Reviewer's report

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

Version: 2 Date: 23 June 2004

Reviewer: Richard Burton

Reviewer's report:

General
The authors have responded in some way to most of the points previously raised, but I am still left believing that the paper is fundamentally unsatisfactory.
As I said before, the paper is largely about the properties of a particular test, a matter of only local interest. As the details are unavailable, the analysis is less interesting than might be. It is only the set of 24 questions that is analysed fully, and I do not see that the discussion of the full 98 contributes anything useful.
As for the 24, this is too short a test to give reliable results, as I illustrated previously. Its unreliability is now briefly acknowledged, but without obvious conclusions being drawn. I suggest the data are too unreliable for the Rasch model to be adequately tested. Showing that departures from the model are not significant is not the same as demonstrating a convincing fit.

I had commented: “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.” However, there are not even mean, SD, range. A detailed comparison of Rasch and classical measures could be a useful contribution (see references in Burton, 2004) - although more so if the test were longer.
“(How well do the classical and Rasch measures of difficulty correlate?)”

I remain unconvinced of the validity of the Rasch model and that is an important issue; examiners should not waste effort on Rasch if it is invalid or little better than ‘classical’. Several studies have indicated that the two methods give similar results.
Rasch is intended for unidimensional psychological traits. In contrast, a typical academic test samples a domain of facts and ideas (including learned problem-solving skills, perhaps) corresponding to a course, with scores limited to the range 0-100% (0 and 100 being scores not permitted with Rasch).
Defining, say, a trait of ‘intelligence’ suitably, there is no upper limit to scores, and the difficulty of a test item is independent of subject.
In contrast, the difficulty of an academic test item reflects the proportion of a class that does not give the correct answer. It is not generally a matter of intrinsic difficulty, but can depend on the attention the relevant fact has received (affected by teaching, epidemics, students’ use of textbooks etc. Notably, the topic of a ‘difficult’ question one year might get special attention the following year.). Therefore the difficulty of some items may well vary from year to year (as both Rasch and ‘classical’ should reveal). (This not about ‘giving an 8th grade maths question to a second grade student.’)

On unidimensionality, I still do not find the text fully convincing. I agree that tests should be unidimensional for additivity. Presumably also, all test results that you combine should, at least in principle, share the same dimension. (Unfortunately, too many examiners are happy to average any kind of disparate test and project scores!)
I would be surprised if everything taught in the course is unidimensional, so should not the test be equally multidimensional?
As said before, “It is well known that the curve of the Rasch model can be a very bad fit to real examination data.” I still do not see how that can be irrelevant. If “the Rasch model defines measurement”, could one not conclude that measurement is impossible? Hopefully not! Is this about wishful thinking, or measurement with a bent ruler? If I am missing the point, perhaps it needs to be spelled out for us all.
Perhaps the choice of model is not critical here because the range of student ability (theta) is narrow.

The Abstract mentions discussion of the disadvantages of Rasch, but this is hard to find.

References

Summary of main points
1. A particular 24-item EMQ test is analysed, but, because the details cannot be revealed, the results are not very meaningful.
2. The test of 24 items is very unreliable, as is acknowledged and as I demonstrated. Thus the data for analysis are inadequate.
3. The data for the full 98 items are not fully analysed, and because the item details are secret, most of the comments based on these lack interest.
4. The Rasch model is not valid for this kind of test. That does not mean that it cannot yield useful results, but the suggested comparison between Rasch and “classical” is not made. Some of the conclusions could have been reached through the classical approach.

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

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

-------------------------------------------------------------------------------
Discretionary Revisions (which the author can choose to ignore)