Dear Managing Board of PCI Registered Reports,

Thank you for your editorial letter dated 1/12/2021 concerning our manuscript *Stage 1 Registered Report: Stress regulation via being in nature and social support in adults - a meta-analysis.* In the current letter we addressed the concerns that reviewers raised by providing point-by-point answers. We have ordered them by reviewer. Additionally, please find an updated version of the current manuscript with track changes, so that you and reviewers can easily spot the changes that have been made to the document.

We thank you and all the reviewers for the feedback that we feel have directly contributed to improving the quality of our Registered Report proposal. We really feel that the Registered Report format offers a superior publication format and has improved our work prior to data collection. We also really appreciate the constructive contributions by both reviewers. We hope that the second version of our manuscript will now match the high standards for In Principle Acceptance with PCI Registered Reports.

Sincerely and on behalf of all the co-authors,

Alessandro Sparacio

Reviewer Feedback

Reviewer 1: Dr. Felix Schönbrodt

I have some comments on the proposal, put forward in a constructive mindset with the hope that they make the proposal even stronger. All of them should be easily doable. I roughly ordered them by their importance (in decreasing order). In particular points 1 and 2 do not have the necessary stringency for a pre registration / RR Stage 1 protocol yet.

1. Estimation techniques:

Comment 1: Multilevel modeling: What if k is very small, e.g. in a subgroup analysis - is a multilevel RE model still appropriate? What if only one study with multiple ES is present, and all others have only 1 - that could break the MLM estimation. I wonder whether you need a fallback strategy for that analysis.

Authors' Response: Thank you for this comment and we completely agree; we had initially not planned for such a possibility. Generally, we think it is always better, by default, to apply a model having a higher verisimilitude to the underlying data generation process. That is why we planned to apply a multilevel RE model regardless of whether it shows a significantly better fit compared to a simpler (but in our expected case a conceptually misspecified) two-level model in a likelihood ratio test. That said, we agree with your point and now will switch to an ordinary two-level RE model under two contingencies: (1) The multilevel model fails to converge in the overall model or any of the subgroups in a subgroup analysis; (2) if the variance components of the model are not well identifiable (specifically, if the log-likelihood does not peak at the variance estimates for both variance components – checked using a profile plot).

Comment 2: Sensitivity analyses: I appreciate the aim to do many sensitivity analyses. But I wonder: are they too many? Do you plan to cross all dimensions of possible variation, or do you always fix the others to one (sensible) value while varying one focal dimension? I envision that reporting, summarizing, and interpreting this plethora of checks will be strenuous and maybe messy.

Authors' Response: Based on your suggestion, we have reconsidered that aspect of the analysis and decided to simplify the system of sensitivity analyses reported in the main manuscript considerably. We removed sensitivity checks for arbitrary decisions that we think would have minimal effect on the results anyway. Specifically, (1) we removed the sensitivity analysis varying the assumed correlation for within-participants designs (as these will likely represent a small proportion of the utilized designs at best); (2) we removed the sensitivity analysis varying the assumed sampling correlations among the effects originating from identical samples (and assume .5 as the constant sampling correlation in the CHE model) as we suspect that this tends to have only a relatively minor effect on the meta-analytic variance estimates; (3) we also removed the sensitivity check for PET-PEESE, using the selection model as a conditional estimator instead of PET (for rationale, please see our response to your point #5).

All these sensitivity analyses will be more closely described and reported only in the supplement. The main manuscript will include only a reference to these analyses. These sensitivity checks will be carried out by varying one dimension and keeping the others fixed at the default values used in the analysis and reported in the paper. The code is/will also be designed to let the interested reader easily modify the inputs to these sensitivity checks.

In the main paper, we will only report one sensitivity analysis – the effect of excluding influential outliers. Effects having a large standardized residual (> 2.58) will be regarded as excessively influential.

Comment 3:. An MLM selection model should be possible in principle, but I haven't seen that yet. So the permutation approach is a viable workaround. The open data will allow re-estimations when new methods are available.

