Author's response to reviews

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

Authors:

Bipin Bhakta (B.Bhakta@leeds.ac.uk)
Prof Alan Tennant (alantennant@compuserve.com)
Gemma Lawton (gemma_lawton@hotmail.com)
David Andrich (D.Andrich@murdoch.edu.au)

Version: 2 Date: 11 Jun 2004

Dear Editorial team.

We would like to thank the reviewers for their useful comments which we feel have greatly strengthened the paper. We reply to each point raised below:

Richard Burton Reviewer:

Reviewer's report:

However, the explanation of equation (1) and Figure 2 could be improved (see below). We have expanded on the explanation of each.

It is not clear how the calculations were performed (using which program). We have added details of the underlying algorithm, and details of the software used.

Although unidimensionality is required for Rasch analysis, it is not necessarily best for an academic test [1].

It depends upon the use to which the test is put. Unidimensionality is a requirement for the summation of items, which is the way in which we use the test. We have added references to support this position, and state so explicitly in the text.

It is well known that the curve of the Rasch model can be a very bad fit to real examination data.

The Rasch model defines measurement, and thus is the standard against which we judge whether or not we have attained measurement. Alternative IRT models, for example the 2- and 3-PL models, lack the property of invariance, and fail thus to deliver measurement - they are explanatory models, not measurement models. We have added this to the discussion.

On thing wrong with the Rasch model is complete knowledge (100% score) is excluded, though possible in a real test.
Indeed this is the case, although any test which had floor or ceiling effects in this way is incorrectly targeted and not valid for the group under test.

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’

The notion of Differential Item functioning is subtly different from distributional aspects. The test is that at any given level of knowledge the probability of attaining a correct response differs, in this case, by term, or successive samples. It does not preclude distributional difference bought about by the vagaries of learning situations.

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.

This applies to all learning situations. You would not give an 8th grade maths question to a second grade student. They have not ‘met’ the maths yet, and so it would be extraordinarily difficult. Again, this is a question of appropriate targeting, and has nothing to do with Rasch. Rasch analysis is all about the difference between ability and difficulty. The property of invariance of the ratio of difficulty across items (that is this ratio between any two items is constant, irrespective of the ability of the students) is again an essential requirement for measurement. We have added this fact to the discussion. Thus the ratio of question difficulty, one to another, does not alter, rather the ability of the student and thus, under the cumulative Rasch model, their probability of success.

A test (subtest) of only 24 items is far too short to give reliable scores [2]. Nevertheless, it would be nice to know the actual spread of scores.

We have added a statement to reflect this, and a reference.

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?

It is true that the sample size does have an effect of the degree of precision of both person and item estimates. It is our understanding that some of the underlying variations in the maximum likelihood approaches which underpin the Rasch computer programs may also influence this (e.g. unconditional versus pairwise conditional), but as these programs have been evolving continuously, this is very hard to establish. In the pairwise conditional approach used in RUMM2020 person estimates are conditioned out (via raw score) and item difficulties are thus estimated independently of distribution.

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.

We agree, and have removed the statement.
"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?

The information is being fed back to the panel who write these questions. We did not have gender recorded for these examinations, but will do so subsequently, and we mention the importance of this in the discussion.

Tom Bramley Reviewer:
Reviewer’s report:

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 justification for this claim arises from the mathematical properties of the Rasch model and, specifically, the separation of item and person parameters consistent with the notion of specific objectivity. We have added these facts to the discussion.

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].

Yes, this would be another way at looking at DIF, but see the concluding remarks of Holland PW, Wainer H. Differential Item Functioning. Hillsdale, New Jersey: Lawrence Erlbaum, 1993 where they argue that “The older approach of fitting separate IRT models to the data from two different groups and then comparing the item parameters that result make no allowance for the fact that the latent traits that are separately estimated may not both have a common interpretation in the two groups”

In addition, the approach we use (i.e. common calibration) allows for identification of both uniform and non-uniform DIF, the latter being group-trait interaction, and which cannot be identified through the separate calibration method.

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.

Yes, this is potentially confusing to readers. We have removed the statement about reliability an 0.8, and have added a paragraph on reliability in the discussion.

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.

Yes, we are aware of this, but make the point that some items, albeit easy, are important to show that the students have mastered the required material.

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).

We have corrected this as we intended to make this a contrast between Rasch and other IRT approaches.

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.

This was an error, and we have dealt with reliability in the discussion.

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. Again, an error, which is corrected as above.

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.

Yes, we see that this is an important point. And have added this to our discussion.

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.

We have included evidence to this effect in the introduction, but pick up the point again in the
discussion and re-emphasis the approach within the framework of the Rasch measurement model.

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.

We have amended the text accordingly.

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]).

To avoid confusion we have standardized upon assumption. We have done this in order to separate out the assumptions underlying the Rasch model, with the requirements of the type of data needed for mathematical operations.

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.

Yes, we cannot give examples taken from the item bank for reasons of security, and we acknowledge that this may affect our ability to describe aspects in detail.