Author's response to reviews

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

Authors:

Adam B Smith (a.b.smith@leeds.ac.uk)
Robert Rush (RRush@qmu.ac.uk)
Lesley J Fallowfield (l.j.fallowfield@sussex.ac.uk)
Galina Velikova (g.velikova@leeds.ac.uk)
Michael Sharpe (michael.sharpe@ed.ac.uk)

Version: 3 Date: 17 March 2008

Author's response to reviews: see over
Dear Sir or Madam,

Re: Rasch Fit Statistics and Sample Size Considerations for Polytomous Data

We are grateful to the reviewers’ for their comments and suggestions, as well as the opportunity to re-submit this manuscript for your consideration for publication in *BMC Medical Research Methodology*.

On the basis of Reviewer 2’s suggestion we have removed Study 2 from the manuscript, and have additionally re-analysed the entire set of samples separately for the HADS-A and HADS-D subscales.

We believe we have addressed all of the reviewers’ concerns, comments and suggestions in full (please find attached).

Please do not hesitate to contact me should you require any further information.

Yours sincerely,

Adam Smith, PhD
Corresponding author
On behalf of the co-authors
Reviewer 1

Major Compulsory Revisions

1. By definition, Type I error refers to the error of rejecting a null hypothesis when it is actually true. In the context of item fit statistic, Type I error refers to mistakenly declaring an item as mis-fitting when it is actually good-fitting. According to this definition, items have to be good-fitting per se. The only way to ensure good-fitting items is to simulate items from Rasch models. This is why this kind of studies uses simulated data rather than real data, because real items will never follow exactly any model and only simulated items can be good-fitting.

Reply 1: The study was concerned with describing the impact of sample size variation on four commonly used fit statistics. Specifically we were interested in observing this impact on misfitting items, rather than on good-fitting items as pointed out by the reviewer. Although we accept the importance of using simulated data to study Type I error rates by definition these datasets do not reflect “true” misfit. It is of importance to test developers and users to know whether and to what extent fit statistics are affected by sample size, similarly it is also critical for test users to understand which statistic(s) will identify misfit correctly and consistently for individual items.

The Rasch model has been previously applied to the HADS (collated from patient data); this has demonstrated some evidence of misfit (Smith et al., 2006; Pallant & Tennant, 2007). This study was therefore designed to further understand how well the four fit statistics identified misfit, and whether this was affected by sample size.

The aim has been redefined to clarify this point:

“Previous research on simulated datasets has focused on the relationship between sample size and fit statistics at the level of groups of items. However, for test users the emphasis is more on which fit statistics are able to identify misfit consistently for individual items. Identification and removal of misfitting items will not only reduce patient burden, but may also improve person measure assessment. [6].

Therefore the aim of this study was to investigate the impact of sample size on four commonly used fit statistics, i.e. Infit / Outfit mean square and their t-statistics for two polytomous Rasch models using data collected from a cancer patient sample.” (p. 5)

2. It is important to distinguish hypothesis testing from effect size measure. ZSTD is for hypothesis testing whereas MNSQ is an effect size measure. A real item will be eventually declared as mis-fitting as long as the sample size is sufficiently large (i.e., a very high power) or conversely, a real item will be eventually declared as good-fitting as long as the sample size is sufficiently small (i.e., a very low power). Hence, effect size measure (i.e., MNSQ) is used to judge the magnitude of misfit. If the magnitude of misfit is not very large, then an item will be declared as practically good-fitting even though it is statistically mis-fitting.

Reply 2: In the literature both mean squares (MNSQ) and fit statistics (e.g. Karabatsos, 2000; Bond & Fox, 2001) are used for hypothesis testing in terms of identifying misfit. The statement by Reviewer 1 that a “real” item will eventually be declared misfitting with a sufficiently large sample size is an empirical one. We have demonstrated that this does not seem to apply for the MNSQ fit statistic; this statistic remained constant across sample sizes, whereas the t-statistic was highly dependent on sample size (and was not able to identify misfit with small samples). In the Reviewer’s terms therefore MNSQ appear to correctly judge the magnitude of misfit.

3. A recent study by Wang and Chen (2005) in the journal of Educational and Psychological Measurement has investigated thoroughly the performance of ZSTD and MNSQ in dichotomous and polytomous items. Being a statistic, the sampling distribution of ZSTD and MNSQ always depends greatly on sample size.

Reply 3: We are grateful to the reviewer for drawing our attention to this important paper and have acknowledged it in our manuscript (p. 4). Wang & Chen demonstrated that the standard
deviation of the MNSQ (for infit / outfit) was dependent on sample size, whereas the ZSTD appeared to be less so, and on the basis of this suggested that MNSQs should be adjusted for sample size accordingly. As we have pointed out above in our response to point 2, our study demonstrated sample size dependence for ZSTD only and not MNSQ. Since ZSTDs are derived from MNSQ and its standard deviation (SD), we believe that the decrease in SD with sample size (as evidenced by Wang & Chen) has a disproportionate effect on ZSTD with a tendency to inflate (or conversely deflate) this statistic as samples increase (or decrease) as has, for instance, been demonstrated by Karabatsos (2000).

We have addressed this point in our conclusion on page 12:

“The potential cause for this sample size dependence for the t-statistics may lie with the standard deviations. The results of previous research have demonstrated that the variability of the mean squares decreases significantly [19] by sample size. As the t-statistics are derived from the mean squares and their standard deviations it appears that t-statistics are disproportionately affected by decreases in variability. The fact that t-statistics are highly dependent on the variance and thereby sample size has also been demonstrated in previous studies with the dichotomous model [15].”

Furthermore, Wang & Chen adduce evidence across items demonstrating that the minima and maxima for MNSQs decrease as sample size increases (Table 5, p. 391). This is similar to the evidence presented by Smith et al. (1998). It is our contention as there is no analysis of individual items presented it is difficult to discern whether this is a global effect of sample size, or indeed whether the MNSQ for individual items are duly influenced by sample size. Our argument, and redefined aim, is that an individual analysis of (mis-)fitting items allows us to address this point.

4. How many replications were there in Study 2? It seems that there was only one replication. If so, then the results shown in Figure 3 may be due to chance.

Reply 4: The methods for Study 2 followed those of Study 1, and there were ten replications for each sample size. However, in line with Reviewer 2’s comments we have removed Study 2 from the manuscript.

5. In the rating scale model and the partial credit model, in addition to the fit statistics (MNSQ and ZSTD) for the overall difficulties, there are fit statistics for the intersection (threshold) parameters. Were these statistics for the intersection (threshold) parameters used to classify items as mis- or good-fitting?

Reply 5: We agree with the reviewer that threshold parameters may also be susceptible to sample size variations, and indeed may impact on fit. However, although we feel that it was not the intention of this study to explore this relationship, the threshold parameters were nevertheless re-assessed for each sample size (and model) for each instrument, and no disordering of thresholds was found.
Reviewer 2

1. Is the question posed by the authors well defined?
The authors claim that the purpose of this study was therefore to explore the relationship between fit statistics and sample size for both dichotomous and polychotomous 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.

Reply 1: Reviewer 2’s primary suggestion was the removal of Study 2 from the manuscript, which we have undertaken. Study 1 is now presented as an independent study. The aim has been rewritten to reflect this change in emphasis.

“Therefore the aim of this study was to investigate the impact of sample size on four commonly used fit statistics, i.e. Infit / Outfit mean square and their t-statistics for two polychotomous Rasch models using data collected from a cancer patient sample.” (p. 3).

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 persons measure, whereas outfit mean squares are more sensitive to deviations further away from a persons expected measure [14]. The term deviations in responses is opaque here, pls express in terms of residuals (and well- and poorly-targetted responses, too).

Reply 2: The wording in the text and Appendix has been amended accordingly:

“Both mean squares are derived from the squared standardised residuals for each item / person interaction (see Appendix 1). The outfit mean square is the average of the standardised residual variance across items and persons and is unweighted meaning that the estimate produced is relatively more affected by unexpected responses distant to item or person measures. For the infit mean square the residuals are weighted by their individual variance (Wni, see Appendix 1) to minimise the impact of unexpected responses far from the measure. The infit mean square is relatively more affected by unexpected closer to item and person measures [11].” (p. 7)

“For the mean square statistics $Z_{ni}$ is the standardised residual variance for each item/person interaction” (p. 16, Appendix).

3a. 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).

Reply 3a: We have included the average age of the sample and reference to the published work from which the data were drawn for further clinical and demographic details.

3b. 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.

Reply 3b: this point has been addressed below in point 6.
3c. 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 subscales; 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).

Reply 3c: There has been some previous Rasch work on the HADS (Smith et al., 2006; Pallant & Tennant, 2007). We have included a description of the Smith et al paper which is more germane to this study given that the data were also collected from cancer patients. As far as we are aware there have been no Rasch analyses of the PHQ-9.

“It has been claimed that scores on the two subscales may also be summed to provide a total score (HADS-T), measuring psychological distress [21]. Previous research in a large heterogeneous cancer population [6] has shown potential misfit on three of the instruments’ items: Anxiety 6 (“I get a sort of frightened feeling”) and Depression 5 and 7 (“I have lost interest in my appearance” and “I can enjoy a good book, radio or TV programme” respectively). This misfit was present both in the full, 14-item version of the HADS as well as for the individual subscales. Although a principal analysis of the residuals did not reveal the presence of any additional factors, given the misfitting items the analysis presented here will focus on the two subscales, HADS-Anxiety (A) and HADS-Depression (D).” (p. 7)

4a. 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.

Reply 4: We have re-analysed the entire HADS samples, creating separate analyses for the two subscales rather than the entire scale previously presented. We have included a justification for this in the text (see reply 3c) on the basis of previous work on the HADS suggesting misfit.

