Dear Prof Bishop,

We would like to thank you for the constructive feedback and the quick turnaround on our work. Below we provide a point-by-point answer to each of the points you raised, ordered by topic. To facilitate review, we have included a version with track changes on, as well as a version with track changes accepted.

Theory

1. **Jacek Buczny:** I wonder why you have not considered applications of the Social Relationship Model (SRM) as an important source of instruments. The SRM posits that social perception/evaluation/traits variance can be partitioned into various components: target variance, perceiver variance, relationship variance, and error variance, and such information can be collected by implementing a round-robin design. I can imagine that specific aspects of social connections (e.g., social support, responsiveness, quality) can be attributed to each type of variance and depend on each other. This gap is a bit puzzling as a thorough understanding of social relationships should account for such specific components.

Authors' Response: We would like to thank the reviewer for that suggestion. We agree with him that the Social Relationship Model is relevant and that in our current approach, we are missing sources of variance that can be attributed to other people's opinions about social connection. The type of psychometric model that the Social Relationship Model relies on is different than the one we focus on. Our current focus is on between-person factor models of individual attitudes of social connection. In the medium term, for such between-person models, we think that attitude network models are the most appropriate way to capture social connection.

In the long term, we are planning to include other sources of variance, but we don't think that the SRM is the most fruitful way to go. We think that complex system models are the best approach to capture not only target variance, perceiver variance, relationship variance, and error variance, but also longitudinal variance.

Proposed action: We have included in our inclusion criteria that our focus will be on people's individual attitudes and have added a footnote explaining that we exclude – for now – approaches like the Social Relationship Model:

> Another useful approach to measure social connection is captured by the Social Relationship Model, which captures target variance, perceiver variance, relationship variance, and error variance. We think that the Social Relationship Model is a useful approach to capture these different sources of variance. At present, we restrict ourselves to a between-person factor model of individual attitudes of social connection. In the medium term, our goal is to switch from between-factor models of individual attitudes of social connection to network attitude models of social connection. In the longer term, we are planning to include other sources of variance (i.e., other people and changes over time) through complex system models.

2. **Alexander Wilson:** On p14, the authors provide definitions that they will use to categorize whether measures are assessing structural, functional and qualitative aspects of social connection. The definition of functional aspects reads "A sense of connection that results from actual or perceived support or inclusion." The definition of qualitative aspects reads "The sense of connection to others that is based on positive and negative qualities." I am curious how confident the authors feel in clearly

distinguishing functional and qualitative indicators of social connection on these grounds, as these similar definitions to me.

Authors' Response: We agree that the definitions of function and quality are close and that could present a problem. We will adapt them slightly, however (see further below). We think it would be useful to talk about our general approach to this review. We chose the definitions as working definitions that best reflect the state-of-the-art of the literature, but they need to be further improved and we can accomplish this through our coding.

We do agree that it can be challenging to categorize these measures when considering them on a case-by-case basis. That is why we will examine intercoder reliability at different stages and adjust the coding categories, if necessary. Furthermore, it is an explicit goal of our review to examine the overlap (or lack of overlap) between these definitions (within and across all components; the latter two points we feel are already represented in the previous version of our manuscript).

Proposed action: We will change the definition of the qualitative component of social connection to "The sense of connection to others that is based on positive and negative affective qualities (e.g., relationship satisfaction, intimacy, or conflict)" and the functional component of social connection to "a sense of connection that results from resources and functions provided or perceived to be available by social relationships (e.g. perceived social support, loneliness)". We will then also add a sentence to recognize that the measurement tools and the categories overlap to some extent and that we will evaluate and potentially update the categories of social connection based on our systematic review: "We recognize that these definitions are overlapping to some extent. We expect the categories and their respective definitions to be updated based on our systematic review.".

Study 1 methodology (literature search)

3. **Richard James:** My main concern for Study 1 relates to the justification for the structural indicators searches. I completely understand that parsing through 400,000+ results is not feasible or an effective use of time. However, the use of a random subsample has potential drawbacks. Specifically, I have reservations that the variety of different types of structural indicator would be captured by random sample of a similar number of results as the number of functional and qualitative indicators. Given the information the authors' have presented, my impression is that there will be much greater heterogeneity among structural indicators relative to the functional and qualitative ones. Second, given the issues reported in the Stage 1 submission so far, it seems fair to expect the results to be far noisier. I wondered whether it might be preferable to stratify the sample to capture a subset of the most relevant results and a random sample (sorted by time), but am also conscious this has its own drawbacks. I would appreciate the author's thoughts on this, and some additional justification of the sampling approach in revising the methods section.

Authors' Response: We agree with the reviewer that there is a potential for noisier results with that category. Whether the structural category is less noisy or noisier is an empirical question.

We further realized that our initial plan to select a random number of page results after sorting the search results by date will not be possible to conduct, as search databases display a limited number of search results (e.g., 10,000 results on PubMed), which makes a wide array of results pages unavailable. We thus decided to perform searches with search results sorted by relevance, stratifying for the year of publication to ensure a fair representation of each year in the list of articles eligible for review. We hope this change proposes a great balance between our initial plan and the reviewer's suggestion.

We acknowledge that our search may not be entirely reproducible, but it should still remain replicable. As explained, we are committed to doing our best to maintain the reproducibility of the search and its results. However, we are aware that achieving 100% reproducibility for our search is not feasible; nonetheless, the outcomes should be replicable, even if the exact population of articles that one analyzes is different.

