Dear Drs Renner and McIntosh,

Thank you for your editorial letter dated February 2, 2024, providing feedback on our manuscript "Evaluating Loneliness Measurements across the European Union". We very much appreciate the quick and thorough turnaround by both of you and all reviewers. We apologize for the delay in responding. Our team was confronted with some health issues.

In what follows, we address your concerns point-by-point.

Our most major revision was to follow your suggestion to already conduct the exploratory analyses and register our final model. We explain in more detail below (and now better in our manuscript what our approach was for the exploratory and confirmatory factor analyses (which in the exploratory set we ran side-by-side to decide the most optimal factor structure).

We hope that this revised version meets the high standards set forth by PCI-RR for In Principle Acceptance.

Please note that the microdata for the EU survey will be released together with a book by Springer and the data is under embargo until then. We propose that the data is available for verification to the reviewers to the following private OSF component https://osf.io/nj69a/?view_only=6570eb840e46463ab0527633edf44355, but that we will only publicly release the data when the embargo is lifted (it is very likely that the publication of the current manuscript, if accepted, will only occur after the embargo is lifted in any case).

In case you have any further questions or concerns, please don't hesitate to contact us.

Sincerely and on behalf of all co-authors,

Hans Rocha IJzerman

Methodological Approach and Analytical Flexibility

Editors Comment 2: Importantly, pre-registration of the full exploratory-confirmatory pipeline is a variation on the traditional pre-registration of the confirmatory analysis after performance of exploratory analysis. This makes it important to establish how you will eliminate analytic flexibility based on knowledge of the results. In the section on exploratory factor analysis (L389-392), you state that "If the factor structure typically used in the literature did not match the most optimal structure identified through exploratory factor analysis, we decided on a structure for the subsequent analyses. Again, our decision aimed to balance theoretical parsimony with model fit."

The proposal here (as also stated elsewhere in the plan) is to make a decision based on a balancing of theoretical parsimony with model fit. This is not an algorithmic decision - it involves a degree of subjective judgement. How would readers of the Stage 2 manuscript be assured that you did not revise your judgement of the best EFA model based on the subsequent CFA stage? Is it possible to unambiguously specify in advance the decision process for choosing the best EFA model? If not, do you think it would be worth running this step first, and submitting the Stage 1 for the CFA instead?

Authors' Response: Thank you for outlining this concern. In the previous iteration, we were planning to run EFA and CFAs side-by-side to determine the most optimal factor structure in our exploratory set. This decision was based on the idea that, while there is a pretty solidified approach to the factor structure in the literature (thus justifying CFA), these factor analyses have not been run in the countries we are examining (thus leaving the chance for considerable cultural and contextual variations in factor structure, thus justifying EFA). Once we decide on an optimal factor structure, we would test the factor structure (after pre-registration) in the confirmatory set. We have used this exploratory-confirmatory pipeline elsewhere in a registered report (see e.g., Wittmann et al., 2023).

Based on how the manuscript has been interpreted and rereading our own words, we understand that not only our explanation wasn't well written, while our approach may perhaps also have been too complicated without much gain on rigor. We decided we needed to streamline this approach. As a result, we follow your suggestion and have already conducted these analyses for the exploratory set (both with EFA and CFA), and are now pre-registering our final model (for which we will only use CFA, not anymore EFA).

In the end, we think that by skipping the registration of the exploratory set we will lose little rigor and simplifying the explanation will help for us as authors, as well as for the readers.

Ethical Review and Bias Control

Editors Comment 3: Finally, our understanding is that the data exist within your research team (i.e., reside with one of the authors). Can you please explain your claim to Level 4 bias control, and whether any additional measures can be put in place to reduce the possibility of bias? See Section 3.6 and Table 1 here for more details: https://rr.peercommunityin.org/about/full_policies.

