**Review**

We thank the editor and reviewers for recognizing the potential of our work. We have addressed the points raised and have indicated the corresponding changes in **bold** within this document and the revised manuscript.

# Editor

I have now received two very thorough reviews of your manuscript, “Measuring Depression and Anxiety with 4 items? Adaptation of the PHQ-4 to increase its Sensitivity to Subclinical Variability”, from experts in the field. I also independently read the manuscript before consulting these reviews. Based on these reviews we believe that the work is very promising, but some revisions are necessary before it is suitable for publication.

The single biggest issue here (as also indicated by points raised by the reviewers) is that the IRT models used are opaque, because, as far as I know, it is not possible to freely estimate all parameters on a pair of items. Indeed, examining the results suggests to me that the model must be fixing parameters to accommodate the small number of items. I have never seen information curves that look like half of the curves generated in your analysis, which makes me suspect that the discrimination parameter is being fixed for one of the items through some algorithm. I attempted to determine what might be happening here but was unable to uncover how the mirt package handles this kind of situation. At a minimum, I and the reviewers need more detail on exactly what model was fit and how parameters were derived in what seems like an impossible situation.  
  
If I am correct about the model fixing parameters to permit estimation, I also have a suggestion for an additional analysis that would help clarify the results: In your second study, you of course have additional items that you think measure the same construct (the BDI and STAI). Conducting an analysis of your PHI items with the items from the complementary scale would eliminate issues generated by having too few items to allow free estimation. Of course, this solution is not perfect, but should help clarify your other results.

This issue appears major, as underlined by both reviewers. We hope to have addressed this point in the first response to Reviewer 1.

Some additional concerns include:

(a) There is contention in the literature about whether the original STAI actually measures anxiety (as opposed to negative affect more generally) and it is unclear to me whether the STAI-5 used here is responsive to this literature or not. Greater clarity would be useful here.

It appears that the editor’s concerns regarding the full STAI also extend to its short form (at least, we couldn’t find any evidence to the contrary). We thus underlined this limitation in the discussion: please refer to the related point raised by Reviewer 2.

(b) The terms “valid” and “validated” are used at various points in the manuscript more loosely than seems wise. For example, the typical biomedical statements are used that measures have been “validated,” which is not really a meaningful statement without further clarification. Validated for what use?

We have now removed the confusing ‘valid’ adjective from our discussion (lines 198).

(c) Line 17, should be “necessarily”;

This typo has now been fixed.

(d) I believe there is some contention regarding how high-quality data from Prolific is likely to be, and it does not appear there were any attention checks in Study 1; more broadly please describe any attention checks and how exactly participants were removed from the study and when this occurred (e.g., was it before analysis?);   
  
(e) please only use abbreviations after defining them (IAS);

This has been addressed.

(f) the description of people being “labelled” as having depression is needlessly in passive voice, and presented as if the label provides additional information beyond the self-reported data it is based on, which is not the case. This could be improved by saying that, for example, you coded people who self-reported x and y as likely having depression;

This has been fixed.

(g) it seems to me that more information from Study 2 would be helpful; the reviewer comments may be a useful guide here;

This has been addressed, see answers to reviewer 2.

(h) it seems inaccurate to assert that adding response options to the PHQ would be cost-free: For example, it adds at least some ambiguity as to whether new results correspond to previous reports in the literature. You may just mean it is relatively easy to do, which is fair enough.

We replaced “cost-free” with “low-cost” and, additionally removed the “no-downsides” statement in the abstract, which was indeed too bold.

# Reviewer 1

I thought this study tackled an interesting idea—one that might be relevant for either clinical psychologists or as a psychometrics tutorial (or both). There were some limitations related to the criterion validity measures as a result of this being part of a broader data collection effort. Most important for me, I didn’t understand the concept of latent variable modeling of a covariance matrix for 2 items (2 item variances, 1 covariance)—I wasn’t aware this was possible. So that’s probably my main concern. I hope the comments below can help the editor & authors move the project forward!