Proposed action: We edited our search strategy for the structural search as follows:

> To ensure the review on structural indicators is feasible to conduct, we extracted a subset of the full search results totaling the average results detected for the functional and qualitative indicators (*N* = XX). We sorted the search results by relevance to reduce noise in the results and extracted search results stratifying for publication year to ensure a fair representation of articles across time. In order to do so, we 1) performed the structural search, 2) retrieved the total number of search results and the number of search results for each publication year, 3) computed the percentage of search results for each publication year, 4) computed the number of search results to retrieve for each publication year. Scripts and spreadsheets necessary to replicate this process are documented on our OSF Page: https://osf.io/stmdb/.

4. **Alexander Wilson:** In review 1a, the authors require papers reporting on functional and qualitative indicators to involve development and/or validation of the measure, whereas for structural indictors, the measure simply needs to be included in the paper (although note that you state this is on lines 277-8 but then state something different on lines 302-3). This seems inconsistent to me, and I wonder if it will bias the kinds of measures you include in the final review. I understand your motivation for this – structural indicators may be just a single question, and so you might not expect development/validation of the measure in a paper. Therefore, I wonder if you should just remove this criterion for all measures in review 1 – i.e., the paper simply needs to measure the construct and not necessarily involve psychometric testing of the measure? In review 2a/b, you could then include this criterion (i.e., that the paper needs to be specifically focused on psychometric), as your aim here will be to the assess measurement qualities.

Authors' Response: We agree with the reviewer that our literature searches should be as consistent as possible across the three indicators of social connection to reduce risks of bias and that the current approach has some disadvantages.

When designing our search strategies, we tried, as well as possible, to weigh the pros and cons of different strategies. While removing the development/validation keywords for the functional/qualitative components would increase homogeneity between literature searches, our trial searches showed that removing these keywords would result in retrieving a large number of (irrelevant) search results. The potential relevant search results we would gain are probably the ones that are created ad hoc (as they do not include the words "validation" or "development"). The current approach to quality and function is therefore likely a conservative judgment of the literature and will make it appear slightly better than it is. If we were to approach it the same way as the structural component, we would again have to draw a subset of results. We may then gain some of ad hoc measures, but we would lose some of the relevant, high-quality measures. Based on our trial searches, we further think that the proportion of irrelevant to relevant results tips more towards irrelevant results (as not many results produce the same measure), whereas this is slightly less so for structure (as many results should produce the same measure).

It thus appears that a tradeoff between homogeneity in methods and quality of data has to be made in the present situation. We think – on balance – our work will be less negatively impacted by a decrease in homogeneity in methodology rather than a decrease in data quality.

Proposed action: We have explained why we prefer the slightly different search terms in Footnote 2:

> We are conscious that the differences in search strategies may produce slightly different search results. Of course, the nature of the measures is usually different for structure on the one hand (often measured by single items) and function and quality on the other hand (often measured through composite indexes and/or Likert scales). One solution could be to remove "development" and "validation" for function and quality. We decided against it after several trial searches. For the structural component, this produced many redundant search results. This was not the case for the functional and qualitative components, which likely means that on balance, we would lose more relevant results for the functional and qualitative components if we were to select a subset of results. By approaching the literature in this way for the functional and qualitative components, we probably lose out on some ad hoc measures, providing a relatively conservative – and thus optimistic – judgment of the measurement quality of the social connection literature.

5. **Alexander Wilson:** Across lines 258-263 (p15), the authors state some search steps they will carry out to ensure the search results are reproducible for the structural indicators. I did not understand how these steps will ensure reproducibility.

Authors' Response: Sorting the search results by date with the oldest articles displayed first ensures that a given page of search results will contain the same articles across searches of the same search string. That is because the publication date of an article does not change over time. As such, searches performed across time should yield the same search results on the same pages. However, as explained in our answer to comment 3, this won't be feasible.

Proposed action: No changes beyond the one proposed to comment 3.

Study 1 methodology (item coding)

6. **Richard James**: The justification for 60% agreement on the item content analysis raises questions. Again, understand given the potential range and heterogeneity of measures how this would be difficult. I think some additional justification of this criterion would be useful with reference to specific studies where this has been a problem.
7. **Alexander Wilson**: As you state, 60% is lower than any conventional level – and as someone reading the paper, this does seem low. I understand you suggest this based on your past experience of difficulties reaching higher agreement – but I wonder if this shows the methodology is suboptimal and it might be best to consider another approach rather than pursue something that may not be reproducible? I accept that it may be defensible if you are simply looking to summarize the measures – however, you plan to use these categories in your analysis to establish the level of overlap between different measures. However, if you cannot reliably establish the construct measured by a particular item, how can you have confidence in establishing the level

of overlap between items, if different people might apply different categories to the individual items?

Authors' response: We thank the reviewers for these comments. We have thought carefully about the reviewers' concerns and have made two changes:

1. We will only allow classification into a single category.
2. We have raised the percentage agreement to 80.

We agree that the lowering of the agreement to 60% poses a problem to the reliability of the task. At the same time, allowing the item to be classified into multiple categories introduces more variability into the process than we would be able to accommodate for if we want a reliable coding process. The change towards only allowing the classification into a single category may further refine the item-categories, but this may produce an even weaker overlap (as we may need more categories to explain the items). We will recognize this by adding our expectations into the text.

Proposed action: We will raise our threshold to a conventional level of .80 Cohen's κ and we will only allow coding into a single category. We will explain it in the following way in the text:

> To ensure that the categories were accurate representations of the items, each of three other authors (XX, XX, and XX) independently cross-checked a separate 10% of the coding list, after which we calculated interrater agreement using Cohen's κ (Landis & Koch, 1977). We repeated this process until we cross-checked at least 60% of the coding list and until the average agreement was at least .80
> [...]
>
> We deviated in our coding strategy from earlier work, where coding into multiple categories was permitted. Even though this coding strategy may make some sense conceptually, in our experience, it is very difficult to have a precise measure of overlap, as it is very challenging to get interrater agreement higher than 60%. In case the item could be classified into multiple categories, we either refined the category, or classified the item into a category that seemed to provide the best fit. We suspect this adjustment may somewhat inflate the final number of categories, potentially leading to an even weaker overlap between measures.

8. **Alexander Wilson**: You propose using Jaccard indexes to quantify level of overlap, but I wonder if these indexes will give an impression of precision whereas what we might really be looking at is fuzzy and subjective (for reasons highlighted above).

Authors' response: We agree that a numerical index can give an impression of precision on what we intend to evaluate. We also agree that the precision we gave previously was rather low, which we have now changed. We hope that this change mitigates the reviewer's concern.

Proposed action: No changes beyond the one proposed in comments 6-7.

9. **Alexander Wilson**: You also state that raters can apply more than one category to an item. I can see this might make sense conceptually, but I wonder if it will artificially inflate Jaccard indexes.

Authors' response: As is probably evident in the response to the reviewers' previous concern (comments 6-7), we agree that allowing coding into multiple categories may inflate the Jaccard index. At the same time, not allowing for this possibility may reduce theoretical coherence (because from a theoretical perspective, and as the reviewer indicates as well, items could, in fact, be categorized into multiple categories). We agree with the reviewer, as is also clarified in the earlier response to comments 6-7.

Proposed action: We changed the text in our earlier response to comments 6-7, largely in agreement with the reviewer's position.

10. **Alexander Wilson**: I have a related concern around a lack of blinding of raters to your hypothesis. A rater's expectation of the level of heterogeneity (or not) across these measures might influence the number of categories included in the codebook and the application of these to the measures. Similarly, you say raters can apply more than one code to an item. If I am expecting high heterogeneity, I might apply multiple codes to the same item.

Authors' response: It is true that the subjective nature of the task will allow bias, particularly when the hypothesis is known to the raters. We hope that the current change – not allowing coding into multiple categories – prevents that problem.

Proposed action: No changes beyond the one proposed to comments 6-7.

11. **Alexander Wilson**: I wonder if a more empirical approach to this process might be helpful? Could you, for instance, include participants (blinded to your hypotheses) in this process to sort and rate items? This would allow you to look at how individuals cluster items of different measures, and how reliable this process is. Just an idea, and you might not have resources for this.

Authors' response: We clearly see the value of the reviewer's proposition and believe that studies conducting similar research would benefit from employing such methodology. Unfortunately, the very high number of items we expect to categorize from more than 30,000 articles would make that process very costly and complex to set up.

However, thinking carefully about the reviewer's point, we do agree, and, in order to further limit biases in the ratings and strengthen the cross-checking process, we decided to leverage artificial intelligence to categorize the items. More specifically, we programmed additional python scripts that will prompt gpt-3.5 (OpenAI, n.d.) with the categories we created and the items to categorize. We back-tested the performance of gpt-3.5 on the data from CORE Lab (2023) and obtained promising results (Cohen's Kappa = .80 between gpt-3.5 and the main coder of the study, across 518 items), suggesting that this artificial intelligence is suitable for this task.

Proposed action: We will leverage gpt-3.5 to partially account for the risk of bias in our coding and strengthen the cross-checking process. Changes to the manuscript will be reflected in the following paragraph:

> As we expected heterogeneity among measures, we could not rule out the possibility that our ratings have been biased toward finding greater heterogeneity. To address that concern to some extent and to further strengthen the cross-checking process, we leveraged artificial intelligence to categorize the items in the categories we created. As the artificial intelligence would be blind to our hypotheses, a great interrater agreement with the artificial intelligence would suggest that our categorization judgments are both bias-free and accurate. Using Python scripts, we prompted gpt-3.5 (OpenAI, n.d.) with the categories we created and the items to categorize (see Appendix X for a template of the prompts we sent). To the best of our knowledge, no

research formally assessed the performance of gpt-3.5 on this specific task. As such, results presented below should be taken with caution, though a test of gpt-3.5 performance using data from CORE Lab (2023) suggests that it is well suited to perform this task (κ = .80, see https://osf.io/stmdb/ for scripts and data). After the final cross-check, average agreement with gpt-3.5 were [not satisfactory/satisfactory], reaching XX. These results suggest that our coding process is [biased/unbiased] and [reliable/unreliable]. A table displaying the agreement obtained with each coder can be found in Table 2.

Study 2 methodology (literature search and study coding)

12. **Richard James**: Study 2a: I agree with the overall approach for the systematic review, and the searches are specifically defined to identify appropriate studies. The only concern I had was that the search strategy relies on the studies clearly flagging this, which in my own experience of gathering data to examine a scale or scope the literature isn't guaranteed. I would like the authors to consider whether there would be value including select forward citation searches of key papers relating to the scales identified (i.e. initial validation papers), to ensure any relevant studies aren't missed. Otherwise, I agree with not conducting further reviews in the structural indicators domain given the use of single item scales. If validated measures do come up though from that search, and the use of forward citation search may be a reasonable adjustment to ensure these studies are properly captured.

Authors' response: We agree that combining the literature search with additional search methods like forward citation searches would certainly help in identifying a number of additional relevant studies. At present – and perhaps due to our misunderstanding of what the reviewer has stated – we are unsure what extra search results we would retrieve by forward citation searches. As we conduct a first search identifying measures, then identify additional results through COSMIN keywords, we anticipate having a large number of search results that identify the necessary validity evidence. For now, we don't adjust the strategy, but if the reviewer can clarify what kind of search results we may miss with our strategy, we'd happily consider it.

Proposed action: No changes.

13. **Richard James**: Study 2a: Reading through this, I wondered whether it would be worth specifically recording whether a non-standard or modified use of a measurement was applied as a variable in the template for extracting sample characteristics. I really liked the use of the COSMIN taxonomy to systematize the quality of the measurements, and think it is a particular strength of evaluating the measurement properties of the scales to be examined. However, I'm also conscious COSMIN doesn't capture some questionable measurement practices that are important in qualifying the use of many measurements, especially where the inconsistent use of measures is a key problem. From my own experience of scoping across a large literature, I find that I quickly begin encountering studies where existing scales have been modified (i.e. different response scales, subsets of questions), and that some scales are more susceptible to it than others (e.g. length of questionnaire, use of many or very few response options). When thinking about the Stage 2 discussion, this might also reinforce some of the evaluation of these measures.

Authors' response: We would like to thank the reviewer for that suggestion. The point he raised (measures being altered across studies) echoes with our own research experience in the field of social connection. We will code that information as we believe it will add relevant value to our research while adding minimal additional work.

Proposed action: In the results section of study 2b, we added that we extracted whether the measure was modified along with the other sample variables. We also added a new column to Table 6, named "Modified measure" that will contain binary values (yes/no) as to whether the measure employed in the study was modified or not.

14. **Jacek Buczny**: Maybe I overlooked it while reading, but it is unclear whether you will evaluate the quality of the theory used to create a specific measurement. I can imagine a scenario in which a good measure is created (reliable, valid, invariant across genders and cultures), but the background theory is rather weak.

Authors' response: We agree with the reviewer that the theory behind a measurement is a critical layer to consider when judging the overall quality of the measurement. When evaluating the content validity of measures using the COSMIN methodology, one grading section is devoted to the development of the measure (see page 17 of https://www.cosmin.nl/wp-content/uploads/COSMIN-methodology-for-content-validity-user-manual-v1.pdf). Though not meant to provide an exhaustive assessment of the theory that originated the measure, it still provides some useful information as to whether the scale "comes out of nowhere" or not.

Proposed action: No changes, but we will develop that point in the discussion.

Writing / Clarification

15. **Alexander Wilson**: I felt the authors needed to define their understanding and approach to "social connection" earlier. For the first page and a half, "social connection" and "loneliness" are discussed loosely, and then aims of the review are stipulated – however, this is before "social connection" is defined. I think this definition needs to come sooner, especially as the authors are focusing on a particular conceptualization of social connection (that of Holt-Lunstad). I would recommend that the authors define this model as early as possible, and highlight its strengths and limitations.

Authors response: We thank the reviewer for this suggestion. In a different manuscript, we had a similar suggestion, where a reviewer asked to define our concepts right at the beginning. We think this is a difference in style – some authors prefer to broadly set the stage and the goals in the early paragraphs, whereas others prefer to define their core concepts. We prefer to go for the former and hope that the reviewer is ok if we keep our structure.

Proposed action: We prefer to keep it as is, unless both the editor, this reviewer, and the other reviewers prefer a change. In that case, we could adopt the following:

> How people form and maintain healthy relationships is one of the most crucial questions in modern social science. Those relationships and the connection to other people, referred to as social connection, are therefore of particular importance. Social connection is a multi-dimensional concept that encompasses structural, functional, and qualitative indicators, explaining how people connect to one another, their feelings of loneliness, and their degree of social isolation. A lack of social connection impacts health and longevity similarly to other clinical risk factors (Holt-Lunstad et al., 2010; Pantell et al., 2013), and other research suggests that living alone (versus

living with others) is associated with a 32 % increased risk of early death (Holt-Lunstad et al., 2015).]

16. **Alexander Wilson**: I would also suggest that the authors consider any previous reviews drawing on this model. On p13, they mention some reviews in passing. I would suggest referencing any relevant reviews in the introduction, and highlighting what this review adds to those existing reviews.

Authors' response: We thank the reviewer for this suggestion. The papers we cite there are a bit of a mixed bag. The Bugallo-Carrera et al., 2023, CORE Lab, 2023, and Maes et al., 2022 are all relevant. The Holt-Lunstad et al. and Valtorta et al., 2016 relate social connection to other outcomes (and are thus not relevant for the introduction). The CORE Lab paper we already referenced in the introduction.

Proposed action: We will add one sentence about the Maes et al. (2022) and the Bugallo-Carrera et al. (2023) papers in the section on "Challenges to the applicability of measures: Heterogeneity, internal structure validity, and intended target groups":

> A high degree of heterogeneity and, thus, a lack of conceptual clarity is at the heart of many measurement problems (Flake & Fried, 2020; Fried & Flake, 2018). A lack of conceptual clarity is often revealed through a plethora of different scales trying to capture the same concept. For instance, the CORE Lab's (2023) systematic review detected 26 different instruments in total that all measure romantic relationship quality. This variety of measurement instruments corresponds to the literature usually reporting exploratory evidence, providing only fuzzy conceptual definitions or often no definitions at all (see also Delatorre & Wagner, 2020). From the 26 different instruments, the CORE Lab (2023) coded 25 different item categories that were only weakly overlapping, ranging in content from affection and love, to agreement about proper conduct, conflict resolution, sexuality, family life/parenting, to compatibility in attitudes/preferences (see also Fried, 2017a, b). More specifically pertaining to the domain of loneliness, two systematic reviews are already suggestive of its heterogeneity of social connection measurements, finding 18 different loneliness measurement instruments (Bugallo-Carrera et al., 2023; Maes et al., 2022).

17. **Richard James**: RQ 3 and RQ4: I didn't think these really came out in the Stage 1 Report as being key aims of the research. I thought the area where this was most clearly referred to was in the abstract. Reading through the report without reference to the RQ table, my impression would be that the results are be reporting the country/population of study i.e. to represent coverage in the literature, rather than the application of these measures to other contexts is meaningful (i.e through use of measurement invariance). The paragraph from lines 190-203 makes the case strongly for the importance of testing these questions, but I thought the end of the paragraph from line 204-212 ought to make it clearer that the aim of this exercise is to assess whether these generalizations are defendable.

Authors' response: We thank the reviewer for pointing this out.

Proposed action: We will rewrite the last introductory paragraph in the following way:

> Take, for instance, a widely adopted theory on social connection, attachment theory. Traditionally, attachment theory focuses on parents, who are typically the primary caretakers of infants in higher-income countries. Among traditional families in

> Madagascar, however, infants interact almost exclusively with peer groups of older children, who end up being infant caretakers (Keller, 2018; Scheidecker, 2017). Attachment theory may have only very limited applicability in certain regions, and this also applies to the measures developed to test it. More generally, strong homogeneity in world regions that developed social connection measurements and in the population groups with which they were developed may be indicative of a miscalibration of those measurements to assess social connection across various human populations. A third focus of this systematic review is thus to identify where measures originated, and, relatedly, a fourth focus is with which kinds of population groups in mind (i.e., age, race, gender, sexuality, religious affiliation, and socioeconomic indicators) the measures were developed.

18. **Alexander Wilson**: On p18, what do you mean by unidimensional and multidimensional? Do you mean measures devised with subscales, or where measures have been empirically shown to include multiple factors? These are potentially quite different things.

Authors' response: We referred to measures devised with subscales, as we think that information is more relevant to provide an informal description of the measures' characteristics.
Proposed action: We added the following footnote to clarify that point (Footnote 10):

> For the number of dimensions, we refer to the subscales defined by the authors of the measurement, not to the empirical factors they may have obtained through confirmatory factor analysis.

19. **Alexander Wilson**: On p18, you state the different properties that you will report on in one order in the text, but then use a different order in your tables on that page. Could these be kept consistent?

Authors' response: We thank the reviewer for catching this inconsistency.

Proposed action: We reordered the variables to make them consistent across the text and the table.

20. **Alexander Wilson**: My understanding is that in the first stages, you will produce a codebook of categories to capture the content of items based on initial review of the measures, and then apply this codebook when returning to the measures. You mention involvement of four coders and calculating inter-rater agreement, however it is not clear whether this is happening in generation of the codebook or after. I would hope this is at both stages.

Authors' response: The reviewer is correct that we calculate inter-rater agreement before and after generating the codebook, until we reach 80%.

Proposed action: We have clarified this in the text:

> If disagreements arose, we either resolved the disagreement, changed the category description, or added new categories. This cross-checking process ensured that both the creation of the codebook and the categorization of items had collective agreement.

21. **Alexander Wilson**: You say that you will aim for 60% agreement. Is this absolute agreement, or a kappa's value?

Authors' response: This refers to the kappa's value.

Proposed action: We have further clarified this in the text:

> To ensure that the categories were accurate representations of the items, each of three other authors (XX, XX, and XX) independently cross-checked a separate 10% of the coding list, after which we calculated interrater agreement using Cohen's κ (Landis & Koch, 1977).

22. **Alexander Wilson**: Over pp29-30, there is some text in yellow in brackets (which I assume you plan to remove?), where you mention using the COSMIN manual to assess the quality of psychometric analyses. However, I think this needs to be in the main method section, and described in more detail. Who is coding this information from papers, and how will you ensure quality ratings given to papers are reproducible?
23. **Jacek Buczny:** The use of COSMIN is a very good idea; however, to me, it is not clear how the tool will be used. Of course, conducting an evaluation based on COSMIN is not a difficult task, but for reproducibility, more details on how you want to use would be welcome.

Authors' response and proposed action: We agree that we are too succinct in describing the use of COSMIN. To address the reviewers' concerns without adding in the manuscript some information that would be too redundant with the COSMIN guidelines, we have added a method section to study 2b that summarizes the key steps of the COSMIN procedure:

> **Evaluation of the measurement properties**
>
> We followed the COSMIN guidelines to evaluate the measurement properties of each measure. Below we provide a summary of the different steps completed for each measure for replication purposes. We refer the reader to the comprehensive COSMIN manual (Mokkink et al., 2017; Prinsen et al., 2018; Terwee et al., 2018) for more exhaustive information on how to complete such an evaluation, as the process described below slightly differs across measurement properties.
>
> For each eligible study, we first rated the methodological quality (as *very good*, *adequate*, *doubtful*, or *inadequate*) of that study using the COSMIN Risk of Bias checklist (Mokkink et al., 2017). We then rated the measurement property (as *sufficient*, *insufficient*, or *indeterminate*) reported in the study against the updated criteria for good measurement properties (Prinsen et al., 2018). [Here we will provide a coding example]. This coding process was carried out by [author] and cross-checked by [author]. We resolved conflicts by soliciting the help of an additional coder ([author]).
>
> Once all eligible studies were coded for a given measure, we first summarized the results on each measurement property, either through a) quantitatively pooling the results using meta-analysis, or b) qualitatively summarizing the results, in case quantitatively pooling was not possible. This process allowed us to provide an overall rating (*sufficient*, *insufficient*, or *inconsistent*) for each measurement property. Wethen accompanied this overall rating with a grading for the quality of evidence (*high*,

moderate, *low*, or *very low*), using the modified GRADE approached outlined in the COSMIN guidelines. [Here we will provide a coding example]. This coding process was carried out by [author] and cross-checked by [author]. We resolved conflicts by soliciting the help of an additional coder ([author]).

Finally, we described the interpretability and feasibility aspects of each measure (Mokkink et al., 2017; Prinsen et al., 2018; Terwee et al., 2018). Interpretability refers to the degree to which one can assign qualitative meaning to the score obtained to a measure. The interpretability of a measure is notably described by the distribution of scores obtained within a study sample. A measure would typically show great (poor) interpretability if its range of possible scores is wide (narrow). On the other hand, feasibility refers to the ease of application of a measure, in its intended context of use and given various constraints like time or money. The feasibility of a measure is notably described by its number of items. A short measure would typically show greater feasibility than a long measure in most contexts. We extracted the different interpretability and feasibility aspects of each measure using the templates designed by the COSMIN group.

24. **Alexander Wilson**: Can you define feasibility and interpretability on pp29-30 please? How will assessment of these properties happen?

Authors' response: These refer to feasibility and interpretability as defined by the COSMIN group (defined on page 44 of this document: https://cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf) and they will be assessed following the COSMIN guidelines, as for the other psychometric properties.

Proposed action: We have provided a definition of feasibility and interpretability in the method section proposed in response to comments 22-23.

"Interpretability refers to the degree to which one can assign qualitative meaning to the score obtained to a measure. The interpretability of a measure is notably described by the distribution of scores obtained within a study sample. A measure would typically show great (poor) interpretability if its range of possible scores is wide (narrow). On the other hand, feasibility refers to the ease of application of a measure, in its intended context of use and given various constraints like time or money. The feasibility of a measure is notably described by its number of items."

25. **Jacek Buczny**: You mention PRISMA for the first time in the result section of Study 2b. Why at this stage? I wonder why because PRISMA is a general framework used for conducting systematic reviews. I would expect that PRISMA is mentioned in the overview of Study 2a. Instead, you want to use COSMIN guidelines only. What was the particular reason not to use PRISMA to design the study?

Authors' response and proposed action: We agree that it would make more sense to refer to it earlier in the manuscript. We thus decided to move the discussion to Study 2a:

The present systematic review aimed to be reproducible and fully transparent. In order to accomplish this goal, we reported the review following the PRISMA statement (Moher et al., 2015). Analysis scripts and the data for this study are publicly available on our project page on the OSF (https://osf.io/wfers/).

26. **Richard James**: I would recommend re-working Section 2 a bit, ideally with a specific sub-heading in the methods for 2b highlighting this is a specific set of analyses, and how these will be reported in the results, with reference to the proposed findings on whether the measures have been validated in countries/populations where it has been applied.

Authors' response: We agreed with the reviewer's comment and decided to rework the naming and structure of Study 2 to make it clearer to the reader.

Proposed action: We adopted the following structure for Study 2:

Study 2a - Systematic review to evaluate the measurement properties of social connection measures

Methods

Literature Search

Results

Results of the articles selection

Study 2b - Measurement properties, country and population of origin of social connection measures

Methods

Evaluation of the measurement properties

Evaluation of the country and population of origin

Results

Results on the measurement properties

Results on the country and population of origin

27. **Alexander Wilson**: I don't understand the distinction between study 2a and 2b. Are these the same?

Authors' response: In Study 2a, we perform the literature search. In Study 2b, we perform the COSMIN coding. We adopted the same structure in Studies 1a and b. We would be open to changing the structures for Studies 1a, b, and Studies 2a, b, and collapse them into Studies 1 and 2, but as we think it is a style preference, we think the effort is not proportional to the gain (if any). Furthermore, we hope the changes we made to the structure of Study 2 in response to comment 26 helps address the reviewer's clarity concerns.

Proposed action: No changes beyond the one proposed to comment 26.

28. **Jacek Buczny**: On p. 25, you write that the analysis code can be found at https://osf.io/wfers – unfortunately, I could not find it. A similar comment applies to this: https://osf.io/n7z4y.

Authors' response: Thank you for pointing this out – this omission was intentional, but we should have said so. Writing scripts beforehand for this part is on the one hand rather straightforward (thus leaving little to no room for flexibility) and on the other hand tricky, as some statistical pooling may or may not happen based on what we find. The selection list of articles for study 2 will be posted during project completion.

Proposed action: We briefly point out that we have not written the scripts or posted the selection list yet in the following footnotes:

- We did not write the analysis scripts yet as their content will depend on the data we retrieve from the literature searches.

- We will upload the selection list of articles once we conducted the literature searches.

External Feedback
Finally, beyond the feedback from the reviewers, we had also sought feedback from other researchers. We will address these in the following notes:

Abstract / Introduction

1. **Mary Louise Pomeroy**: I'd hypothesize that these measures are not well validated.
   [Overview table - Hypothesis RQ2]

Authors' response: We agree that this hypothesis would make sense, especially in light of the results obtained by CORE Lab (2023) on romantic relationship quality, a subcomponent of social connection.

Proposed action: We formulated the following hypothesis in the overview table for research question 2:

> We predict that measures of social connection will show insufficient evidence of great measurement properties.

In addition, we formulated that hypothesis in the introduction section:

> This evaluation is the second focus of this systematic review and we expected social connection measures to mostly show insufficient evidence of great measurement properties, largely in line with the results obtained by CORE Lab (2023).

2. **Maximiliane Uhlich**: I would assume the most represented country in the literature are the US.
   [Overview table - RQ3]

Authors' response: We also agree that this hypothesis would make sense. There is ample evidence in the literature that most research in psychology is conducted in higher-income countries, and especially in the US.

Proposed action: We formulated the following hypothesis in the overview table for research question 3:

> We predict that measures of social connection mostly originate from the US.