AND

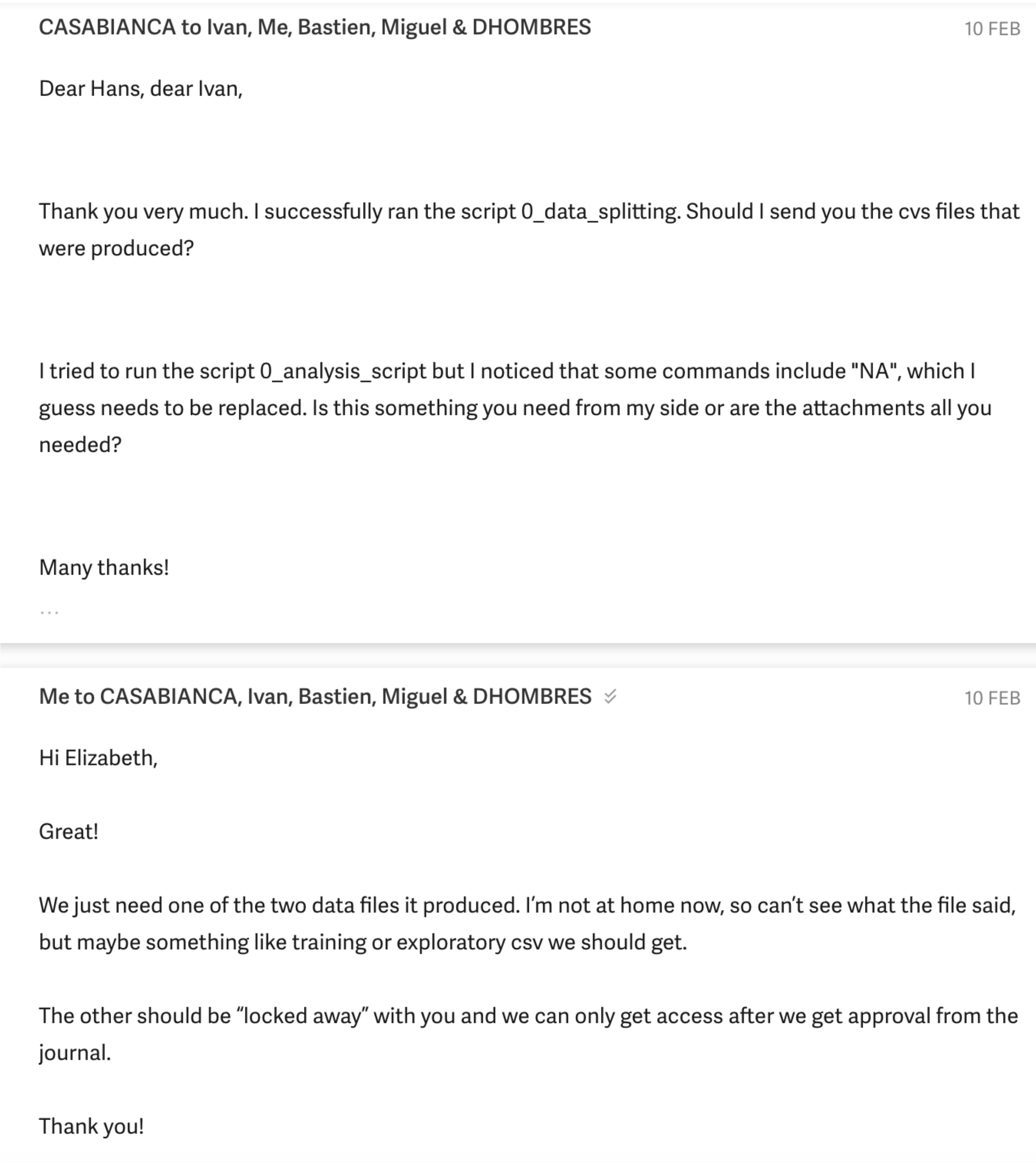
Drew Altschul: The authors distancing of themselves from the sample partitioning procedure is appropriate, and the code matches the functionality required, though I hope this might be described in a bit more detail in the actual text in a future version.

Authors' Response: Thank you for pointing to this issue, as this is an important concern. The data is indeed already available and is currently held by the Joint Research Centre. To minimize flexibility, Miguel Silan, Bastien Paris, Ivan Ropovik, and Hans IJzerman didn’t have access to the full data. Elizabeth Casabianca has the full data, and is in charge of splitting the folds into exploratory versus confirmatory. As Elizabeth Casabianca and Béatrice d’Hombres have seen the data before, they are not involved in the process of making inferences about the data. In our previous manuscript, we had described this in the following sections as follows:

Author note: “Béatrice d’Hombres and Elizabeth Casabianca have reviewed the data before the submission of this Registered Report. The final decisions for data analysis, hypothesis, and inferences are all with Bastien Paris, Miguel Silan, and Hans IJzerman.”