We thank the reviewer for raising this important point regarding the application of IRT to two items (which would indeed be problematic in a linear factor analysis framework, as there would be insufficient degrees of freedom to estimate model fit and test all parameters). We would like, firstly, to clarify how the IRT model differs. We will then present an additional analysis that compares the IRT-estimated latent variables against their ground-truth measures available in study 2 (BDI and STAI), showing high correlation and consistent patterns. Model identification is achieved by fixing the scale of the latent trait theta (i.e., by setting its population mean to 0 and variance to 1).

**Details on IRT**

The IRT model used here, the “Graded” model (Samejima, 1969) suitable for polytomous response categories does not solely rely on the item covariance matrix for parameter estimation. Instead, the estimation is done using full response pattern information, which under the hood estimates the probability of all observed response patterns (in our case, for 2 items with 5 response categories, there are 5x5=25 distinct response patterns). The likelihood function is constructed based on the model using these probabilities. Moreover, our primary goal in this study was to apply IRT to understand the 'coverage' of each response option relative to the underlying depression or anxiety continuum (the latent theta score), achieved by examining Option Characteristic Curves (OCCs), which are a direct output of the IRT analysis, rather than other IRT parameters and tests.

We clarified that in the text as follows (lines 131-140):

*“One natural methodological limitation pertains to the application of the IRT framework to pairs of items. While this is statistically sound (the graded model utilizes the full response pattern information and does not solely rely on the item covariance matrix for parameter estimation), it is important to underline that in our study's context, "the latent anxiety/depression dimension" merely corresponds to the amalgamation of the two items of the anxiety or depression subscale, and not to a more general and independent latent anxiety or depression factor.”*

**Valid, but is it meaningful?**

While the approach is statistically sound, we nonetheless agree with the more general point that reviewers might be implicitly making about the practical usefulness and meaningfulness of analyzing a 2-items scale with IRT. And indeed, we do agree with the limitation of this approach. In particular, we can underline the fact that in our case, the estimated latent factor is very sensitive to any changes in the response to the items. This, in a way, “inflates” (or rather magnifies) the specificity of each response option and its influence on the general score. Thus, the fact that the new option is being used prevalently is somewhat unsurprisingly accompanied by a distinct coverage of the latent score reflected by the OCCs. In other words, the specific coverage of the new response option could be seen as a natural consequence of the finding concerning the prevalence of the new response option (showing that this option is often picked).

