How robust are your examination questions? The psychometrics of student assessment.

Introduction.

Pressures for accountability, transparency and consistency make reliability in assessment a compelling requirement. Yet, how can we be confident in the robustness of our assessment methods? What measures can we take to maintain the standard of difficultly of assessments within our programmes?

The Rasch Measurement Model may have some useful answers and comparing this modern psychometric tool with more traditional tools is a way to illustrate its usefulness. Traditional psychometric analysis of assessments used to examine undergraduate attainment is based on classical test theory whereby the focus of analysis is on the total test score (to determine pass marks, grades of distinction etc); frequency of correct responses (to indicate question difficulty); pattern of response to individual questions; reliability of the test and item-total correlation (to evaluate discrimination at the item level). Although these statistics are used widely in academia, one important limitation is that they relate to the specific sample under scrutiny and thus are sample dependent. In contrast, the Rasch measurement model focuses on the analysis on item responses such that the calibration of the items (questions) is independent of the sample of students taking the examination. The analysis also investigates key attributes such as unidimensionality which is a prerequisite for using total scores to judge whether a student has passed the test.

The Rasch Measurement Model

The Rasch model is based on a probabilistic form of Guttman scaling. It asserts that the easier the question, the more likely it will be answered correctly, and the more able the student, the more likely they will answer an item correctly compared to a less able student. It assumes that the probability that a student will correctly answer a question is a logistic function of the difference between the student's ability [θ] and the difficulty of the question [b], and only a function of that difference. (The equation of the Rasch model is shown on the right.) From this, the expected pattern of responses to a set of questions is determined given the estimated θ and b. When the observed response pattern does not deviate too much from the expected response pattern, then the questions constitute a true Rasch scale.

$$p_{ni} = \frac{e^{(\theta n - b_i)}}{1 + e^{(\theta n - b_i)}}$$

Where p_{ni} is the probability that person n will answer item i correctly [or be able to do the task specified by that item], θ is person ability, and b is the item difficulty parameter.

A main advantage of the Rasch model is that the item difficulty and person ability parameters are derived independently and therefore the item analysis is not dependent upon the student sample from which it was taken. Where data do not conform to the expectations of the Rasch model, the main challenge is *not* to find a model that better accounts for the data, but to identify and explain statistical misfit. By understanding the lack of fit of the data to the model, the examination writer can reflect on the validity of individual questions and decide whether to construct more valid questions.

Person and Item fit

To determine how well each question fits the Rasch model, and contributes to defining a single dimension, a set of 'fit' statistics are used. These statistics include overall fit statistics as well as fit statistics for individual questions (and students). Statistics indicating fit to the model test how far the observed data match the model expectation. Misfit of an item indicates a lack of the expected probabilistic relationship between the item and other items in the scale. This may indicate that the item does not contribute to the trait under consideration. That is, something other than what we are trying to measure has a much stronger influence on the response than we would like.

Differential item functioning (DIF)

Within the framework of Rasch measurement, the exam should work in the same way, irrespective of which group is being assessed. Thus, the probability of a student answering a question correctly – at a given level of ability – should be the same for younger or older students, males and females, and so on. This type of analysis is given the name Differential Item Functioning (DIF). A DIF Analysis can be easily carried out within the Rasch framework to determine whether there is any systematic bias among the items for whichever groups are being assessed.

Conclusion

The Rasch Model can provide an excellent foundation with which to investigate robustness of student assessments. Rasch Analysis can provide a wealth of information relating to each individual student and exam item, as well as for the assessment as a whole, with regard to whether an assessment is measuring what it is meant to measure and is free from any external bias. As Rasch Analysis is not sample dependent, it is also a useful tool to develop calibrated item banks that help to ensure that high-stakes examinations remain of a comparative difficulty between different cohorts of students, meaning that standards can be maintained.

Further information about the Rasch measurement model and contact details are provided on our website http://www.leeds.ac.uk/medicine/rehabmed/psychometric/index.htm

Ref: Bhakta B, Tennant A, Lawton G, Horton M, Andrich D. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. BMC Medical Education 2005; 5:9.

Bhakta BB[1], Horton M[1], Levesley M[2], Tennant A[1]
[1]Academic Department of Rehabilitation Medicine, University of Leeds, UK.
[2]School of Mechanical Engineering, University of Leeds, U.K.