Sampling plan (Table on P. 3): “Elizabeth Casabianca, who is not involved in drawing inferences from the analyses, will supervise the splitting of the folds.”Elizabeth Casabianca, who is not involved in drawing inferences from the analyses, will supervise the splitting of the folds.”

As further support that this is our procedure, we are submitting a screenshot of our communication:



CASABIANCA to Ivan, Me, Bastien, Miguel & DHOMBRES 10 FEB

Dear Hans, dear Ivan,

Thank you very much. I successfully ran the script 0_data_splitting. Should I send you the cvs files that were produced?

I tried to run the script 0_analysis_script but I noticed that some commands include "NA", which I guess needs to be replaced. Is this something you need from my side or are the attachments all you needed?

Many thanks!

...

Me to CASABIANCA, Ivan, Bastien, Miguel & DHOMBRES 10 FEB

Hi Elizabeth,

Great!

We just need one of the two data files it produced. I'm not at home now, so can't see what the file said, but maybe something like training or exploratory csv we should get.

The other should be "locked away" with you and we can only get access after we get approval from the journal.

Thank you!

We realize that this may not have been 100% clear, so we adjusted our author note as follows (which now includes our adjustment of having seen the exploratory fold):

> Author note: Béatrice d'Hombres and Elizabeth Casabianca have reviewed the data before the submission of this Registered Report. The final decisions for data analysis, hypothesis, and inferences are all with Bastien Paris, Miguel Silan, Ivan Ropovik, and Hans IJzerman. Paris, Silan, Ropovik, and IJzerman did not have access to the full data prior to In Principle Acceptance. They received data for the exploratory fold from Casabianca, who kept the confirmatory fold until after In Principle Acceptance.

In addition, in the text on p. 19-20, we have explained the splitting of the data as follows:

> Elizabeth Casabianca, an author not involved at the level of data contingent choices, chose a fixed random seed number and used a dedicated R script to automatically partition the dataset into two folds—exploratory and confirmatory—of equal sample sizes. Stratification was performed based on the country variable to maintain a consistent representation of countries between folds. We first conducted the analyses of the measurement properties of the loneliness instruments on the exploratory fold. Once we had analyzed the exploratory fold, we then wrote our conclusions and – based on the findings – pre-registered resulting hypotheses prior to testing them in our confirmatory fold.

To move one step higher in the PCI level of bias classification is not possible, as this would require that none of the authors had access to the data.

That said – and as also pointed out by Drew Altschul – to reduce possible bias as much as possible within the given rung of bias classification, we employed a cross-validation procedure, splitting the analysis into two stages, exploratory and confirmatory, while pre-registering the hypotheses and analysis procedure before the confirmatory analyses.

Joe Bathelt: However, a statement about independent ethics review is missing. It is recommended to include a statement regarding independent ethics review to enhance the report's comprehensiveness.

Authors' Response: Thank you for pointing out this shortcoming. We added on p. 16 the following statement:

> The JRC Research Ethics Board (REB) reviewed the project for the data collection.

Data Handling and Analysis Details

Drew Altschul: Inclusion and exclusion criteria are generally good, but there are some missing bits – e.g. the authors handling of missing values/NAs. I was able to understand what the authors intend to do through looking at the analytic code, but there is almost no mention of this in the text (this sort of thing should be improved).

AND

Again, I would like to see a bit more about how the authors plan to handle unusual data, like NA's, potential outliers (or if we shouldn't expect outliers…).

Authors' Response: We agree this aspect of analysis is always important to clearly report in the manuscript and apologize for this omission. We handled the missing data by listwise deletion when estimating latent variable models and by pairwise deletion for zero-order correlations. Of course, we apprehend that listwise or pairwise deletion is an acceptable solution only when data are missing completely at random, which they seldom are. In the case of our research, we had a choice between two mutually exclusive choices. Either we properly model the ordinal character of the loneliness scales items using the WLSMV method, letting the model estimate the thresholds and use listwise deletion, or suboptimally model the Likert-type items as continuous indicators but use Full Information Maximum Likelihood to simultaneously estimate the model and impute missing data. We decided that the former option is more optimal as there were only 1.9% of missing data overall. We anticipate that in such a scenario, the choice of missing data treatment has substantially smaller effect than improper modeling of ordinal-level items as continuous.