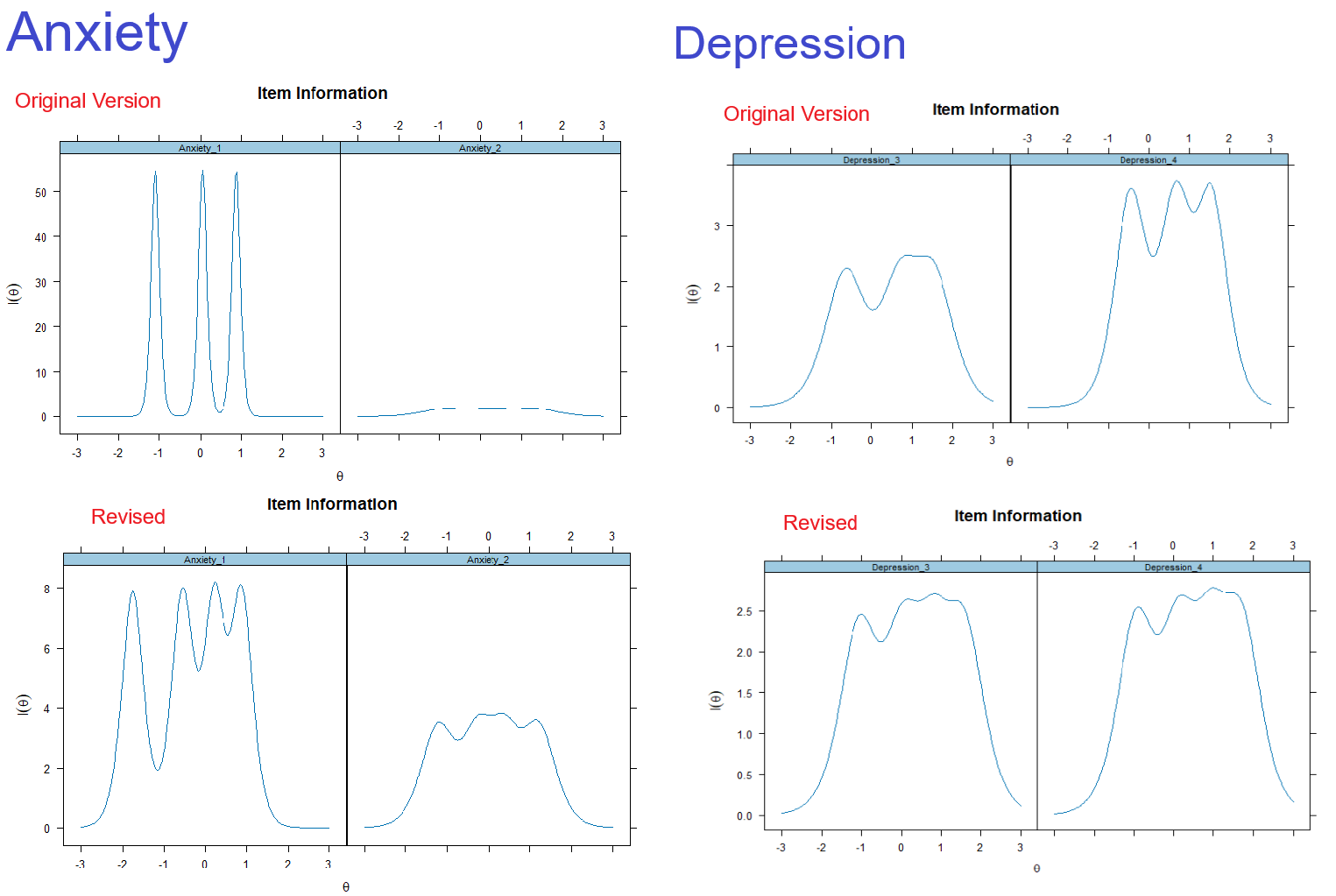
In general, we propose to keep the IRT analysis for completeness's sake, but downplay its importance as a key result of the study by underlining its limitations more explicitly, and by refocusing the conclusion on the most basic finding (lines 141-145):

“*In summary, the main take-away of this first study is that the "Once or twice" response option appears as a popular choice, which begs the question of its usefulness in capturing more fine-grained variations of the underlying dimensions as measured by independent tools.”*

We do indeed believe that, for study 1, the descriptive results on prevalence make for a more intuitive and insightful “first line” of evidence for the inclusion of the new response option that is then more properly supported in the follow-up study.

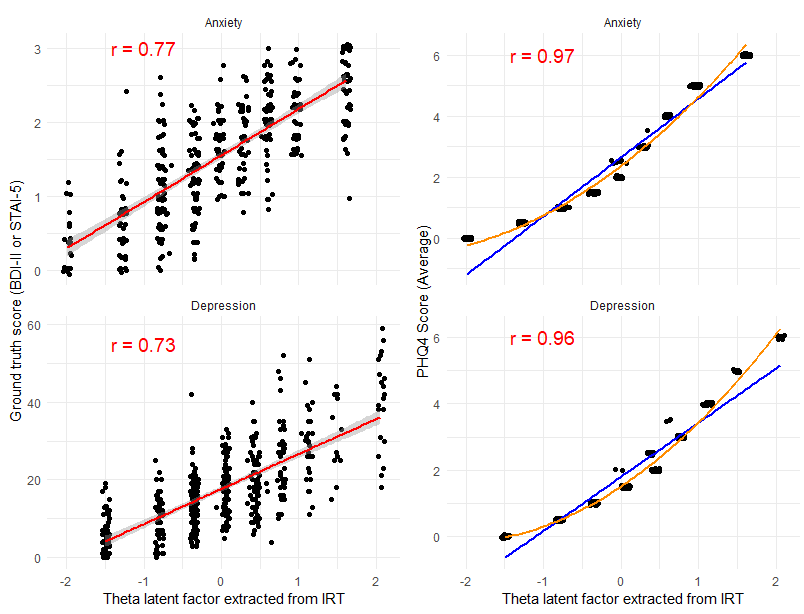
**Additional Analysis**

**Information Curves.** As suggested by the reviewer in the comment below, we leveraged the data of study 2, which contained both the original and revised versions, to visualise side-by-side the information curves of the original and the revised PHQ-4:



