**Peer review of “A systematic review of social connection inventories”**

I found this paper interesting to read and reflect on. It appears systematic, and could make a worthwhile contribution to the field. However, there were several issues that occurred to me as I read, which I outline below.

**Introduction**

1. 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 conceptualisation 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.
2. 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.

**Methods**

1. On p14, the authors provide definitions that they will use to categorise 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.
2. 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.
3. 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 assess measurement qualities.
4. 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.
5. 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?
6. In review 1b, it sounds like you are planning to assess though mixed methods the construct(s) measured by the social connection measures. This is worthwhile exercise, but I am concerned that the methods may not be optimal.
   1. 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.
   2. You say that you will aim for 60% agreement. Is this absolute agreement, or a kappa’s value?
   3. 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 summarise 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?
   4. You propose using Jacard 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).
   5. 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 Jacard indexes.
   6. 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.
   7. 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.
7. I don’t understand the distinction between study 2a and 2b. Are these the same?
8. 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?
9. Can you define feasibility and interpretability on pp29-30 please? How will assessment of these properties happen?

Dr Alex Wilson

Clinical Psychologist, with research background in social communication and psychometrics.