Authors' Response: We agree and we thank you for affirming. In this respect, we tried to design the code in a rather modular way, allowing one to easily edit/change these methods. As the entire workflow from data wrangling to HTML report generation is/will be implemented in R, one will be able to easily evaluate the effect of any change to the analytic apparatus.

Comment 4: "To use a measure of precision that is uncorrelated with the effect size, we used $\sqrt{(2/N)}$ and a 2/N terms instead of standard error and variance for PET and PEESE" -> Please provide a reference or a rationale why you chose this approach (and deviate from the standard procedure).

Authors' Response: We apologize for the accidental omission of the reference in the text (it was only referenced in the code). This approach was based on a simulation study by James Pustejovsky published on his blog (2017). The reason to use the $\sqrt{(2/N)}$ and a 2/N terms instead of standard error and variance for PET and PEESE is that N-based predictors do not induce a correlation with the effect size like that variance or standard error do (as these are calculated using the ES), where exactly this correlation is the estimand of PET-PEESE. Models fitted using N-based predictors exhibit practically no bias in the absence of publication bias and a markedly lower false-positive rate (Pustejovsky, 2017).

Please see https://www.jepusto.com/pet-peese-performance/.

We added the left-out reference and a short rationale for choosing n-based predictors instead of variance-based ones. The current version of the code also allows the user to change the approach from N-based predictors to the original procedure using a simple TRUE/FALSE argument.

Comment 5: "Additionally, we also used the 4PSM as a conditional estimator for PET-PEESE" -> How does that work? Conditional on what? I did not understand this.

Authors' Response: Our idea was to also use the selection model as a conditional estimator for PET-PEESE, instead of PET. As is well known, the original PET-PEESE uses PET result to decide whether to use standard error (PET) or variance (PEESE) as a model term. The 3PSM, however, tends to have more favorable error rates under many conditions than PET, and that is why we also wanted to try to apply a better estimator than PET to decide whether to reject the null and thus apply PET or PEESE.

After considering your other points, we finally chose not to overcomplicate any further. Especially given the fact that PET-PEESE was used rather as a secondary bias-adjustment method, we felt that this additional experimental layer was an overkill and removed it from the manuscript. The R code still allows anyone to set 3PSM as a conditional estimator, the function argument is, as before, set to FALSE by default though. The code will, however, allow the user/reader to change the default setting of conditional estimator for PET-PEESE, along with other tuning arguments (this is reflected in the manuscript, as we have removed the following sentence:"Additionally, we also used the 4PSM as a conditional estimator for PET-PEESE and explored the effect of such decision on the resulting inference.").

Comment 6: You have a fallback strategy from 4PSM to 3PSM, depending on the number of p-values in each bin. What happens if there are <4 p-values in the 3PSM bins? (Which can easily happen at high publication bias).

Authors' Response: Our initial plan was, following a fallback from 4 parameters, to estimate the 3PSM, disregarding the distribution of p-values across the two bins, and report the results if it converges. Along with the results, we will also report the total *k* and whether any of the bins contained <4 p-values (cautioning the reader that the model results rest on shaky ground). That said, we do not have a great deal of conviction about this approach. We are thus open to changing our approach and following your advice if you strongly believe our approach is suboptimal and thus refrain from estimating the model altogether when any k_bin < 4.

Comment 7: How do you do inference in the permuted 4PSM models? I understand that the median estimate is used for interpretation, but how is inference done?