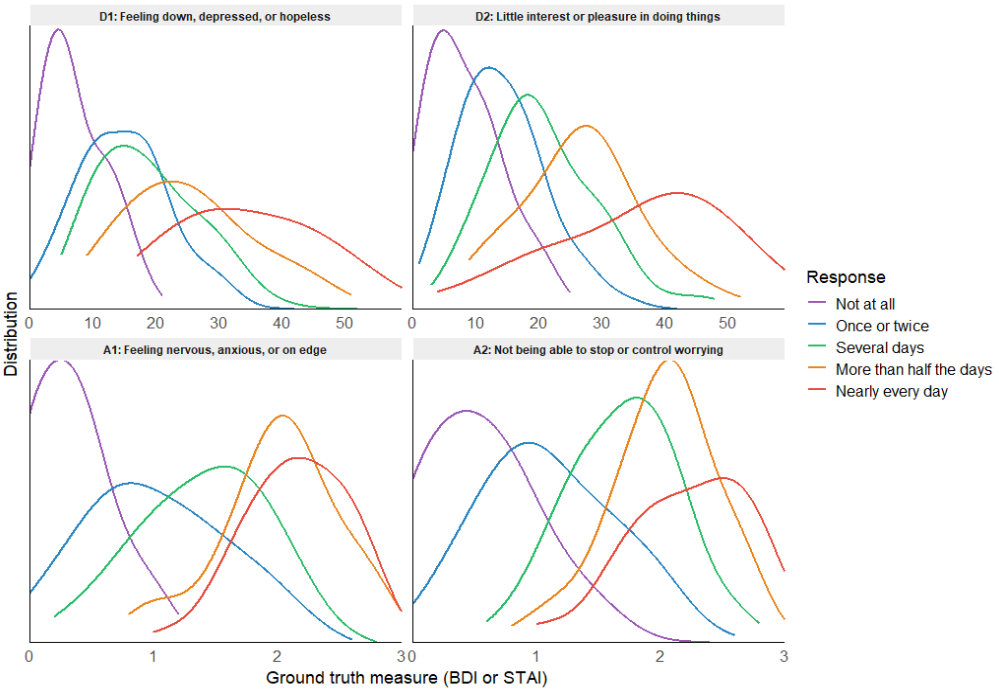
The main takeaway here from our perspective is that the original version, particularly the anxiety 1 item, has a highly specific coverage. The coverage is somewhat smoothed out by the introduction of the new response option (nothing ground-breaking, as it introduces more granularity).

**Correlations.** Additionally, we made use of our external measures of depression (BDI) and anxiety (STAI) available in study 2 to better understand the meaning and validity of the graded model. We tried to reinforce our confidence that the “latent scores” computed by the IRT were meaningful by comparing the extracted IRT theta score with the mean score and its ground truth measures:



As seen in the figure above, the composite score computed internally by IRT correlates highly with the respective ground-truth, which is reassuring - but is explained by the fact that the theta scores are very strongly correlated to the PHQ facet scores (computed by averaging) - although, interestingly, they are not exactly equivalent, but rather share a curvy relationship.

Finally, to further provide evidence for us and the reviewers for our main hypothesis that the new response option carries specific information, we analysed “empirical coverage curves” by plotting the distribution of ground-truth scores (BDI and STAI) as a function of different response options for every PHQ-4 question:



This is in our opinion the most intuitive and compelling evidence showing that the new response option added to the PHQ-4 does lead to a unique “coverage” (distribution) of an independent measure of the same construct.

While we added these results to the supplementary materials files available online, we think it is preferable to only include in the manuscript the current analysis, which we find more formal and streamlined.

**Motivation.**

What is the utility of more precisely describing subclinical variation? The authors didn’t quite explain that in the Intro, but it seems important for readers’ understanding of the study’s potential contribution.

We’ve added the following paragraph in the introduction (lines 26-36):

*“Enabling a more precise assessment of low-severity alterations is important, as milder symptoms are significant predictors of future clinical disorders and are associated with present functional impairments and reduced quality of life (Judd et al., 1998, Cuijpers et al., 2004). In addition, the growing reliance on large-scale, often online, psychological and epidemiological surveys, to monitor population-level mental health, evaluate interventions, or track responses to global stressors such as pandemics or geopolitical crises, demands tools that are both brief and sensitive to small but meaningful fluctuations.”*

I wasn’t sure if this article were meant as a sort of psychometric tutorial or whether it was meant to inform research and clinical practice with internalizing problems. I ask because I think for improved interpretability the article would ideally give more background/information/context (particularly in the Intro, Method, Discussion) in that specific area.

This article does fall somewhat in-between the methodological and psychopathological psychometry fields. We hope that having added the following and aforementioned paragraphs in the intro and the discussion helped with providing a bit more context and details related, in particular, to the importance of subclinical disorders. YOLO

**Method.**

I was super confused about fitting latent variable models to the (co)variation of a pair of items. I actually didn’t know this was possible in isolation (i.e., without borrowing information from indicators of other factors), because of degrees-of-freedom issues underpinning model estimation. How is this possible? And what do the parameters (factor loadings / discriminations, intercepts / difficulties) even mean; do they have the same interpretation as usual? I would doubt it? There’s a vague message about this at the end of study 1 (line 120), but it’s not explicit.

We hope to have clarified the approach in the response above.

I didn’t quite understand the reasoning for scoring the new option as 0.5 points. What are the other options? What are the implications of choosing 0.5 instead of assuming equal distance among all response options (including the new one)?

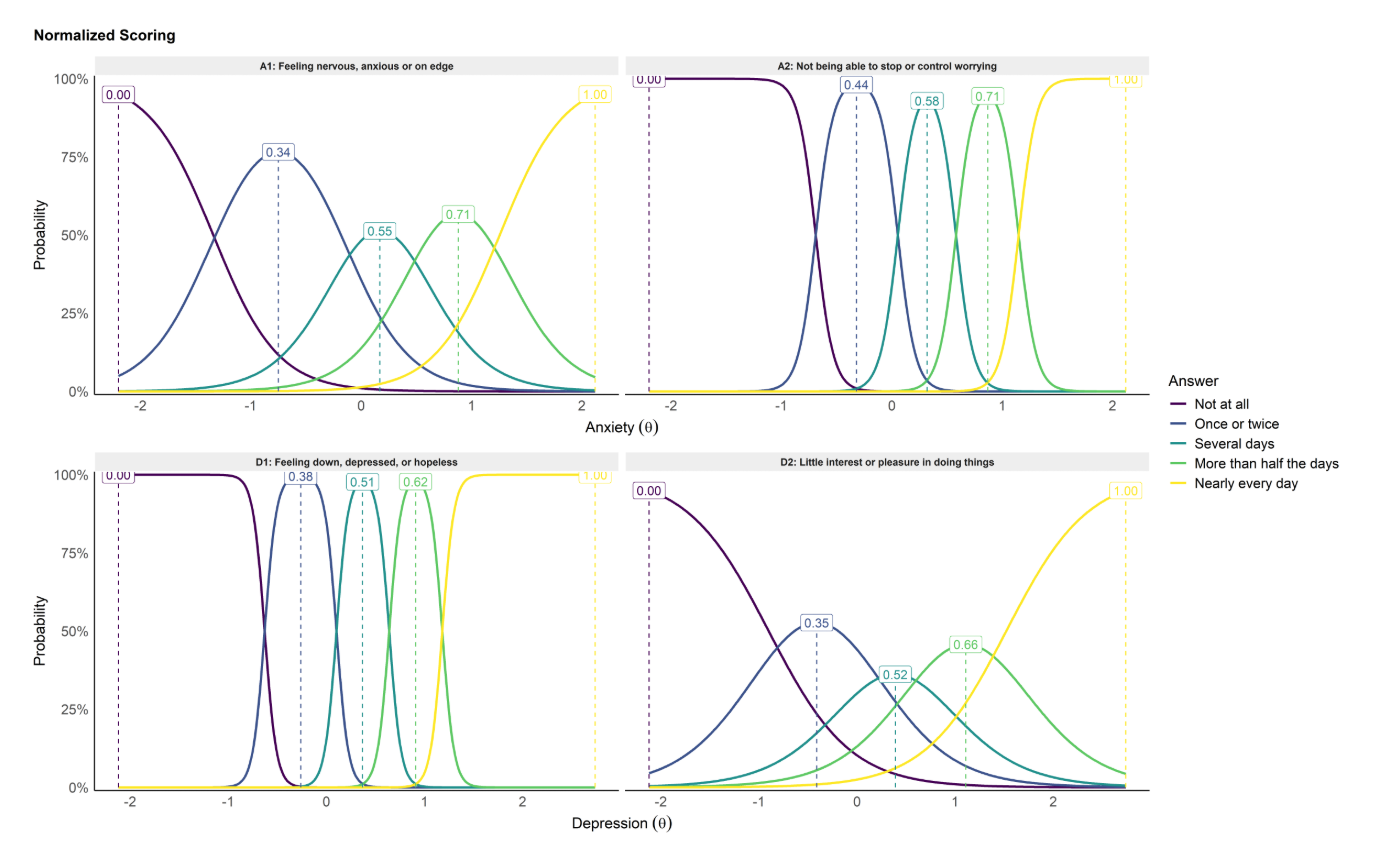
The conversion from a discrete option to a numeric score is always a delicate issue. In our case, we operated under a tradeoff of pros and cons, with a focus on preserving as much as possible the simplicity of the scoring approach and its comparability with the original scale.

