Reviewer's report

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

Version: 3  Date: 7 December 2004

Reviewer: Richard Burton

Reviewer's report:

General
This version is not much changed from the 2nd and 1st versions. Many of my comments of my 1st and 2nd reviews have not been acted on, even when they point out simple clarifications for the reader. I shall not go over most of these points again. If such suggestions are inappropriate, I should be told why. My main objections remain (1) that Rasch should not generally apply to academic tests, even in theory, (2) as acknowledged, 24 items and 193 students are far too few to give reliable (meaningful) results, and (3) the paper is too much about a particular test that we are told too little about.

Paragraph 1 of authors'™ response
Yes, there is interest in the EMQ test, but my comment referred to the particular test that is being analysed. It is the particular test that is of only local interest, especially as test items are not stated. We learn little new about EMQ.

As I suggested, means and standard deviations have now been provided, but â€œso that readers can have something familiarâ€. Surely the reader should be told something of the spread of marks for other reasons. I do not see the theoretical objection to calculating means.

Paragraph 2 of authors'™ response
The suggested detailed analysis is not provided, nor any convincing argument given for not doing so. It is not essential, but, with the intention of being constructive, I suggested it as a possible way of extracting useful information from the study.

As I suggested, means and standard deviations have now been provided, but â€œso that readers can have something familiarâ€. Surely the reader should be told something of the spread of marks for other reasons. I do not see the theoretical objection to calculating means.

Paragraph 3 of authors'™ response
The second sentence is logically right, so I seem to be hoist by my own petard. I am guilty of having expressed myself loosely and incompletely there. Tests should indeed be unidimensional for strict additivity. However, I still believe that teaching is, in general, multidimensional, and would be surprised if that were not true of the course actually being tested. Though I disparaged the summing of â€œany kind of disparate test and project scoreâ€™, we cannot generally escape the need to combine scores from different tests somehow and to add up scores for individual tests. Tests can often be seen as testing (by sampling) something like â€œthe proportion of required facts and ideas that have been taughtâ€™. That model is quite different from the Rasch model and from tests of unidimensional psychological traits (for which Rasch is appropriate).

The messiness of real tests and classical theory does not prove the validity of Rasch. (By the way, I am no fan of alpha, but believe that it is often used when it should not be. Indeed, I criticized the use of alpha in the first version of the paper.) Is the last sentence intended to imply that useful knowledge that cannot generate items fitting the model should not be tested? Probably not - given the second sentence of the Introduction.
Paragraphs 4 and 5 of authors’ response
Elegant does not necessarily mean correct. Rasch is no better than the classical approach as a basis for item banking. I pointed out that the ‘difficulty’ of an item may vary from year to year as a response to variations in teaching, epidemics etc. The authors’ agree, but dismiss it as an ‘explanatory issue’. They do not say whether items should be omitted from the bank when they are affected by these variations, or whether recalibration is required every year.

Paragraph 6 of authors’ response
The authors accept my concern about the reliability of a 24-item test (page 11). However, they still do not discuss the possibility (even likelihood) that the whole study is thereby invalidated.

Paragraph 7 of authors’ response
There might be some truth in this, but we cannot really conclude much about the different specialities in terms of generalizations, because we are told nothing of the actual questions and teaching methods/standards.

Figure 7 (added)
This needs explanation. We are left to guess at its meaning. Excel allows the irrelevant ‘Item 3’ box to be deleted.

Classical item-discrimination indices deserve a mention. The unreliability of these (Burton, 2001, Assessment & Evaluation in Higher Education 26, 213-) would apply to the (Rasch) data here too.

I had suggested inclusion of more actual examples of item response curves, because it is hard to find examples elsewhere. I realise now that the 3-point curves are not what I had in mind in writing that.

Terms in the equation are still undefined.

Figure 2 still has those unexplained numbers and abbreviations in blue: if they are useful, explain them; if they are irrelevant, remove them.

Presumably ‘bad’ questions have been removed from the bank: could these be usefully revealed?

Table 1. How is this helpful if we do not know the questions? Why not summarize the points of interest as text?

Page 10. It is not really explained what one does with the three groups, I think.

Page 15, line 3. ‘would be’ or ‘is’?

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

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)