The manuscript has been extensively rewritten to reflect the new analysis. Significant changes to the original can be seen in pp. 8 – 10. In addition, Tables 1 and 2 have been modified to present the re-analysis separately for the HADS subscales. Furthermore, we fully accept the reviewer’s suggestion on the use of consistent terms. Throughout the manuscript we have used infit / outfit mean squares (MNSQ) and t-statistics (t).

4b. 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.

Reply 4b: This important point is addressed in our reply to 5.
4c. 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.)

Reply 4c: We agree with the reviewer that similar scales should be used throughout. We have simplified (and reduced the number of) figures, combining mean squares and t-statistics and pairing the infit and outfit to produce a clearer representation of the results.

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

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

5a. 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): 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.

Reply 5a: The principal aim of this study was to assess Type I errors, however we recognise the importance of Type II errors, although the study was not designed to address these errors, nevertheless we believe that the reviewer’s statement that the latter may have a greater impact on health is one that deserves empirical investigation. In our discussion, we have included work we undertook in a previous published study (Smith et al., 2006) which suggests that the inclusion of misfitting items (full HADS scale)(Type II error) had no detrimental effect on the efficacy of the instrument to detect cases of psychological distress (as measure by the gold-standard psychiatric interview).

Text:

“The presence of underfitting items in instruments may have a potentially significant impact by severely degrading the measures, whereas overfitting items will tend to overestimate differences in raw scores [11]. The former may lead to an under-detection of health problems (e.g. low levels of screening efficacy), the latter may interfere in comparisons within and between individuals. Clearly the need to accurately identify misfitting, particularly underfitting items is paramount. This study demonstrated that low Type I error rates were evidenced by mean square fit statistics, which appeared independent of sample size. The clinical impact of erroneously removing misfitting items has not been directly investigated, however research suggests that the converse problem of retaining misfitting items (Type II errors) has little or no impact on the efficacy of, for instance, instruments used to screen for psychological distress [6, 28]. Research on both the HADS [6] and the Geriatric Depression Scale [28] suggest that misfitting items may be removed from the instruments whilst maintaining, if not improving screening efficacy (in terms of diagnosing cases of anxiety or depression) when compared with a gold standard psychiatric interview. Although the clinical implications of Type I and II errors needs to be explored further the results suggest that correctly identifying misfit has a direct benefit to patients by reducing the burden of the number of questions needing to be answered (whilst maintaining efficacy of the instrument).” (pp. 13-14)

We believe that our study demonstrates that the use of mean-squares does not lead a significant number of misfitting items being missed. Our study demonstrated that mean-squares remained independent of sample size, whereas the ZSTD was unable to detect misfit at low sample sizes.
We agree with the reviewer that it is important to understand the Type I error rates associated with fit statistics. We believe that this study has demonstrated that these error rates are greater for the ZSTD statistic. Furthermore, the greater Type I errors for the ZSTD suggests that a greater number of items may be erroneously discarded at large sample sizes. The impact of this on a screening instrument, for instance, is not known, although the limited evidence available suggests for the moment at least that their inclusion is not detrimental.

5b. 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.

Reply 5b: Text has been amended accordingly (refer to reply 5a).

5c. 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.

5c. This is an important point, which we are grateful to the reviewer for raising. We have addressed the issue of the potential impact on diagnosis in our response to 5a. In addition, we have included a discussion of the wider impact of Type I and II errors on treatment.

The text has been amended accordingly (refer to reply 5a) to address this point.

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

Reply 6: We have addressed the limitations more extensively (p. 13):

“There are a number of limitations to this study: 1). The primary limitation is that “real” data directly derived from patients were used rather than simulated data. Previous work on the HADS in particular had demonstrated the presence of misfitting items in the scale [16]. The aim was observe how effectively the four fit statistics identified misfit and whether and to what extent this was affected by sample size. However, we acknowledge that estimates of Type I error rates are more optimally derived from simulated data where fit and misfit may be artificially manipulated. Further limitations reflect the fact that the data were restricted to cancer patients only, and only included mental health questionnaires; the relationship between sample size and instrument length was not explored, although there were modest differences in test length between the HADS and PHQ-9; and finally any potential interactions with dimensionality and item difficulty [15] were not explored.”

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

Reply 8: The title and abstract have been modified as suggested.

- Discretionary Revisions (which are recommendations for improvement but which the author can choose to ignore)

Comment on clinical importance of findings.

Reply: These have been addressed in our reply to point 5.
- Plotting pairs of fit stats on to single graphs.

Reply: The figures have been simplified and fit statistics have been represented on the same figure for comparison.

- 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 Baum

Reply: Reference 11 is now dated 2001, and “Baum” has been omitted from reference 12.

- 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.

Reply: As described above the HADS-A and D have been analysed separately, and underfit / overfit described individually in the analysis and results.

Consistent use of terms, e.g.,

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

Reply: The text has been modified so that Infit Mn Sq; Outfit Mn Sq; Infit t and Outfit t are used consistently throughout.

- Use comparable scales across all figures.

Reply: All figures have now been rescaled.