The additional response option ("Once or twice") was scored as 0.5 to reflect its conceptual position between “Not at all” (0) and “Several days” (1), without altering the scale’s total range or cut-off points. This allowed us to preserve comparability with the large body of research using the original PHQ-4 scoring system, where established cut-offs guide screening and prevalence estimates. Assigning it a value of 1 (and shifting all others upward) would have altered the scale's interpretation and its maximum score, reducing compatibility across studies. Inserting the new response option as an in-between the two first options was a way to prevent this major issue. It also followed the conceptual idea of introducing variability “in-between” two existing established categories.

However, we actually do agree with the overarching point made by the reviewers, and the fact that this scoring arbitrarily assume that the new option sits perfectly in-between the two adjacent options, which might be not reflective of its true position in measuring the latent construct (and in fact, probably is not, based on the coverage curves of IRT).

However, this issue is in our opinion not addressed correctly by introducing it as a new integer option. Indeed, the current established scoring of the PHQ4 already carries this problem by converting fundamentally categorical choices into integers and assuming equidistant separation. This “continuous” treatment of discrete options is a well-established issue in the psychometric literature. In our opinion, it should be better addressed either by:

1. Treating the PHQ4 answers as categorical in statistical models by using monotonic or discrete models, which is doable for single items but not straightforward for Likert scales (which requires the amalgamation of multiple items into one score).
2. Create an empirically-sound categorical-to-numeric-conversion that would map each response to a number that accounts the non-equal distance between response options. Interestingly, this second option was actually explored in an earlier version of the manuscript, and the corresponding analysis is available in the supplementary materials:



We essentially assigned normalized non-equidistant values for each answer based on the peak of the IRT coverage curve. However, we decided against presenting that in the paper due to the limitation of our IRT model mentioned above, and the fact that this new scoring approach required a validation on its own, which appeared to be out-of-scope given the goal of the study, that of introducing a minimal and “low-cost” change with respect to score interpretation and literature comparability.

However, we do wholeheartedly agree with the reviewers and the literature about the fundamental psychometric issue pertaining to the discrete-to-numeric approach used prevalently in psychology and medicine, which we hope might be addressed through better awareness and new analysis methods.

