Author's response to reviews

Title: Variability in depression prevalence in early rheumatoid arthritis: A comparison of the CES-D and HAD-D Scales

Authors:

Tanya Covic (t.covic@uws.edu.au)
Julie F Pallant (jpallant@unimelb.edu.au)
Alan Tennant (a.tennant@leeds.ac.uk)
Sally Cox (S.Cox@leeds.ac.uk)
Paul Emery (p.emery@leeds.ac.uk)
Philip G Conaghan (P.Conaghan@leeds.ac.uk)

Version: 2 Date: 17 December 2008

Author's response to reviews:

Dear Editor,

Thank you for the opportunity to revise this paper. We provide a detailed response to reviewers below.

Please not not hesitate to contact me if you require any further clarification.

Kind Regards

Alan Tennant

Details of changes made in response to reviewers comments

Manuscript Title: Variability in depression prevalence in early rheumatoid arthritis: A comparison of the CES-D and HAD-D Scales

We thank both the reviewers for their detailed and thoughtful review of the manuscript and provide below details of the changes that have been made to address each of the points raised.

Reviewer 1

1. Aims of the study (Major Compulsory Revisions)

There is a discrepancy in the manuscript between the stated aims of the study. In the abstract the authors write: “The objective of this study was to test if the CES-D and HADS-D measure the same construct and identify similar levels of depression.” In the end of the introduction the authors write “Consequently, the aims of this study were to: (a) assess the psychometric properties of the CES-D and HADS in a cohort of early RA patients; and (b) to
co-calibrate the items from both scales on to a single metric to determine if they do measure the same construct, and to compare their respective cut points for depression.” And finally the authors state in the first paragraph of the discussion: “The aim of this study was to use Rasch analysis to test the construct validity of two widely used measures of depression in patients with early RA and to co-calibrate these scales onto a single measure of depression to demonstrate that they target the same construct, and to compare their cut points.”

After I have read the paper I think that the purpose of the study is to test the two measures for unidimensionality in the context of Rasch modelling and to test if any items possess DIF. Another aim in project is to merge the items from the two instruments fulfilling the criteria for Rasch-unidimensionality and test if these items fit the Rasch model.

Reply: We have clarified and modified the aims and objectives so that there is now consistency across the text. The objective of this study was to test if the CES-D and HADS-D satisfy current modern psychometric standards for unidimensional measurement; to see if they measure the same construct (i.e. depression) through co-calibration onto the same metric, and to test if they identify similar levels of depression.

My interpretation of the term construct validity is that it is an overarching validity encompassing all relevant aspects of validity. Therefore, content validity is also a part of the construct validity. In the present study the content relevance and the content coverage of the two measures are not assess. Therefore, I think it is misleading to use the term construct validity in the present paper. Alternatively, the authors could use the term statistical construct validity or even better just (Rasch) unidimensionality – and thereby additivity. No matter which terms the authors wants to use it is important that the aims described in the manuscript are consistent and described what is actually done in the study. This is not the case at the moment.

Reply: The Rasch model tests the internal construct validity of a scale – that is the domain or factorial structure and unidimensionality. It cannot for example determine if the scale is measuring depression, just that it measures something in a satisfactory manner. The term internal construct validity has been long established in the literature and we have clarified this in the text, and added appropriate references.

2. Screening (Discretionary Revisions)

In the manuscript the issue about screening for depression is mentioned a number of times. The authors discuss the relation between cut-off points of the different scales and the prevalence of depression in the target population. Another important ethical and clinical problem about screening for depression with a self reported instrument is the predictive values of the scale. This issue is not mentioned in the paper. I think that the manuscript would improve if this topic
was mentioned either in the introduction or in the discussion.

Reply: Yes, this is a good point, although for example, the positive predictive value of the screening instrument cannot be ascertained from the current study as there is no criterion validity to act as a gold standard. Nevertheless, we have added the issue to the discussion.

3. Disordered thresholds (Major Compulsory Revisions)

Ten of the CES-D-items were found to have disordered thresholds, suggesting problems with the 4-point response format. The authors therefore decided to rescore all items by merging the two middle categories (‘some or a little of the time’ and ‘occasionally or a moderate amount of the time’). This problem about the response categories not functioning as intended is not further discussed in the paper. Actually, the authors conclude that the reduced 13-item CES-D scale can be used. However, if there is problems with the four response category format design (disordered threshold can be cause by several different problems) how can the authors conclude that this design is OK to use?

Reply: Yes, we recognise this and have added text to state that further work will need to be done on this matter

4. Reliability (Major Compulsory Revisions)

The authors state in the method section: “Finally a measure of reliability is provided in the form of a Person Separation Index, which can be interpreted as a Cronbach’s alpha. Values of 0.7 and above are suitable for group use, and 0.85 and above for individual use.” However, the Person Separation Index and Cronbach’s alpha both depends on the numbers of items in a scale. Therefore, acceptable or suitable values of the PSI and Cronbach’s alpha are always relative and not absolute as stated in the manuscript. Because the authors are arguing for the use of the two scales as screening instruments I also think they should have suggested as implication for further research to conduct a test-retest reliability study. Furthermore, such a study would be interesting to conduct with a 13-item CES-D scale with reduced numbers of response categories hypothesising an increasing reliability compared to the design with four response categories.

Reply: Yes, this is absolutely correct, and we have added this to the discussion. However, there is a technical problem with the comparison of the three and four items response options as if the latter has disordered thresholds, then fit may be worse, and the theta estimate less accurate, rendering direct comparisons somewhat difficult.