Authors' Response: Not just in the 4PSM but also in other methods implemented using a permutation-based procedure, we will average over the iterations by *picking the model* with the median ES estimate (and the median z-score for the right-skew estimate of the full p-distribution for p-curve). In other words, we will pick the median estimate from the parameter distribution and, with it, the corresponding model that the estimate was originating from. For the selection model, for instance, this preserves the mutual consistency between the estimate, z-value, CIs, and p-value. The inference will then be approached in the exactly same way as if one interprets the results of a single ordinary selection model. In the revision, we have tried to describe this procedure in a clearer way (please see below) and explicitly mentioned that both the interpretation and inference will be based on that median model:

"To deal with dependencies in the data and avoid arbitrariness in the selection of effects within studies, we applied a permutation-based procedure, iteratively selecting only a single focal effect size from each independent study, estimating the model in 5000 iterations, and averaging over the iterations by picking a model having the median ES estimate (where both, the interpretation and inference will be based on that median model).[4] ”

[4] That is, we picked the median estimate from the parameter distribution and, with it, the corresponding model that the estimate was originating from. The goal of this procedure was to preserve the mutual consistency between the estimate, *z*-value, *CI*s, and *p*-value.

Comments 8 & 9: ”If the results of the 4-parameter selection model disagreed with the more general Bayesian model-averaging approach”: What is the inferential criterion for the RoBMA results? The HDI? A BF? How is “disagree” defined? What if both show “significant” positive results, but disagree in magnitude? I think the final ”inference algorithm” should be defined more clearly in the preregistration. Currently, it seems to leave a lot of researcher degrees of freedom.

Inference: I think it should be clearer and more stringent, which models are used for interpretation and inference. E.g., you write „To estimate the range of effect sizes that can be expected in similar future studies, we calculated the 95% prediction intervals. For each analysis we conducted, when the included effects (k) were less than 10, we did not interpret the estimates.“ -> this relates to the non-bias-corrected model. PET-PEESE is not used at all for inference (except as a part of the RoBMA approach). Could you give some justification on why you capitalize on 4/3PSM, ignore PET-PEESE, and use RoBMA as a „validator“? (To be clear: I think this is a reasonable approach, but some justification for the reader would be nice. Maybe also mention that RoBMA is a quite new approach that probably has not been fully vetted by independent experts and has not been stress-tested in practice). To summarize: I think inference and interpretation should be based on the same model. Make clear what the status of the non-bias-corrected results are: Are they reported just for completeness? Or will they be interpreted? Why not interpret the bias-corrected estimates which are the basis for inference?

Authors' Response: After thinking about the different options, you convinced us that the approach based on the principle that “inference and interpretation should be based on the same model” is the most rational. Our initial thought was, indeed, to examine if the result of a single ordinary multiple-parameter selection model is “validated” by a more general model-averaging approach. We agree though, that the method has not been sufficiently vetted in independent simulation studies so instead decided to do the following:

1. Based on the premise that (albeit still largely imperfect) the selection models represent a less misspecified representation of the selection process, we will base the inference on the result of a selection model, rather than on an unadjusted ordinary RE model.
2. To stay on the conservative side, we will base the inference on the result of the 4/3-parameter selection model solely, as its performance under various conditions has been studied more extensively (compared to the RoBMA approach) and the informed readers can thus have a better calibrated notion about the expected degree of bias and precision of this method in suchlike analytic conditions.
3. The estimation will be based on the same single model. The unadjusted RE estimate will be reported just for completeness, the substantive interpretations will be based on the bias-adjusted model.
4. PET-PEESE and RoBMA models (as well as the Vevea and Woods step function models with a priori defined selection weights denoting moderate/severe/extreme selection) will be estimated and reported in a descriptive manner to offer the reader a different perspectives using a model based on a different approach and posing

> different assumptions (PET-PEESE) and on a arguably more general model-averaging approach (RoBMA). The results of these models will not weight in with respect to the substantive inferences. We will descriptively note, though, whether the primary 4/3-PSM estimate fell within the 95% credible interval of the RoBMA estimate.

In the revised protocol, we have removed the mention of the RoBMA-based inferential contingency and added the following:

"Substantive interpretations were guided by the estimates and inferential results of the 4PSM solely. The other exploratory bias-adjustment methods served a descriptive purpose, to provide the reader with a more comprehensive view on bias adjustment under quantitatively and qualitatively different assumptions (Vevea & Woods models and PET-PEESE, respectively) and using a more general model-averaging approach (RoBMA)[6]."