Regarding outliers, we chose not to identify or exclude any outliers due to two reasons. First is the nature of the measurements producing data with bounded variance, where no excessively influential values are possible. Second, due to liberal inclusion criteria, we had no theoretical reason to expect that some of the participants were not members of the target population.

In the revision, we report the proportion of missing data and clearly outline and justify the missing data treatment method, separately for latent variable models on p. 22:

> For all latent variable models, we handled the missing data using listwise deletion, as only 1.9% of the data for loneliness measures were missing. Here, we preferred the ability to directly model the ordinal character of the data using WLSMV over imputing the little amount of missing data by Full Information Maximum Likelihood.

and for zero-order correlations on p. 26:

> For the estimation of zero-order correlations of factor scores, we handled the 1.9% of missing data using pairwise deletion.

Joe Bathelt: Making the code available prior to the analysis is another strong point. However, this code could be better documented. While the input and output parameters are clearly defined, the purpose and parameters of various helper functions are less clear. Improving the documentation, especially regarding the purpose and parameters of various helper functions, would enhance the code's clarity. It would also be helpful to know if the code was tested with simulated data.

Authors' Response: We have gone through, revamped many parts and commented on the entire analytic R code. We believe it is now much more clear what part of the code does what. We have now decided to no not test the code using simulated data, since the exploratory fold data served that purpose.

Transparency and Code Sharing

Drew Altschul: I would encourage the authors to put their code on github or some other open versioning system so if changes are made, they can be transparently followed.

Authors' Response: Thank you for this suggestion. Now, we employ Github for archiving the code and version control. Additionally, we linked the Github repository with the project's OSF page so that both repositories are updated synchronically. Please see https://osf.io/7u4e8/.

Specific Analytical and Writing Feedback

Drew Altschul: "given that no comprehensive data exist on its factor structure" – what exactly does "it's" refer to hear?

Authors' Response: Thank you for pointing this out. The word "its" referred to the given measure (i.e., that there is little evidence about the factor structure of DJGLS-6 and T-ILS). In the revision, we make this point explicitly. We have now revised this to clarify it is about both questionnaires in the following text: "Given that no comprehensive data exists on the factor structure of the DJGLS-6 and T-ILS in samples from the European Union, we are not very certain of these a priori hypotheses."

Drew Altschul: "Following these analyses in the exploratory fold, we will decide – per country" – decide based on what? This isn't particularly clear here; it is somewhat better in the main text but really this ought to be fleshed out more to reduce researcher degrees of freedom.

Authors' Response: Thank you for pointing this out. We suspect this was mostly a shortcoming in how we explained our analyses, as well as the complexity of our procedure. To simplify (again without compromising rigor), we have now already run the exploratory analyses, so that this point, hopefully, has automatically been addressed. In the revision, we tried to be clear about what the analysis plan was. Following the analyses in the exploratory fold, we plan to decide – separately for each country – what is the best-fitting factor structure for DJGLS-6 and whether a unitary factor model fits for T-ILS. We will then attempt to cross-validate the empirically identified factor structure for the DJGLS-6 and the one-factor structure for T-ILS in the confirmatory fold using confirmatory factor analysis. As the T-ILS consists only of 3 items, no other factor structure theoretically and pragmatically makes sense.

Drew Altschul: LNs 64-73 are a good first paragraph, but the authors jump too quickly in the next paragraph. What is loneliness and why is it an issue now? The authors cover some of this in the 2nd section of the introduction, but I'd really like to see some broader scoping text that contextualizes this issue as more than just a "strategic priority".

Authors' Response: We appreciate this comment a lot. There are two comments here: the order of explanation and the completeness of the information.

As it concerns the order of explanation, there is a balance to be struck between providing a short – and non-jargony – introduction while at the same time giving complete information. We believe that this is a stylistic preferenc: we prefer a short introduction, where we flesh out the landscape of loneliness and our current understanding of it only later, in the body of the text. We thus prefer to keep this section as is.

