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)
Alan Tennant (alantennant@compuserve.com)
Gemma Lawton (gemma_lawton@hotmail.com)
Michael Horton (m.c.horton@leeds.ac.uk)
David Andrich (D.Andrich@murdoch.edu.au)

Version: 3  Date: 24 November 2004

Author's response to reviews:

Dear BMC,

Thanks you for the opportunity to further revise this paper. We respond to the reviewers comments in detail below,

Reviewers report: Richard Burton

We are not at all convinced by the arguments, scientific or otherwise, put forward by this reviewer. The EMQ test itself is increasingly used within undergraduate medical education in Europe, and thus there is intrinsic interest. Furthermore, the paper is also about an introduction to an approach to assessing this form of exam question and will be of general interest and have general applicability.

We have added some classical analyses so that readers can compare to what they are familiar, and have also added means and standard deviations of the test scores, again so that readers can have something familiar. However, we must point out that we believe this type of data do not really support such arithmetic-based or normally distributed operations.

The reviewer comments that examiners should not waste effort on Rasch analysis if it is invalid or little better than the classical approach. The fact that the reviewer is arguing that typical academic tests sample a domain of facts, and are thus are implicitly multidimensional, but then agrees that tests should be unidimensional and additive, is consistent with the weakness of the argument. It is a mathematical requirement that any summated set of items be unidimensional. This is stated explicitly in the text. It is rare to see any justification for a summed raw score from classical approaches; that is the test forms a unidimensional scale and consequently that the summed score is valid. Thus the Rasch approach recognises the mathematical assumptions underneath the summation of a set of test items; and formally
tests this. Classical approaches either ignore this or rely upon alpha as a surrogate. In either case the position is in error (for example, you can have high alpha with a multidimensional scale) and, consequently, reflect poor science. Whatever is taught, to summate a set of items, the test must be unidimensional.

Furthermore, the Rasch approach offers an elegant unified approach to all the issues such as unidimensionality; Differential Item Functioning and distractor analysis. It also provides the basis for item banking onto a calibrated metric, something that classical approaches cannot deliver.

The fact that exogenous matters may affect item difficulty from year to year is independent of classical or Rasch approaches. This is an explanatory issue, not one of basic measurement. Again, the Rasch approach will identify if the unidimensionality of the test, and thus the validity of the raw score, is compromised in any particular year (and what item(s) contribute). Just why that is the case is then a matter of investigation and explanation..

The reviewer expressed concern about the reliability of a 24 item test, based upon his own work. There is clearly a problem of reliability in such tests and we have stated this explicitly in the text.

We also believe that readers would be interested in the exploratory analysis of the 98 questions, as this describes all the medical specialties, and allows for a simple comparison of specialty-specific item difficulties across the trait.

Reviewers report: Tom Bramley

We have modified the test to emphasise the variability of learning situations as opposed to the intrinsic difficulty of the test items across administrations.