We added the following details in the manuscript (lines 177-186):

*“Participants were randomly assigned to complete either the original or refined version of the PHQ-4, which included one additional response option ("Once or twice"). To preserve comparability with the original PHQ-4 scoring system and avoid altering the scale’s total score range or established cut-off thresholds, we assigned the new response option a value of 0.5, placing it midway between "Not at all" (0) and "Several days" (1). While this scoring assumes equal spacing between response options - a common but imperfect convention in ordinal scales - it offers a pragmatic compromise between conceptual fidelity and applied utility.”*

Additionally, note that we have moved the mention of this new scoring to study 2, as it wasn’t used in study 1 (where it was treated as ordinal).

Would it have helped to show histograms & item information curves for the original vs refined PHQ side by side?

The item information curves for the original vs. refined are shown above for the reviewers, but we opted for not including them in the main manuscript to keep it concise and not redundant.

**Open materials.**

Thank you for taking the time to do this so well. The materials were very easy to navigate.

We thank the reviewer for recognizing our efforts in this area.

**Minor issues**

A. I didn’t really understand the first part of the title. I wasn’t sure it was super closely connected to the main objective of the study.

Assuming the reviewer refers to the question at the beginning of the paper’s title, it is related to the findings of the study (especially study 2), adding supporting evidence for the PHQ-4’s validity, even for low-severity symptoms. The (rhetorical) question format was also meant to attract a readership interested in choosing suitable tools that answer this specific goal (namely, to measure mood disorders with few items in subclinical samples).

B. On line 31, the authors say “latent measure”, but I think this is probably an oxymoron.

We replaced it with “of the construct”.

C. Typo here: “The two depression items captured 82.8% of the variance of its latent trait (𝜃𝑑𝑒𝑝𝑟𝑒𝑠𝑠𝑖𝑜𝑛), and the opposite pattern was found: the first item had a higher precision

(𝛼 = 16.46) than the first (𝛼 = 2.41).”

This has been fixed.

# Reviewer 2

This manuscript examines the changes in the functioning of the PHQ-4 when a low severity response option is added. Within classical test theory, increasing the number of items will increase internal consistency. Within modern measurement methods, such as IRT, increasing the number of response options will increase the range of the underlying trait for which information will be assessed. The results of this work bears this out empirically. Relying on two moderately sized sample enhances the rigor of the work. The visualizations of the associations in scatterplots and density plots aid in evaluating the associations. There are only a couple of comments for the authors to consider in enhancing the presentation.

In both studies, there was a decision to rely on a score of 0.5, deferring to the ‘original version.’ I am unsure what is meant here. Given the empirical focus, why not estimate factor scores that incorporates difficulty parameters? This would take unevenness in distances between response options (which is not possible in more classical test scoring methods). Some comment about retaining this decision/recommendation could be provided in the recommendation.

See response above to reviewer 1.

Given the ordinal scaling of the variables, it is not clear whether internal consistency estimates have been computed using this information. It is not clear how the RMSEA values are being estimated within the paragraph on reliability.

We thank the reviewer for noticing this confusing bit: indeed, the RMSEA values presented are from the IRT model, which is only introduced in the next paragraph. For Cronbach’s alpha, we firstly replaced the mention of “the reliability” to “the consistency”, which is more appropriate. However, there is to our knowledge no particular issue with applying it to ordinal data (https://stats.stackexchange.com/questions/4691/internal-reliability-for-an-ordinal-scale). We have clarified both aspects by moving the RMSEA mention in the IRT paragraph and adding (line 94):

“responses were treated as ordinal”

I am also unsure of how to evaluate an IRT model with two indicators. If this were in a fully formal SEM/CFA framework, you could estimate a model with two indicators if you constrain (1) the variance of the latent factor to 1 and (2) the factor loadings to be equal. In the IRT model, it is not clear whether this is occurring under the hood. If it is, these imposed assumptions should be clarified in the text.

See response above to reviewer 1.

