Reviewer's report

Title: Rasch Fit Statistics and Sample Size Considerations for Dichotomous and Polytomous Data

Version: 1 Date: 8 February 2008

Reviewer: Trevor Grahame G Bond

Reviewer's report:

BioMedFitReview

This is an interesting and overdue paper for the field. I support the authors’ intentions, but the paper slips considerably between conception and execution and again between execution and reporting. I have taken more time than I would have wished to comment on Study 1 of this report and suggesting the two studies be reworked and resubmitted as separate but related papers.

1. Is the question posed by the authors well defined?

The authors claim that the “the purpose of this study was therefore to explore the relationship between fit statistics and sample size for both dichotomous and polytomous data.” In fact the focus is more felicitously described as examining the impact of sample size on four routinely used fit statistics for two polychotomous Rasch models. The inclusion of the dichotomous analysis is quite subsidiary to the apparent key purpose.

2. Are the methods appropriate and well described?

Yes: sampling techniques are well described and appropriate. But, given that fit statistics are at the core of this investigation, the authors could take a little more care with their descriptions / definitions.

They claim (p.4): Both mean squares are derived from the squared standardised residuals for each item and each respondent (see Appendix 1).

Better: for each item/person interaction.

The infit mean square is weighted by the variance for each response string [11].

Better (from their cited reference [11]): Residuals are weighted by their individual variance (Wni) to lessen the impact of unexpected responses far from the measure (p285).

Infit mean squares are more sensitive to deviations in responses to items near a person’s measure, whereas outfit mean squares are more sensitive to deviations further away from a person’s expected measure [14].

The term “deviations in responses” is opaque here; pls express in terms of residuals (and well- and poorly-targetted responses, too).

3. Are the data sound?

The data are quite suited to the purpose. We would expect that the age
demographics of the sample would be mentioned in passing (admittedly this is not central to the analysis). Given that the authors compare their results with those from purpose-generated simulated data; some acknowledgements of the limitations of using a genuine data set for examining fit, i.e. only in simulated data sets do we know what the mis-fit of data actually existed.

The HADS and the PHQ seem apt for this task.

p5: The scale consists of 7 items forming a Depression subscale (HADS-D), and 7 items forming an Anxiety subscale (HADS-A). Scores on the two subscales may also be summed to provide a total score (HADS-T), measuring psychological distress [21].

This strong claim presumes additivity within and between sub-scales; yet the results in Table 1 provide ambiguous evidence for this claim. Have the Rasch measurement credentials for these two subscales (HADS-A and HADS-D), the total scale (HADS-T) as well as the PHQ been canvassed in the literature? Please cite. Some evidence of this sort seems to be a pre-requisite for the study. (see comments on Results, below).

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?

Results: Presentation and Interpretation

p7 Items demonstrating more variation than predicted by the model can be considered as not conforming to the unidimensionality requirement of the Rasch model.

Agreed, but the apparently overlooked result in Table 1 shows that HADS-A6 mis-fitting on every indicator on both Rasch models and three of seven HADS items performing poorly. Surely, it would be more conservative and appropriate to analyse / examine the HADS subscale separately. While the mis-fitting and overfitting items are totalled separately in Table 1, they are merely lumped together on p8. Seems there is a prima facie case for insisting that a rewritten version of this paper label and count the mis-fitting and overfitting occurrences separately. Similarly, consistency in use of fit terms within the text, figures and tables would assist the reader.

More importantly, the clinical consequence of item mis-fit (one on HADS-A; least one on HADS-D; and perhaps one on PHQ) remarks completely un-mentioned.

Presentation of results as figures would be more helpful if appropriately similar scales were used on the vertical axes; e.g., Fig 1a has a scale for Outfit t of -1 to +2 (75% of the acceptable range) but from 0 to 2.00 for Outfit Mn Sqr (over 300% of the acceptable range). Requiring the authors to plot the two versions of the fit stats together on the one graph for the figures 1 a-e should oblige them to clarify the representations for themselves to to communicate them more clearly to their readers. Similarly Figs 2-4 need rethinking (leading to re-scaling.)

5. Are the discussion and conclusions well balanced and adequately supported by the data?

The discussion and conclusions are weaker that they should be at a number of
points because:

a. The authors are concerned only with Type I errors (false rejection of items). Yet type II errors: failure to exclude (erroneously including) misfitting items, has potentially far more serious health consequences. The differences between the two error types are not clear here (p.4): â##using mean square statistics may lead researchers to missing significant numbers of misfitting items, which may have an important impact on the development of unidimensional instruments, and that there is, furthermore, a need to understand Type I error rates associated with critical values for fit statistics.â##

b. The authors do not distinguish between the two forms of mis-fit: under-fit (erratic, unpredictable responses) indicated by high fit stats degrades the measures severely; over-fit (Guttman-like, deterministic responses) indicated by low fit stats are generally not harmful to measures (lack of parsimony means potentially greater patient response load) but will overestimate differences in raw scores.

c. Given this is a BioMed Journal, I would expect the issue of potential impact on diagnosis or service evaluations to be canvassed. i.e., what the consequential validity (see Messick, 1989; 1995) of making Type I (and, of course, Type II errors) on HADS / PHQ scores / measures / diagnoses / treatments? Perhaps, this would not be so important in a measurement journal.

6. Are limitations of the work clearly stated?
Limitations are generally over-looked.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?
Yes.

8. Do the title and abstract accurately convey what has been found?
Should be revised as suggested above.

9. Is the writing acceptable?
Yes.

In conclusion, if these issues were dealt with comprehensively, then Study one could stand on its own as an important contribution to the literature. Then Study 2 should be re-configured in light of the comments about study 1 and re-presented as a separate but related study.

- Discretionary Revisions (which are recommendations for improvement but which the author can choose to ignore)
Comment on clinical importance of findings.
Plotting pairs of fit stats on to single graphs.

- Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
reference 11 should be 2001; reference 12 omit â##Baumâ##
- Major Compulsory Revisions (which the author must respond to before a decision on publication can be reached)

Separation of mis-fitting (underfit) from over-fitting (Guttman-like) items for counting and interpretation.

Analysis of HADS-A and HADS-D separately.

Consistent use of terms, e.g.,

â Infit Mn Sq; Outfit Mn Sq; Infit t and Outfit t â could be used consistently throughout the text, figures, tables. Of course, other terms could be chosen, but absolute consistency is important.

Use comparable scales across all figures.

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests