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: 19 April 2004

Reviewer: Richard Burton

Reviewer's report:

General
The paper is generally very well written and presented. However, the explanation of equation (1) and Figure 2 could be improved (see below). This is true of many other accounts, so readers may well appreciate help - or a reference to a clear and accessible account. It is not clear how the calculations were performed (using which program).

The paper is largely about the properties of a particular test. That is presumably of only local interest, but the use of Rasch is usefully illustrated and important ideas are aired. Nevertheless, I am not sure that there is much novel in that. What is rare is to see actual examples of item response curves and more might be made of that.

Theoretical problems
I have given reasons elsewhere why I doubt the usefulness of Rasch or IRT analysis in relation to academic examinations [1]. As it stands, this paper does not win me over, so I would see it as very useful if it did so Here some points relating to that:

*Although unidimensionality is required for Rasch analysis, it is not necessarily best for an academic test [1]. Indeed the Introduction acknowledges this for the totality of assessments. One could therefore argue that misfit questions should not be excluded if they are otherwise ‘good’.

*It is well known that the curve of the Rasch model can be a very bad fit to real examination data. Testing this point can require much more data than are here available, so the best one can do is test for significant departures. Failure to find these is not of course the same as validating the model. This does not mean that Rasch analysis with few data is worthless, but it does raise the question of whether classical theory would be as good. (I do note the paragraph starting “The use of the Rasch model entails...”.)

*The Discussion says that the 5 question responses dictated by term should be modified, but I do not see why. Could it not be that the learning situation differed term to term? One may try to standardize learning situations, but epidemics, holidays etc. may prevent that. This brings us to the important issue of ‘difficulty’ This is a clear enough concept with some tests of mental attributes (as also with weight lifting) and it can be obvious that particular medical ideas are easier to grasp than others. However, the measured ‘difficulty’ of an idea can just reflect how many students have met it and paid attention to it. Thus an easy idea is ‘difficult’ if no student has met it. In this way difficulty can vary from one group of students to another and this undermines Rasch analysis. Are there useful points to make about what determines difficulty in these particular questions?

A test (subtest) of only 24 items is far too short to give reliable scores [2]. Consider a simplified model in which 193 students take a 24-item test, with every student having the same % knowledge of the subject (i.e. the sampled knowledge domain) and all possible items being equivalent. Let us take 50% knowledge (not far off the actual mean?). Thus, every student has 0.5 probability of answering each item correctly. I calculate that most students should then score within
the range of 8-16 out of 24 (33-67%), with 0.06 probability of scoring outside that range. I ran the model on Excel, generating answers randomly (all $P = 0.5$): in the first run, the mean mark was $11.8 + sd 2.5$ (range 5-17) and classical item facilities were not identically 0.5, but 0.40-0.55. This is not intended to mimic the actual results, but to exemplify the effects of the vagaries of sampling. Nevertheless, it would be nice to know the actual spread of scores. The paper is not about person abilities (theta) (here not reliably measurable), but uncertainties regarding theta must influence estimates of beta. How important is this? Is there an established/well-known answer to this?

An internal consistency reliability of 0.8 is stated to be ‘very good’, but a test with that value can actually be quite unreliable. This is a common misconception: as the authors recognize in other contexts, reliability coefficients depend very much on the spread of scores, and hence on the spread of student ability. There is no need for the statement anyway.

Other
“These EMQs should be reviewed for ambiguities ...”: Why not do it?
“DIF by gender also needs to be considered”: again, why not go ahead, if this could yield a finding of general interest?

References

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

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

1. Angle the paper to make it more clearly of general interest. I am not sure how to accomplish this, but one possibility might be to compare Rasch and classical analyses (using all the data).
2. Explain the analysis, equation (1) and Figure 2 more clearly.
3. Discuss the issue of unidimensionality in relation to academic tests.
4. Discuss difficulty more fully.
5. If possible, demonstrate (argument and/or data) that Rasch is appropriate and better than classical theory.
6. Discuss the unreliability of theta and the implications of this for beta.

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

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

7. Classical measures should be stated for the 24 as well as the 98, e.g. spread of raw scores, perhaps relating raw scores to abilities.
8. Terms in equation 1 are not all defined. The relation of equation 1 to Figure 2 and ‘location’, needs more explanation. What are ‘item 1’ and ‘locn’? Relate ‘location’ to ‘theta minus beta’? The ‘expected value’ in Figure 2 is a probability.
9. Table 1: State here what ‘Prob’ is.
10. Correct the few typos, e.g. apostrophes (as after EMQs), 25 for 24.

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

Discretionary Revisions (which the author can choose to ignore)
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 limited interest

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests: none