[6] Apart from reporting the results of these bias adjustments, we examined whether the primary 4/3-PSM estimate fell within the 95% credible interval of the RoBMA estimate (being based on a more general model).

Comment 10: Text order. It was confusing to read about the bias assessment *after* reading about the fact that studies will be excluded based on that assessment - maybe shift that section before the analysis section?

Authors' Response: We thank the reviewer for the suggestion. We think that the analysis part describing the traditional unadjusted models (along with more general analytic decisions) should precede the section describing the adjustment of these models. Instead, we chose to move the part in which we excluded studies based on the bias assessment in the "Quality of evidence assessment". We hope in this way to have made the structure of the analysis part clearer.

Comment 11: Exclusion criteria: Do you also exclude studies with inconsistent n? If studies are excluded based on risk of bias etc.: Are they a priori excluded for all analyses, or do you look at this assessment as a moderator (e.g. to show inflated ES in biased studies)?

Authors' Response: In general, we tried to be very conservative about excluding otherwise eligible studies/effects. Therefore, we chose to exclude studies based on reporting inconsistencies and high risk of bias only in a moderation analysis, where we will examine the effect of doing so. Apart from several edits to the Inclusion criteria part (please see p. 11-13), we made that particular aspect of the analysis clearer in the present version of the RR, please see below.

"Finally, we ran two moderation analyses to assess whether studies with high risk of bias and mathematically inconsistent means or standard deviations showed inflated effect sizes".

Comment 12: Why these two interventions? I understand that you have to start somewhere, but it would be interesting what guided the choice. Are these interventions the most often applied? Do they provide the strongest evidence so far?

Authors' Response: We thank the reviewer for making this point. We have now specified in the text why our choice fell on these two strategies:

"The reason why we chose these two strategies is similar to what guided the choice in our previous work: The decision was partly based on the fact that we were interested in analyzing scalable, non-invasive and cheap strategies that could be used by an extended number of individuals and partly arbitrary as to where we start with our approach."

Comment 13: Personality traits as moderators. This is mentioned on p. 3, but never picked up again. Why would you expect such a moderation? Is that incorporated in your analysis in any way?

Authors' Response: We thank the reviewer for pointing this out. We wanted to include this analysis because we believe that there is a strong link between personality traits and the way individuals react to stress. For instance, some studies found evidence that people that have low scores in Neuroticism and high scores in Conscientiousness are the ones with the most favorable combination of traits when it comes to cope with stress (Vollrath & Torgersen, 2000). However, in our previous meta-analysis on the efficacy of self-administered mindfulness and biofeedback on stress regulation (Sparacio et al., 2022) we saw that personality traits and individual differences were oftentimes measured, but not taken into account in any of the focal analyses (they were solely used as exclusion criteria). Thus, we wanted to assess whether in the literature of being in nature and emotional social support these personality traits are not only measured, but also used in moderation analyses to assess whether they play a moderating role in the processes of stress regulation. We did not mention this moderation analysis in the analysis paragraph originally; we have now added it in the revised version of the manuscript.

“To check whether the named strategies have an effective role in reducing stress levels we conducted a meta-analysis with the following objectives: (...) to determine whether personality traits were used as moderators in stress regulation studies.”

Comment 14: Consequences of stress: If I understood correctly, the authors would also include studies that do not include one of the three components of stress (the mediator), but only consequences of stress. Then, I guess, they have to include the entire literature on depression, well-being, and much more, as ”affective consequences of stress” can be really a lot. I am not sure whether under these conditions the scope of the meta-analysis is clearly enough defined. What if studies measure well-being (as a consequence), but otherwise have no relation to ”stress” as the mediating factor, neither by measuring it, nor theoretically? Would that fulfill the inclusion criteria? (I am aware that ”stress” always is included in the search term, but the primary study still could be quite distant).