4. Additional test of unidimensionality including PCA and t-test (Major Compulsory Revisions)

The results of these additional tests are only reported in the result section. The methodology of the tests should also be described in the method section
Reply: We have added details of this test to the methods section.

5. On page 12 the authors write:

“In contrast the cut-off value (16) to identify depression for the 20-item CES-D scale [9] gives a prevalence of 45.3%, while the RA-specific cut value of 19 [31] gives a prevalence of 35.9%. As such the depression prevalence is twice as high for the CESD than the HADS-D.” However, the authors are justifying that the 20-item version of the CES-D was not unidimensional and thereby the raw scores of the items were not additive.

Despite this contradiction, I think that the calculation of a sum score of all the 20-CES-D items is relevant for the reader. However, the revealed none-unidimensionality of the 20-item CES-D version is most likely to contribute to the reported high prevalence of depression – because it is not the actual prevalence of depression that is measured it is at least two different constructs. I suggest that the authors discuss this contradiction and what it can cause.

Reply: Yes, we were keen to provide the prevalence based upon the original 20 item scale, even though this is not a valid scale. We have added text to emphasise the limitations of the 20 item version, and the risk that multidimensionality may inflate the prevalence figure.

6. Conclusion (Major Compulsory Revisions)

I think that the conclusion is too long and not all stated in the conclusion is a conclusion, e.g. normally, quotation a not included in a conclusion and I would therefore suggest that the issue about testing other depression scales should be removed to the discussion; page 15 (e.g. Beck’s Depression Inventory[44])

Furthermore I see the part below as suggestions for further research and practice and I would there suggested the authors to split the conclusion into a conclusion-paragraph and a paragraph titled; "Implications for research and practice”

Reply: We have moved the suggestion for testing other scales into the discussion. We have also included a separate paragraph at the end of the discussion addressing the implications for research and clinical practice. We have also expanded the remaining conclusion to make sure we emphasise key findings.

7. Figure 1 (Minor Essential Revisions)

There seem to be a ceiling-effect for the CES-D 13-item version. If so, please provide the reader with this information and if relevant is could be discussed.

Reply: No ceiling effect is present in the current study.
8. Partial Credit Model (PCM) and Rating Scale Model (RSM) (Major Compulsory Revisions)

The reader is introduced to these two Rasch models. However, no results on these analyses are reported or discuss.

Reply: We have added to the results that the unrestricted or partial credit version of the model was used for the analysis.

9. Differential item functioning (DIF) (Major Compulsory Revisions)

Item 3 in the CES-D is revealed to possess uniform-DIF (both in the 13 item version of the CES-D and in the merged scale). In a study conducted by me and my colleagues we also found that this CES-D item possessed DIF. However, we found it possessed none-uniform. The authors decide to keep item 3 in the scales despite the revealed uniform-DIF. The consequences of including an item with uniform-DIF in a scale should be discussed and I would also suggested the topic about adjusting for uniform-DIF to be mentioned (which is actually also described in our DIF-paper).

Reply: Yes, we have added the issue of retaining a DIF item to the discussion.

10. Targeting (Discretionary Revisions)

I would like to suggest three additional figures to be added to the manuscript. In software programme RUMM2020 there is a feature where a distribution of the respondents and the items can be illustrated. To be able to assess the coverage of the continuum measured and to be able to assess any possible redundancy of the items I would suggest item threshold-person-distribution figures added to the manuscript for the three different measures: the 13-item CES-D, the HADS-D and the merged scale.

Reply: As the separate analysis will in effect be on a different metric, (of which the 20 item CES-D is invalid) we thought it more appropriate to add the co-calibration data showing the distribution of persons and items and the hierarchical ordering of items of the co-calibration. We have removed the original Figure 1 as this does not seem to add anything to the text.

Reviewer 2:

Minor Essential Revisions

1) Page 7: write out full meaning of DMARD. Is there a supportive reference for this protocol?

Reply: The full meaning of the abbreviation DMARD has been added to the Methods section. The protocol is an internal clinical protocol and is not published.

2) Page 8, first paragraph, there is no close brackets.
Reply: This has been corrected.

3) Page 8 sentence about meaning of significant Chi-Square – please clarify if deviation from the model means ‘poor fit’.

Reply: This additional clarification has been added.

4) Page 9 second sentence, again please clarify that ‘fit to model’ means a ‘good fit’. In both these instances I think a judgment statement (poor, good) helps to clarify the point being made.

Reply: Yes, we have further clarified this.

Discretionary Revisions

5) Page 6 second paragraph. Please consider clarifying that it is cost / access to resources that prevent screening of all patients with a clinical diagnostic interview and that the confidence in the screening instruments relate to them identifying the same level of depression as the clinical interview.

Reply: A good point, and we have added this to the text.

6) Page 7 last line; can you provide an explanation as to why the higher cut off level is deemed appropriate and does this relate to all RA patients or just early RA patients (in other words do levels of depression vary with duration of disease – early RA peak in depression as adjusting to new condition / acute flare or chronic pain / disability mean longer duration of disease results in greater depression). As the point of your paper is early detection of depression it seems important to have the right cut off for a screening tool for this specific sub group of RA patients.

Reply: The scale and associated cut point should be invariant, irrespective of the duration of disease. The two cut points are indicative of different levels of probability of depression, and this is quite common across depression screening scales. We have added text to make this clearer.

7) Page 10, fourth paragraph. In which order were the items removed? And does it matter?

Reply: This is now added to the text – they were removed according to the magnitude of misfit

8) Can you tell us what the seven items were (or at least the four that were not named in the discussion).

Reply: Details of these items have been provided.