As it concerns the completeness of the information, we have followed the reviewer's suggestion and split the paragraph that in the previous manuscript started on LN 86 into one that covers loneliness' effects on health and longevity and one on loneliness' economic costs, adding the following information:

Loneliness impacts health and longevity similar to other clinical risk factors (Holt-Lunstad et al., 2010; Pantell et al., 2013). Research suggests, for instance, that a one-point increase in loneliness is associated with a 26% increased risk of early death consistently across different demographic groups (Holt-Lunstad et al., 2015). Loneliness is associated with cardiovascular disease, hypertension (Hawkley et al., 2010; Valtorta et al., 2016), and with a greater decline in activities of daily living and motor performance (Perissinotto et al., 2012; Buchman et al., 2010).

These impacts on physical health translate to economic costs…

We hope that this revision suffices. We do agree it is important to point out loneliness' consequences, but, at the same time, don't want to take away from the important discussion on measurement and its potential shortcomings.

Drew Altschul: LN 89 – please clarify: "increased spending in mental health care" by whom?

Authors' Response: This is an important comment. We went back to the Meisters et al article, but could not find who was paying the bill. We suspect that this is related to the complexity of estimating this number. In the Netherlands, people are typically insured through a combination of public and private insurance. That means that the government pays part of the costs. What remains is paid through private insurance, and a part of that is funded through a person's deductible. It’s probably not impossible to estimate who pays what, but it wasn’t reported in the Meisters et al article.

Drew Altschul: LN 90 – not sure this is the best word choice – is there such a thing as “wanted” loneliness?

Authors’ Response: Agreed. We have removed the mention.

Drew Altschul: LN 113 – something to consider more here are these sampling differences. Or if not here, then in the discussion, since sampling differences come in many varieties, and can be extremely important, but this report only really covers national differences.

Authors’ Response: We do agree on this point. We tried to fit it into the discussion, but when we tried to rewrite, it felt that we would go onto too big of a tangent. We propose talking about sampling differences, and guidelines for reporting about samples, in the discussion.

Drew Altschul: LNs 147-156 – this is all fair criticism, but in the interest of being balanced before the authors present the results, I think there needs to be a bit more on what single item measures can do, or have done, or what we have learned from them. More brief lit review of this along the lines of the following paragraph on what “insights” composite indexes have yielded (LNs 157-162).

Authors’ Response: We have rewritten the concerned part. In our view, a balanced presentation of the psychometric and pragmatic tradeoffs involved in the choice of single-item or multi-item measures is one that provides all substantial arguments/facts. In the revised version, we have integrated the psychometric differences between the two types of measures into a single paragraph and try to balance the psychometric arguments (which are naturally in favor of multiple-item measures) with presenting the choice as a tradeoff that in reality has to conform to pragmatic constraints and consider the diminishing returns of adding items.

Our main argument that we try to convey is not that single-item measures are inherently bad (although they tend to be suboptimal to multiple-item measures), but that we have far fewer psychometric tools to analytically examine their qualities for measuring the underlying constructs.

Drew Altschul: LNs 195, 196 – here the authors write “alpha” and “tau”, but elsewhere they use the proper Greek symbols. I suggest the authors choose one or the other for consistency, and I also suggest they choose to use Greek characters.

Authors’ Response: In the revised manuscript, we now consistently use Greek letters, except for the first mention of both alpha and omega, where we give the English name of the letter along with the Greek symbol for clarity.

Drew Altschul: LNs 250-251 – it's given elsewhere but I think it should be here too – state that 0.6 is the cutoff value, or tell the reader where to find it.

Authors' Response: In the given part, the cutoff is now defined explicitly.

Drew Altschul: LNs 256-257 – "to the level of shape and CI of the correlations" – I don't understand this phrase. Could the authors please rewrite this and clarify?

Authors' Response: We apologize for being unclear. We have rewritten the concerned part. What we intend to do take the results of the nomological network analysis and preregister these latent correlations as the predictions to be cross-validated. We will then use the confirmatory fold to examine whether the results from the exploratory fold replicate. In the context of cross-validating the nomological network results, we will test and assess the replication success as follows: We will apply Fisher's *z*-transformation to the correlation coefficients from exploratory and confirmatory fold and calculate the *z*-score for their difference. We will then use a BIC approximation (implicitly assuming a unit information prior) to compute Bayes factors (Wagenmakers, 2007) to assess to what degree do the data support the H0 of no difference between the correlations. We will deem the given correlation effect successfully replicated either if both correlations will be significant, above .10, and in the same direction, or in case the BF_{01} (in favor of the null) will be larger than 3 (taken as an indication of equivalence of the correlation coefficients). We have changed that from the rather nonspecific plan to test the CI overlap.