Authors' Response: We see dr. Schönbrodt’s point. However we decided to restrict the focus of our meta-analysis to trait anxiety and depression. This is indeed another arbitrary decision dictated by constraints of time and resources; we could not extend the field of research to all the affective consequences of stress. Thus, we decided to make a selection picking up the two consequences that are more often targeted for interventions, namely trait anxiety and depression. We specified this in the revised version of the manuscript:

“We decided to pick depression and chronic anxiety as relatively arbitrary starting points for constraints of time and resource and because those are traditionally the most investigated outcomes for these interventions”

Comment 15: Hypotheses: I am not sure if it is necessary to formulate hypotheses, given that the focus is on estimation. Sure, at the end p-values will be computed and reported; but I think the hypotheses could be dropped without much/any loss.

Authors' Response: We see the reviewer's point. We recognize that formulating hypotheses is uncommon for a meta-analytic approach and that removing the section would not invalidate the robustness of our approach. Thus, we followed the suggestion of Reviewer 1 and we dropped it from the revised manuscript.

Comment 16: Inclusion criteria: To be clear: Do you only include experimental studies?

Authors' Response: We apologize for not being clear about that. We included not only experimental, but also observational studies estimating the exposure to being in nature or emotional social support.

Naturally, the observational designs carry high risk of bias due to possible causal confounding. Lack of randomization, however, automatically puts these studies into the high risk category (by the Risk of Bias 2 tool) and so the effects originating from these designs are excluded in a sensitivity analysis.

Comment 17: Subgroup analyses: Why k=10 as cutoff? How did the authors arrive at that number?

Authors' Response: We detailed in the revised version of the manuscript why we chose the threshold of 10. It is again a somewhat arbitrary choice to avoid large expected sampling variability when there are too few effects. We hope to have more precise results with the addition of this rule. We now provide a clearer justification, yet are willing to discuss alternatives.

"For each analysis we conducted, when the included effects (*k*) were less than 10, we did not interpret the estimates. Similarly as we did for our previous meta-analyses (See Sparacio et al., 2022; IJzerman et al., 2022), we have chosen this threshold arbitrarily, because of the large expected sampling variability of such estimates, leading to imprecise results in smaller sets of effects."

Comment 18: Existing meta-analyses: Maybe it would be interesting to report a reproduction attempt of the existing meta-analysis (in particular when they have been done by other authors). Did you extract the same effect sizes? Do you arrive at a comparable estimate/ conclusion? Although the new, more encompassing MA supersedes the old MAs, it could be interesting to what extent the old stuff is reproducible.

Authors' Response: Some existing meta-analyses exist on the two strategies we included in our project, but the inclusion criteria do not really overlap. For instance the meta-analysis of Schwarzer et al., (1989) is centered on social support and health, meaning that both the strategy (i.e., social support) and the dependent variable (i.e., health) are broader than our strategy (emotional social support) and of our dependent variable (i.e., stress). The meta-analysis of Antonelli et al., (2019) is comparable to our same strategy (i.e., being in nature), however their dependent variable (i.e., cortisol) is narrower than ours (i.e., stress). It will thus be very difficult to compare the conclusions and effect sizes from these meta-analyses to ours.

Minor points:

Comment 19: p. 3: „We intend to shed light on the mechanisms underpinning stress regulation by employing a workflow incorporating various publication bias-correction techniques" —> how can the latter shed light in the former?

Authors' Response: We see how this was unclear. We changed this to:

"We intend to shed light on whether being in nature and emotional social support has stress reducing effects or not through our meta-analysis and how big the effect - if any - is. Our combination of publication bias-correction techniques can provide a less biased estimate of the effects of interest (Cf., IJzerman et al., 2022; Sparacio et al., 2022)."

Comment 20: It might be helpful to explicitly state that the authors (of course) include all studies from the existing meta-analyses.

Authors' Response: We thank Reviewer 1 for making this point. We indeed included studies from existing meta-analyses. Of course, given what we have said above, we could not include all, but only those that satisfy our inclusion criteria. We have excluded those who do not satisfy our criteria. We now clarified this choice in the manuscript :

“We included studies of existing meta-analyses that satisfied our inclusion criteria.”

