge0000424

Thériault, R., Ben-Shachar, M. S., Patil, I., Lüdecke, D., Wiernik, B. M., & Makowski, D. (2024). Check your outliers! An introduction to identifying statistical outliers in R with easystats. *Behavior Research Methods*, *56*(4), 4162–4172. https://doi.org/10.3758/s13428-024-02356-w

Viechtbauer, W. (2010). Conducting Meta-Analyses in *R* with the **metafor** Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Watson, D., Levin-Aspenson, H. F., Waszczuk, M. A., Conway, C. C., Dalgleish, T., Dretsch, M. N., Eaton, N. R., Forbes, M. K., Forbush, K. T., Hobbs, K. A., Michelini, G., Nelson, B. D., Sellbom, M., Slade, T., South, S. C., Sunderland, M., Waldman, I., Witthöft, M., Wright, A. G. C., … HiTOP Utility Workgroup. (2022). Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): III. Emotional dysfunction superspectrum. *World Psychiatry: Official Journal of the World Psychiatric*

Association (WPA), *21*(1), 26–54. https://doi.org/10.1002/wps.20943