Drew Altschul: LNs 260-261 – so how will the authors proceed if they have no predictions? I have an idea since I've read the whole paper over, but I think the authors need to give a little more information here about what they're going to do.

Authors' Response: The goal was to examine whether the psychometric meaning of the measured constructs is equivalent across different cultural contexts by systematically testing what level of invariance do the data generated by the measures support. We thus engaged in ordinary measurement invariance testing, which does not require any a priori predictions. We have extended the description of the goal of invariance testing and also carried out some revisions in the Measurement Invariance section to make it clear what we've done and why.

Drew Altschul: LN 272 – what are "ex-post weights"?

Authors' Response: Apologies for unclear language. These are ex-post (calculated after data collection) sampling weights to account for sampling probabilities that do not match the population proportions. The dataset that we worked with include sampling weights accounting for the following population characteristics: age, gender, educational attainment, and NUTS region of residence based on available data from Eurostat. In our analysis, all aggregate (across countries) latent models employed sampling weights for each country, to balance out unequal sampling probabilities caused by the fact that sample sizes across countries were similar (while country population sizes vary widely).

Drew Altschul: LNs 314-317 – I'm not sure these Omega values should be included here, it goes against the flow. For instance, Omega_u-cat: the u_cat part hasn't been described yet and the reader may not know what this is. It also isn't clear just what segment(s) of the sample these Omega values are for. Please move them, or clarify in case I'm missing something.

Authors' Response: We agree and have removed the omega values from the given part. These reliability estimates are now presented in full detail in the Results section. We have also simplified the reporting of omega coefficients. Previously, we planned to report omega_u-cat for unidimensional categorical item scales and omega_h-cat, the hierarchical estimate for multiple-factor solution. This is probably suboptimal as far as practice is concerned – likely, the given scales will be used individually and it is more practical to evaluate reliability evidence for each scale separately. Now, the manuscript uses only a single omega coefficient (the former omega unidimensional categorical), simply denoted as ω.

Drew Altschul: LNs 349-350 – this is actually not what the authors propose to do with the T-ILS in their analytic code, and I agree with the analytic code. So this should be rewritten – it is only for the DJGLS-6 that exploratory factor analysis will be used.

Authors' Response: Of course, you are right. Also, a three-item unitary-factor structure is just-identified (has zero model degrees of freedom) so fit to the data cannot be evaluated. For T-ILS, we therefore carried out the same analysis except for the factor analysis part, where the three-item structure does not allow for formal testing of the factor model and we assessed the internal structure by the adequacy of factor loadings only.

We have revised the concerned part where we now explicitly separate the DJGLS-6 and T-ILS analysis pipelines in that respect.

Drew Altschul: LNs 368-377 – I leave this decision up to the authors, but I don't personally think this approach is necessary. The authors could just look at the EFA's of 1, 2, and even 3 factor solutions and then evaluate model fit, rather than using 2 somewhat subjective criteria. But ultimately I leave this at the authors' discretion.

Authors' Response: We agree that a procedure that relies on a single method provides a more straightforward and formally stronger inferential power. In the revised version, we chose to determine the most optimal factor structure solely based on the Empirical Kaiser Criterion (Braeken & Van Assen, 2017). Compared to parallel analysis, this method tends to perform better with measures such as ours. Parallel analysis will only be used as a robustness check. Most likely, these two methods will yield the same factor structure. In case they will not, the reader will get the chance to see the additional layer of uncertainty associated with our results and consider how much similar/different would our results be if another – still psychometrically sound – method would have been used. We think this plan provides for both, unequivocal inferential procedure and transparency with regards to methodologically justifiable alternative approach.

We have revised the concerned part where we describe the inferential method for the exploratory factor analysis.

Drew Altschul: LN 383 – this is somewhat correct and somewhat incorrect. Firstly, the authors leave space for RMSEA confidence intervals in the results section, which is a good thing to do, but those 90% CIs are akin to pCLOSE, so ultimately they are still using significance testing here. That's fine, though the authors should definitely mention the fact that they're going to be using RMSEA confidence intervals here, before the results.