Comment 21: p. 7 „For emotional social support we conducted two additional subgroup analyses: The type of social support (e.g., physical) and the source of social support (e.g., known person or stranger).“ —> is the e.g. exhaustive? Can you already define what the subgroups will be? From a preregistration point of view, this would be desirable. Or write explicitly that the categories are not fixed yet and will be created during the coding phase.

Authors' Response: We apologize for not having clarified this. We have now specified this in the manuscript:

“For emotional social support, we conducted two additional subgroup analyses: The type of social support (0=not specified, 1=physical, 2=verbal, 3=mixed, 4=other) and the source of social support (0=not specified, 1= stranger, 2=known person; see for more details our coding sheet; https://osf.io/4cjux/). Although we believe that this coding is exhaustive, if we realize when we start the data collection that our coding sheet is inadequate, we may change our coding scheme, which we will document it in the Appendix A: Protocols and deviations sheet.”

Comment 22: p. 10 „For the affective consequences of stress, we used the same procedure we used for the affective components of stress.“ —> I am not sure to what procedure this sentence relates to.

Authors' Response: Thanks for pointing this out.We have clarified this aspect better in the revised version of the manuscript:

“For the affective and cognitive components as well as the affective consequences, we relied on self-report measures.”

Comment 23: p. 11: Exclude studies where participants were below 18 years of age: Any participant? (If it’s only one?)

Authors' Response: As per exclusion criterion we will exclude studies where at least one participant is below 18 years of age. We have now better defined this exclusion criteria in the text:

“A study was excluded if (...) The sampling frame of the study explicitly involved participants below 18 years of age.”

Comment 24: p. 11: Namely, for being in nature, we excluded studies in which participants engaged in physical activities besides walking (e.g., running or exercising).“ —> What if the control group is „running indoors“ (vs. running outdoors). Shouldn’t that be eligible?

Authors' Response: We will exclude each study in which participants engage in any type of physical activity other than walking. This is because in such a scenario it would be difficult to disentangle the stress-reducing effect of physical activity and that of being in nature. We will thus exclude studies such as the one described by Dr Schönbrodt.

Comment 25: p. 12: „the number of citations of the paper" —> according to which database?

Authors' Response: We meant number of citations from Google Scholar. We have clarified this part now in the new version:

"We extracted data for the following variables: publication year, the number of citations of the paper by Google Scholar at date of extraction".

Comment 26: p. 15: „by varying the assumed severity of bias, modeling moderate, severe, and extreme selection." —> how did you model this? I don't want to look into the code for that information.

Authors' Response: We had not included this detail, as we felt it would be overkill. We did however specify what we did a little more clearly in the revised version of the manuscript:

"First, we tried to assess the variability in adjusted estimates under different assumptions of the publication selection process using Vevea and Woods' (2005) step function models with a priori defined selection weights (instead assessing them via estimates of maximum likelihood). These step function models allowed us to explore the results by varying the assumed severity of bias, modeling moderate, severe, and extreme selection."

We agree that it is important to make the reader understand the interpretation of the Vevea & Woods model. However, we chose not to incorporate the V&W model itself as we felt it to be too much for a model that is not our primary model. For now, we are including the specification of the step function into the functions.R script file on our GitHub RePo (https://github.com/alessandro992/Registered-report-meta-analysis/blob/main/functions.R; Line 19 of code). However, if the reviewer really strongly feels otherwise, we can include the table below:

Steps	.0025	.005	.0125	.025	.05	.10	.25	.50	1
Moderate selection	1	.99	.97	.95	.80	.60	.50	.50	.50
Severe selection	1	.99	.97	.95	.65	.40	.25	.25	.25
Extreme selection	1	.98	.95	.90	.50	.20	.10	.10	.10

Reviewer 2: Siu Kit Yeung

Comment 1: Generally clear and transparent. Meaningful topic. I learned something new about publication biases tests and tools for meta-analyses. Thank you! I am grateful for this opportunity and learning experience. I suggest revision and resubmission, with relatively minor changes.The below are suggestions. It is up to you to adopt or not.