In addition, we formulated that hypothesis in the introduction section:

> A third focus of this systematic review is thus to identify where measures originated, and, relatedly, a fourth focus is with which kinds of population groups in mind (i.e., age, race, gender, sexuality, religious affiliation, and socioeconomic indicators) the measures were developed. In line with previous research, we expected social connection measures to mostly originate from the US (CORE Lab, 2023; Delatorre and Wagner, 2020; Henrich et al., 2010).

3. **Maximiliane Uhlich**: I would assume realistically, most measures were developed using a student sample.
   [Overview table - RQ4]

Authors' response: Though an important body of research in psychological science is conducted on university students, we have more reservations in formulating hypotheses on that point, as a significant number of social connection measures have been developed for use on populations of old adults.

Proposed action: No further action.

4. **Mary Louise Pomeroy**: Another component could be if contact is in person or online/device-mediated
   [This comment referred to the term social connection and its definition]

Authors' response: This is a great point. Currently, our search strategy does not target many forms of distance relationships, which may have great prevalence nowadays. To improve on that matter, we added keywords for the structural search.

Proposed action: We decided to add the following keywords in our structural search:

social media interaction
social media engagement
social media participation
social media use
internet relation*
distance relation*
digital relation*
internet access
social media access
phone access

5. **Mary Louise Pomeroy**: Given that social connection is an umbrella term that encompasses various facets of social relationships, it would be helpful to provide a few examples of the various aspects up front (e.g., social isolation, loneliness, social support, social engagement, etc.), if word count permits.
   [This comment referred to the abstract]

Authors' response: Thank you for the suggestion, we will take it into account.

Proposed action: We edited the abstract accordingly:

> "To date, a plethora of instruments exists to measure social connection, assessing a variety of aspects of social connection like loneliness, social isolation, or social support."

6. **Mary Louise Pomeroy**: Specify the time frame for included articles. The conceptual model that collapses aspects of social connection into structural, functional, and quality domains has been adopted in more recent years; terms that capture similar concepts in older literature could be missed with this search algorithm.
   [This comment referred to the abstract]

Authors' response: We have incorporated the suggestion into the updated abstract. Additionally, we expect our keywords to encompass both old and new literature. Our revised search strategy, as addressed to the feedback of the PCI-RR reviewers' team, aims to retrieve all relevant literature, ensuring a comprehensive spectrum through the chosen search terms / keywords.

Proposed action: We have specified the time frame in the abstract:

> in Study 1a, we conducted a systematic review to create a database of social connection measures (N=xx) for its structure (N=xx), function (N=xx), and quality components (N=xx), spanning [YEAR] to [YEAR]

7. **Mary Louise Pomeroy**: Would recommend adding "social isolation" "loneliness" and perhaps "social support" to the keywords for indexing purposes

Authors' response: We see the value of adding these keywords. Thank you for the suggestion.

Proposed action: We have added the keywords in the edited manuscript. The new keywords are: measurement, social connection, social isolation, loneliness, social support, systematic review, quality assessment

8. **Mary Louise Pomeroy**: But would we expect these to be highly correlated? Perhaps, but they are different constructs, so I wouldn't be too alarmed if they were not. [Comment was regarding line 142 and further in the old version]

Authors' response: That is a noteworthy point and we will refer to it in the discussion. We think we can keep this example in the introduction and come back to this aspect in the discussion depending on the results of this review.

Proposed action: No further action.

9. **Mary Louise Pomeroy**: I would be curious about the level of heterogeneity across measures developed for different populations. Did we decide to measure such constructs similarly? How much does the population at hand affect the way we measure certain constructs? It's great you are looking into this, and much needed. You might consider citing this: Taylor, H. O., Cudjoe, T. K. M., Bu, F., & Lim, M. H. (2023). The state of loneliness and social isolation research: Current knowledge and future directions. BMC Public Health, 23, 1049. https://doi.org/10.1186/s12889-023-15967-3

Authors' response: Thank you for providing this relevant article. We will cite it in our manuscript.

Proposed action: We added the following paragraph at the end of the section '*Challenges to the applicability of measures: Heterogeneity, internal structure validity, and intended target groups*':

> Relatedly, Taylor et al. (2023) recently pointed out that research on loneliness and social isolation—two subcomponents of social connection—has been conducted unevenly across world regions and among different populations (e.g., ethnic groups, immigrant communities, or cultural groups). It is probable that similar problems are present in the broader field of social connection, which extends beyond psychology.

Study 1 Methodology

10. **Maximiliane Uhlich**: In the social support literature this is often even further differentiated by the subjective expectation to actually receive social support if

needed
[Comment was regarding the functional indicators of social connection]

Authors' response: We have changed the definitions and addressed these changes in our answer to Comment 2 of the PCI-RR reviewers' team.

Proposed action: No changes beyond the one proposed to comment 2 of the PCI-RR reviewers' team.

11. **Maximiliane Uhlich**: Does this also exclude for instance individuals with chronic illnesses? I´m asking because it´s very common among older people and likely affects their ability to socially connect
[Comment referred to Selection criteria]

Authors' response: Thank you for pointing this out. We will exclude populations targeted because of specific illnesses. However, samples of older adults may include individuals with chronic illnesses. As long as the study authors did not intend to collect data specifically on participants with illnesses of any kind, these samples are acceptable for the study selection. We considered changing the criteria in 'The populations on which they were tested were aimed to be non-clinical populations' but decided to omit it to avoid potential confusion about the phrase 'were aimed to be'. Nonetheless, this is still a valid point, and we plan to address this aspect in the discussion.

Proposed action: We will address this point in the discussion.

12. **Mary Louise Pomeroy**: I'd recommend more specificity here to make it clear that you are looking at consistency from measure-to-measure both within each of the three domains (hopefully high consistency) as well as across the three domains (hopefully low consistency/high discriminant validity). In general, this point should be made clearer early in the manuscript (I wasn't sure that this was what you were doing until the methods section).