Authors' Response: Yes, you are right, 90% RMSEA CIs can be used for null hypothesis significance testing. In our study, however, we do not use the RMSEA CIs for any inferential decisions and they are reported solely for completeness. Likewise, although we do not assess the fit of the models to the data based on the chi^2 test, we report the chi^2 values, dfs, and p-values because it is a good practice and also because we think that the results for the chi^2 model test or RMSEA confidence intervals are always informative. In the manuscript, we try to make sure the inferential criteria are clear and chi^2 test or RMSEA CIs are not among them. If you think it would be better if we explicate which of the reported indices/statistics are not used for inference, we will gladly add this information.

Drew Altschul: LN 454 – why does the unraveling happen only at the scalar level? Or am I misunderstanding this? It would be good to know more about what the implications of switching to the mixture multigroup analysis are.

Authors' Response: We focus on the scalar level because it is justified to interpret mean differences across groups (countries, in this case) only if scalar invariance holds. This allows for the interpretation of differences in loneliness prevalence. Under scalar non-invariance, the validated measures would be inadequate for any cross-country comparisons because it is unknowable whether the observed differences are due to true differences in the level of the underlying construct or a measurement artifact. In the revised version, we make this reason explicit.

Regarding the implications of switching to the mixture multigroup analysis: in case the given measure turns out to be non-invariant with respect to country, it is important from both the substantive and the pragmatic perspective to know for which countries the scores are comparable. Identification of clusters of countries where scalar invariance holds provides researchers with necessary empirical input that is needed to devise and test substantive theories/hypotheses that address the differential functioning of the loneliness measures. From a practical standpoint, it provides empirical grounds to readily compare loneliness prevalence within clusters of EU countries. To clarify the use of mixture multigroup analysis, we have included the arguments above in the revised manuscript.

Drew Altschul: A final note about the construct validity sections, drawing on the code. In the code, the authors appear to be fitting many CFA constructs for the wider nomological net, which is good. But I'm left wondering why the correlational analysis is so simple. Westfall and Yarkoni

(2016) demonstrated that correlations between latent constructs are easily inflated, but when one accounts for measurement invariance as one does in SEM, we can get much closer to the true correlations. Westfall and Yarkoni also show how this sort of analysis can be done with single indicators as well, when a multi-indicator latent variable cannot be formed. I suggest that authors consider this approach, or at very least they refrain from using Pearson correlations for this set of analyses. Pearson correlations really only ought to be used on continuous variables, and adding up the items of these constructs doesn't really make them continuous, there is still clear ordinality in these sorts of data. I strongly suggest the authors use Spearman or Kendall correlations instead.

Authors' Response: We fully agree with this point. Instead of using simple Pearson's correlations of observed scores, we have revised the analysis plan and now use factor scores. For multiple-item measures, we have fitted a CFA model using WLSMV estimator, explicitly modeling the items as ordinal, and extracted the measurement error-free for the unitary latent factor. For single-item measures, we conservatively assumed ~50% reliability, modeling a latent variable having a single indicator by fixing the factor loading to .70. We had to commit to some factor loading otherwise the measurement models would be under-identified. Instead of bivariate Pearson's correlation between the observed mean scores, the relationships between loneliness measures and nomological network constructs is now estimated by using factor scores, which are a superior approximation to the true relationship between the underlying constructs.

Translation and Cultural Considerations

Mary Louise Pomeroy: Some of her comments that she left in the document we directly revised there.

Mary Louise Pomeroy: Line 75. I had expected to see a brief conversation (around here or elsewhere) about the psychometric issues that may present themselves when applying existing loneliness measurement tools across different countries, languages, and cultures. I never fully found such a discussion in the present manuscript, except for some procedural information about how the tool was back-translated in the methods. Such a discussion seems highly relevant given that the authors' short-term goal is to investigate the rigor of these tools' psychometric properties in a population that spans the EU, with a long-term goal of guiding the selection of measurement tools for EU population-level surveys that seek to estimate and address loneliness. I.e., considerations such as: is loneliness defined differently in different countries? Are questionnaires' wording interpreted the same way by different cultures? Does content validity change when the tool is translated into other languages? etc. I believe Roger O'Sullivan may have done some work in this area. Any literature or background information you can provide on the matter would likely satisfy my desire for this discussion, assuming that the authors are not able to tackle these issues in analyses.

