Reviewer's report

Title: Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education

Version: 1  Date: 27 April 2004

Reviewer: Tom Bramley

Reviewer's report:

General
This manuscript gives a useful and detailed illustration of how the diagnostic statistical information produced by applying the Rasch model to item-level data from a medical examination can provide insight into the psychometric properties of the examination. It will be of interest to those involved in developing fair and valid tests in this field.

-------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)
1) The first 'key learning point' (pages 19-20) readers are to learn from the paper is that the Rasch model provides the 'test free measurement' which is a prerequisite for establishing a bank of valid calibrated questions. There is not enough in the analysis presented in the paper to justify this claim. The use of 'term in which examination was taken' as a student classification variable for analysis of Differential Item Functioning (DIF) is unusual and rather contrived - the impression is given (on page 16) that somehow something in the content of the item is responsible for its differential performance across terms. It is much more common to use candidate variables such as sex or native language, when significant DIF might more justifiably lead to a revision of the item.
I think that both these criticisms could be addressed by treating the samples taking the test in different terms as separate samples (which seems a more natural approach). Then the invariance of item calibrations across samples could be investigated and demonstrated by plotting the calibrations against each other and superimposing the relevant 'error tramlines' [1]. It may well be that the five items displaying DIF no longer seem out of line within the limits of measurement error.
In summary, I recommend dropping the DIF analysis and converting it to a demonstration of sample-free item calibration to support the first key learning point.

2) The abstract claims that the exam had good psychometric properties, yet the person separation index (reliability) was extremely low at 0.50. On page 10 the authors note that values above 0.8 are considered very good. The low reliability is very briefly ascribed (on page 13) to the homogeneity of the students but this needs to be discussed further, perhaps in terms of classification consistency around the pass/fail cut-score, or in terms of decisions about mastery / non-mastery of the material. It is not sufficient simply to comment that traditional tests of reliability are not appropriate for criterion-referenced tests. The information in Figure 3 could also be useful in this context - it shows that 5 items were extremely difficult and 12-16 items were extremely easy for this particular sample and therefore did not contribute much to their measurement.

-------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
1) At the bottom of page 5 is a paragraph about traditional item analysis. References 8-11 are given with this paragraph, but these references appear to be specifically about standard setting, not item analysis.
2) On page 6 it is claimed that the Rasch model entails a different paradigm from the traditional approaches. This should be clarified, because reference 15 relates to the comparison of Rasch with other IRT approaches, not the traditional approach (if ‘traditional’ here refers to classical test theory).

3) At the bottom of page 9 a sentence about reliability appears in the middle of a paragraph about general tests of fit. This sentence should be in a separate paragraph, unless there is an intention to relate person separation to the statistical power of any tests of fit.

4) On page 16 the last sentence of the first paragraph of the discussion (beginning “This is to be expected…”) appears to be a non-sequitur.

Discretionary Revisions (which the author can choose to ignore)
1) One particular strength of the paper is the inclusion of a principal components analysis of residuals to check for local independence. It is quite rare to see this level of rigour. However, it is not specifically linked to the use of EMQs as opposed to standard multiple-choice questions. One obvious potential criticism of EMQs (likely to occur to readers on reading point a) on page 19 in the discussion) is that the responses to individual items based on the same theme will not be locally independent. The authors should raise this potential objection to show that their analysis has refuted it in this instance! It might even be possible to test specific hypotheses about within-theme / across-theme residual correlations.

2) More clarification could be given about a) where the analysis shows an advantage of EMQs which would not be obtained with a conventional MCQ, and b) about which extra diagnostic statistical information can be gained from applying the Rasch model - for example, some of the analysis of response options (page 15) can be achieved within a conventional item analysis.

3) I am not sure what level of psychometric or statistical sophistication is expected of the reader, but the reference in parentheses to ‘sufficiency’ on page 6 seems rather cryptic. I think Rasch’s main focus was on ‘specific objectivity’ rather than statistical sufficiency.

4) On pages 8 and 19 the word ‘assumptions’ is used regarding the Rasch model, but under the different paradigm of Rasch referred to on page 6, these are not assumptions, but requirements (as one of the authors has stressed elsewhere! [2]).

5) Figure 1 would be much easier to understand if the actual content of the questions had not been removed - but perhaps this was necessary to preserve test security. Similarly it would be interesting to see a discussion of the content of the three items which displayed the misfit but again perhaps this is not possible for the same reason.

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

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No
Declaration of competing interests:

None