Authors' response: Thank you for this recommendation. We will rewrite the first research question and make this point clearer right at the beginning.

Proposed action: We changed the formulation of the first research question in the overview table, research question 1:

> To what extent do the items of the measures of social connection overlap within and between the different aspects of social connection (i.e., to what extent do the different measures assess the same or distinct constructs)?

Furthermore, we reformulated the study description in the first introductory paragraph:

> in Study 1b, we assessed the heterogeneity of these measures through an analysis of item-content overlap of the measurement instruments within and between different aspects of social connection.

13. **Mary Louise Pomeroy**: Consider citing work by Elena Portacolone on living alone, particularly if you are interested in cognitive decline as an outcome of limited social networks.

Authors' response: We agree that this is relevant work. However, we think that the examples in our introduction are sufficient and that additional examples are not necessary.

Proposed action: No further action.

14. **Mary Louise Pomeroy**: First paragraph started by focusing on social connection, but at this point it feels like the focus is more on loneliness. I would aim to rephrase as efforts to augment social connection overall, including various aspects (e.g., loneliness).
[Comment was regarding line 80 and further in the old version]

Authors' response: We agree that the bridge between those paragraphs was missing and that some rewriting is warranted.

Proposed action: We omitted two sentences (starting from line 80 in the previous version) while maintaining the subsequent sentences unchanged to establish coherence between the preceding and subsequent paragraphs:

> Despite the increasing efforts, such as a worldwide initiative to combat loneliness (the Global Initiative on Loneliness and Connection), the UK government dedicating a minister to social contact (UK Government, 2018), and near-immediate access to others through social media, trends remain concerning. For example, estimates suggest loneliness has either remained stable (Hawkley et al., 2019) or increased slightly (Buecker et al., 2021).
>
> Perhaps one of the reasons that loneliness has not decreased may be that the monitoring of it is poorly understood. Loneliness, and its overarching concept of social connection is, after all, multi-dimensional with no clear consensus on a single definition of the concept.

15. **Mary Louise Pomeroy**: Do you only include loneliness under function? I wonder if it its squarely under function or whether aspects of loneliness might also fit under quality (e.g., emotional loneliness, etc.). I am not familiar enough with the loneliness literature, but just a thought.

Authors' response: This is a really good point. Yes, loneliness has different facets (e.g., emotional vs. social loneliness), and maybe some loneliness measures would be more appropriate for the qualitative aspect of social connection. However, we will work with the state-of-the-art definitions, and if they are not appropriate, we will identify this during the coding and as a result and, therefore, change the definitions according to our review.

Proposed action: No further action for now. We expect that this will be addressed in the coding process.

16. **Mary Louise Pomeroy**: Will you include a reference review to catch any missed articles?

Authors' response: Comment 12 of the PCI-RR reviewers' team proposed a similar suggestion. We will already review some of the references of the existing articles highlighted in our previous text.

Proposed action: No further action.

Study 2 Methodology

17. **Maximiliane Uhlich**: If social connection measure originate from different cultural backgrounds, heterogeneity regarding measurement might also be a result cultural differences. Are you thinking of accounting for what social connection means across cultural contexts?

Authors' response: Thank you for highlighting this aspect. We will address this point in the article's discussion. Due to the scope of this study, there is not enough space to adequately address this important and interesting aspect theoretically and empirically. However, regarding the results we obtain, we can theorize why specific measures from different countries differ and mention that this is an interesting research question for future studies. Therefore, we will come back to it in the discussion.

Proposed action: No further action but we will address this aspect in the discussion.

18. **Maximiliane Uhlich**: I´m curious how much the variables overlap across studies. It could be interesting to have table listing all variables/outcomes across studies

Authors' response: We agree that this would be really interesting. We think that a meta-analysis would be the most appropriate way to answer this question. Unfortunately, this is beyond the scope of our current manuscript.

Proposed action: No further action.

19. **Maximiliane Uhlich**: Maybe also include characteristics like minority populations such sexual/gender minorities or refugees. If studies specify that maybe countryside/remote areas vs urban areas/cities could be important regarding geographical barriers. Furthermore, education level, social media use and personality (especially extraversion-introversion) could be important.

Authors' response: Thank you for the suggestions. These characteristics will be extracted in Table 9. We will include a new column for sample types, such as refugees (as suggested) or university students (for instance). Sexual/gender minorities won't be added separately as these aspects are already covered in the columns for sexuality/gender. Additionally, education levels are captured within the socioeconomic indicators and won't require a separate column. As for social media use, personality, and geographical areas, we don't assume distinct conceptualizations of social connection depending on those aspects and thus won't be including separate columns for these factors in the table.

Proposed action: We will include a new column for sample types in Table 9 *Summary of Social Connection Measures' Sample Characteristics.*

20. **Maximiliane Uhlich**: Maybe also some studies examined life transitions (e.g., transition to parenthood, retirement etc.) rather than populations

Authors' response: If we understand it correctly, the author meant populations that had experienced a certain life event? In this case the sample type could be coded as an event sample, and this would then be covered in the added column 'sample type' in Table 9.

Proposed action: No changes beyond the one proposed to comment 16 of the additional feedback.

21. **Maximiliane Uhlich**: Examining re-test reliability could also be important if studies tested their measures longitudinally

Authors' response: Thank you for the suggestion. This aspect is accounted for in COSMIN.

Proposed action: No further action.

We thank you, the PCI-RR reviewers, and our own reviewers for your/their extensive feedback. We think they made our manuscript a lot stronger.

Yours sincerely and on behalf of all the co-authors,

Bastien Paris and Hans IJzerman