Authors' Response: We thank Reviewer 2 for the kind comments.

Comment 2: Issue: It appears that CRediT – Contributor Roles Taxonomy is used, but no citation is provided. Suggestion: Please add citation (Allen & O'Connell, 2014) and provide the full name of CrediT there.

Authors' Response: We added the requested citation:

"Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Credit where credit is due. *Nature*, *508*, 312–313. https://doi.org/10.1038/508312a."

Comment 3: Issue: I can't find a statement regarding moderator analyses or subgroup analyses Suggestion: I believe it is better to add statements regarding these. It would be more interesting to read an Abstract with those elements.

Authors' Response: We see the reviewer's point. On the basis of the suggestion provided by Reviewer 2 we have changed our abstract adding statements as related to subgroup analyses:

"We carried out several subgroup analyses to investigate the heterogeneity caused by variations in population characteristics or conceptual aspects of utilized study designs and we found [no evidence for x subgroup analyses and/or evidence for x subgroup analyses]"

Comment 4: For the discussions regarding affective, physiological and cognitive consequences of stress, it appears the authors stated two studies of affective consequences, but only one study of physiological consequences and one study of cognitive consequences. Suggestion: It is probably better to be more balanced. Two studies for each perhaps.

Authors' Response: We apologize if this was not clear, but we will try to clarify here in reply to the reviewer. If the reviewer still thinks it is not clear, then we welcome suggestions on how to further clarify. Both for being in nature and emotional social support we reported a study for the three components of the stress response (affective, cognitive, and physiological) and one study for the affective consequences of stress. It is important to clarify that the affective component of the stress response is different than the affective *consequences* of stress. Thus, for being in nature we cited the study of Beil and Hanes' (2013) that found an effect of being in nature on self-reported stress (affective *component* of the stress response), while the subsequent cited study (Marselle et al, 2014) found evidence of being in nature on symptoms of depression (affective *consequence* of stress).

For emotional social support we did the same; we reported a study for the affective component of the stress response and a study for the affective consequences of stress. We hope this clarifies. We did not make a change in the manuscript, but would be happy to consider one if the reviewer deems it necessary

Comment 5: Suggestion: Curious, are there null findings or mixed findings in the literature? It would be better to discuss those to communicate uncertainties in the literature Issue: Great that issues regarding replication crisis are mentioned. Suggestion: Would be better to add citations regarding replication issues in mental health or environmental psychology research, perhaps with examples if possible.

Authors' Response: We thank the reviewer for making this point. We think that answering this question is far from easy; null findings might be present in other areas of the health psychology literature that are less relevant to this work. However, we are not aware of mixed findings in regards to replications and null effects (and may not be aware of them due to publication bias). We specified this state of affairs in the manuscript:

"The psychological literature therefore contains an unknown proportion of unreliable and false positive findings that also characterize the field of stress regulation. For instance, in our previous meta-analysis we analyzed whether self-administered mindfulness and biofeedback were effective strategies to decrease stress. We detected an effect for both strategies. However, when we applied the same publication bias techniques as we intend to apply here, we found no more evidence that self-administered mindfulness and biofeedback were successful in reducing stress. The originally detected effect was thus largely due to publication bias (Sparacio et al., 2022)"

Comment 6: Great that possible differences between different conditions are discussed, but those descriptions are too brief in my opinion. Suggestion: Add more specific information regarding those studies and theories, e.g. "stress recovery theory (Ulrich, 1983)" and Social Baseline Theory (e.g. Beckes & Coan, 2011; Coan & Sbarra, 2015)" (p. 8).

Authors' Response: We thank the reviewer for this suggestion. We have now made some changes, such that stress recovery theory (Ulrich, 1983) has been described more in depth:

"According to the "stress recovery theory" (Ulrich, 1983), nature provides a restorative influence helping individuals recover from stress. Ulrich's (1983) theory relies on a psycho-evolutionary theorizing: Humans evolved in the course of centuries in natural places adapting both psychologically and physiologically to these types of environments. The argument is that when a stressor is encountered, an unthreatening natural environment might evoke feelings of pleasantness, decrease stressful thoughts, and promote physiological restoration (see also Ulrich et al., 1979)."

For what concerns Social Baseline Theory (e.g. Beckes & Coan, 2011; Coan & Sbarra, 2015), we added a paragraph (p.6) in which we explained why social resources regulate the perception of a threat in presence of a stressor:

"One particular theory, "social baseline theory" (e.g., Beckes & Coan, 2011) offers an account that can provide a mechanism for the stress-buffering hypothesis, as it suggest that social support and proximity to others reduces the perceived threat of a stressor and people can thus exert less effort in regulating stress (Coan & Sbarra, 2015; Ein-Dor et al., 2015). Stress reduction, according to the theory, is reduced because individuals can distribute the efforts needed to achieve particular goals with other people (e.g., partner, friends, family members, or even strangers), a phenomenon known as "load sharing". In one particular study illustrating this phenomena, people held hands with a partner or a stranger and were confronted with the threat of a (mild) electric shock. When people held hands with someone, areas related to stress were less activated when confronted with the electric shock and the reduction of stress was greater the more familiar the partner (Coan, Schaefer, & Davidson, 2006; Coan et al., 2017). "

Comment 7: Issue: It appears no justification is provided regarding the minimum no. of studies needed for subgroup analyses, and it is not clear. Suggestion: Please explain why 10?

Does this mean if there are 5 studies on physical social support and 15 studies on veral social support, subgroup comparison analyses won't be concluded? So the minimum no. of *each* subgroup is 10?

Authors' Response: We thank the reviewer for raising this point, which was also underlined by Reviewer 1 and we responded to there. Just to be clear, 10 is to be considered for the total effects by type of subgroup analysis, not by category. For instance, if there are 5 studies on physical social support and 15 on verbal social support we will conduct the relative subgroup analyses, as the total number of effects is 20. However, if the total number of effects is below 10, we will not run that subgroup analysis.

Comment 8: Issue: "Finally, we used social networks (Facebook Groups and Twitter) and mailing lists (SPSP, EASP, ESCAN)" (p. 10) – full names of mailing lists are not provided Suggestion: Please provide full names

Authors' Response: We have now provided full names of the mailing lists.

Comment 9: Issue: No justification for the exclusion criteria "participants were below 18 years of age" (p. 11). Suggestion: Please justify briefly.

Authors' Response: We thank the reviewer for this comment. We decided to focus our meta-analysis on a population of adults, as this is the segment of the population that is sampled more frequently in psychological research. This was an arbitrary decision and there were no particular theoretical reasons to select participants that are aged 18 years or older.

Comment 10: Issue: The "general coding" and "Main outcome" tabs are clear and well done, but there is no data tab (sorry if I misunderstand your coding sheet) with all the columns yet. Suggestion: I believe it is better to prepare in advance the data tab. You may check Yeung et al. (2021): https://mgto.org/exp-ma-rr-template-coding that we are developing. It is for experimental meta-analysis so feel free to adjust. It would be better to include columns for providing information about page numbers/justifications for coding.
Suggestion: It would be better to include a tab for article list, decision for inclusion/exclusion and record of contacting authors (see our template).

Authors' Response: We thank Reviewer 2 for making us discover this useful template for meta-analyses. We updated the file "CodingSheet - Being in nature & Emotional social support" by adding a sheet called "Actual coding (post data collection)" in which we will code the data of the articles that will be retrieved in stage 2. We also added another sheet in which we document attempts to contact the authors (we adapted the template you provided of Yeung et al., 2021).

Comment 11: Clear and transparent, well done! Issue: The README is too brief: https://github.com/alessandro992/Registered-report-meta-analysis/blob/main/README.md Suggestion: Please provide filenames and information for all files.

Authors' Response: We thank Reviewer 2 for the kind words for what concerns our script. We updated the README by providing filenames and information for all files.