There is some verbiage that also does not appear correct. The factor model is set to reflect the indicators; thus, the factor is explaining proportions of variance in the item sets. Or, are the authors presenting latent factor reliability?

We thank the reviewer for spotting this mistake! Indeed, we fixed it to the correct interpretation, that for example *“the latent anxiety dimension captured 89.2% of the total variance across the two items”* (and not the other way around).

The item characteristic curves could be nicely complemented by the item information curve. This is especially true for Study 2 where this could provide a strong visual comparison of the range and amount of information assessed between the two versions of the PHQ-4.

We have made and copied the figures (in the response to reviewer 1), however, we would prefer not to include them as it might inflate the perceived importance of IRT in this paper, which as we mentioned above can only give limited insights anyway. We prefer to present its application in study 1 as providing first elements of evidence, which leads to a more robust verification in study 2 (which is the most important part of the paper).

The reporting of the interaction was hard to follow. The interaction term itself did not appear to be reported.

We thank the reviewer for the suggestion, as our presentation was indeed confusing: in fact, the 2 parameters we reported are the “effect of PHQ version [on the intercept]” and the “effect of the PHQ version [on the slope]” (which is colloquially referred to as \*the\* interaction effect). However, both parameters represent a modulation (of the intercept and the slope, respectively) by the refined version relative to the original version. We clarified by adding the following explanations (lines 233-240):

*“A linear model testing the interaction effect Δ of the refined condition on the intercept (representing by how much the value of the outcome when the PHQ score is 0 changes for the refined version compared the original) and slope (its increase or decrease by the refined version compared to the original) of the relationship between the PHQ-4 depression score and the BDI-II total score was fitted.”*

Note that we did not include the parameters (intercept and slope) for the original PHQ version (only the effect of the refined condition) as the values are irrelevant with regards to our question, but the reader can find all information in the parameters tables available at: https://dominiquemakowski.github.io/PHQ4R/study2/analysis/2\_analysis.html#relationship-with-bdi-2

I would advise the authors to note that the work is limited by the reliance on the BDI. There are studies showing that the BDI is pretty insensitive to low severity depression (https://doi.org/10.1016/j.jclinepi.2013.04.019; https://doi.org/10.1002/mpr.1348). With this low sensitivity/reliability of information, there is also reduced power to detect associations at lower levels of severity. It is possible that the STAI may show similar limited sensitivity at one end of the dimension.

We thank the reviewer for underlining this and pointing to these papers, which indeed slipped our literature review. Although we could interpret the observed increased sensitivity for low-severity observed with an insensitive ground-truth measure such as the BDI as a strengthening of our claims (as we would expect that the increase of sensitivity would be even more pronounced with a more sensitive ground-truth measure), we included this fact as a limitation and future directions (lines 337-355):

*“One of the potential limitations of our study includes the choice of the measures used as external "ground-truth" of depression and anxiety, namely the BDI-II and the STAI-5. For instance, there is ongoing debate about the STAI's discriminant validity (which might extend to its short form used in the present study). A recent meta-analysis reported that trait anxiety scores were more strongly associated with depressive than anxiety disorders (Knowles et al., 2020), suggesting that the scale may rather capture general negative affect than specifically anxiety. Regarding the BDI-II, existing evidence that suggests its relative lack of sensitivity for low-severity depression (Olino et al., 2012) - comparable to similar instruments of its size (e.g., CES-D), and an improvement over the BDI-I (Wahl et al., 2014) - put into question its choice for our goal of showing increased sub-clinical sensitivity. Thus, future studies should verify these findings with alternative measures of anxiety and depression and clinically assessed populations.”*

More could be said about whether this work is focusing solely on applications of the PHQ-4, or whether these findings should generalize to other measure with similar kinds of measurement properties.

We added the following point in future directions (lines 378-386):

*“While adding granularity to the response format appears useful for the PHQ-4, it is possible that similar benefits could be found with other scales and measures. While the use of a limited, highly discriminating set of response options is understandable for specific applications (e.g., clinical diagnosis), we recommend future studies to investigate response format and its potential improvement for other scales used in online surveys and general population research.”*