Authors' Response: We completely agree with the reviewer that these issues are important. We had planned to include information related to conceptual issues in the general discussion. After all, we will have a set of measurement invariance tests that are relevant for the issue of the understanding and measurement of loneliness across the EU. We will then also include the work by Roger O'Sullivan. We are, however, open to doing so differently if the reviewer feels strongly about including this information in the theoretical introduction.

Thuy-vy Nguyen. The method of data collection, including translation of the original English version, was explained. However, I would like the author to add in details related to translation process; for example, were texts translated verbatim, or were there steps to also gather cultural input from native-speakers of each country? I understand that data has been collected, so I am only asking for clarification so future research can look for ways to improve, in the case that texts were only translated verbatim.

Authors' Response: We thank the reviewer for pointing this omission out and apologize for not having explained this important point. We have now included the following information:

> The survey was originally drafted in English. Once the English version was finalized, professional translators forward-translated the entire survey into the national language of each member state (with the exception of Ireland and Malta, where only an English version of the survey was used). Thirty-one out of the 82 survey questions of the main questionnaire were back-translated. Back translation was reserved for more complex questions. For the remainder of the questions either existing translations (4 questions) or forward-translation were used. Instructions to translators are provided in the survey on our OSF page: https://osf.io/unfrc/.

Descriptive Statistics and Sampling

Thuy-vy Nguyen. I would like the authors to consider more the descriptive statistics of the scales. There was a discussion around how to deal with 3-point scales and not to treat them as continuous scales. However, what if there is a ceiling effect where most participants in any specific countries are either on the very high or low end of the response scales, would this affect the planned analyses in anyway (I do not know because I am not an expert in these analyses so this is a question in case it needs to be considered). In that case, will data need to be recoded for that country, for example?

Authors' Response: Indeed, given the prevalence of loneliness, the data from the three loneliness measures are expected to be zero-inflated and right-skewed, like what can be seen with the vast majority of psychopathology symptoms. So, although there is no doubt that floor effect usually represents loss of information, in this case, it is a perfectly normal feature of the given data. Apart from the ordered categorical nature of the data, the positive skew of the data is the other reason for choosing the Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator. This method is a standard choice when modeling such data.

That said, we echo your point about the need to clearly lay out the descriptive characteristics of the measurements across all countries and in aggregate, which we do in Table 1. In the discussion, we will note if we find any peculiar pattern in the data.

Conceptual Clarifications and Literature Review

Mary Louise Pomeroy: Line 103. The conceptual difference between loneliness and social isolation should be clarified (and perhaps presented earlier). A history of research articles conflating social isolation and loneliness also contributes to the conceptual barriers in loneliness measurement. For example, studies that purport to measure social isolation but use questions with language such as “feel socially isolated” (indicating loneliness). This problem has improved much over the past five years but did create confusion in recent history. I provided additional thoughts for this paragraph in track changes in the manuscript.

Authors’ Response: First of all, thank you for your detailed feedback! We have incorporated almost all your comments into our manuscript. Then, we have revised the explanation about social isolation versus loneliness. Part of the problem in this survey is also that the item “Feel isolated from others” is included for the T-ILS. We will return to this problem in our general discussion.

Mary Louise Pomeroy: Line 115. It is worth tipping the hat to some of the conceptual work that has been conducted for loneliness by Tom Prohaska, Linda Fried, and colleagues. I have provided some suggested citations in the manuscript’s comment bar.

Authors’ Response: Thanks again. We have included these references and discussed their work.

Mary Louise Pomeroy: Line 178. Please provide a distinct opening paragraph or a few sentences regarding any psychometric properties of loneliness measures that are well-established, or that perform particularly well or poorly, prior to discussing findings that are mixed and gaps in knowledge.

Authors’ Response: We’d love to provide this recommendation, but don’t yet feel very well equipped to do so! We are currently conducting a systematic review of measures, and we feel that that systematic review will provide that information (well-established measures will provide evidence of measurement invariance, good concept-to-measure mapping, test-retest reliability, and so forth. A formal test of this will be best). We propose adding a brief reference to this systematic review in our general discussion.

Mary Louise Pomeroy: Line 314.It would be helpful to the reader to provide a combined table that displays each of the three tested loneliness measures, so that readers can compare their question prompts, specific items, and underlying language. I would find myself wanting to examine those differences when interpreting the study’s results.

Authors’ Response: We agree. We have included Table 1 to provide the loneliness measures for easy access to the reader.