Rasch measurement: a response to Payanides, Robinson and Tymms[1]

Abstract

A response is made to a paper that urges the use of the Rasch model for educational assessment. It is argued that this model is inadequate, and that claims for its efficacy are exaggerated and technically weak.

[1] This paper was submitted to the British Educational Research Journal as a response to the published article. Rather surprisingly it was rejected on the advice of an editor without even being sent to review. The reason given was "on criteria of appropriateness and relevance". Given that it was a response to a previous article and that the journal often takes such responses, this seems a very strange decision. Hence its appearance on this web site so that those interested can read both viewpoints.
As a matter of interest the current editors are: Vivienne Baumfield (Glasgow), Ian Mentor (Glasgow), Christine Skelton (Birmingham), Gary Thomas (Birmingham).

Introduction

Payanides et al (2010) seek to resurrect the so called Rasch test score model, discussing the history of its use in the UK and arguing against those who have been critical of its use. My own writings about this feature largely in their critique, and there are several issues that I would like to respond to.

The authors deal very briefly with the period around 1980 when the utility of using the Rasch model was debated within the DES. They mention two seminars held by the Assessment of Performance Unit (APU) and complain that the National Foundation for Educational Research and the APU 'bowed under pressure' to drop the use of Rasch. What they fail to mention is that those seminars included several leading assessment experts at the time and it became clear at those seminars that the advocates of using Rasch, notably Bruce Choppin, had a weak case and essentially lost the argument. It was this failure to make a convincing case that led to the dropping of the use of this model for the APU andalso in other areas.

The technical weaknesses of the Rasch model for national assessment were discussed by me at the time (Goldstein, 1980) and it was this analysis that helped to inform the debate. Since the 1980s things have certainly moved on, as Payanides et al point out, but the essence of the criticisms remains and centres around the claim that the model provides a means of providing comparability over time and contexts when different test items are used. If such a claim were true then there would be no problem with making statements about changes in 'standards' or comparing individuals in different educational systems who take different versions of a test etc. In fact, this all remains very much an area for debate (see for example, Newton et al., 2008)

I do not wish to rehearse the detailed arguments here. I would, however, like to correct some misconceptions and technical inaccuracies in the Payanides et al paper.

Misconceptions and inaccuracies

First, the so-called 'classical' test score model and the more recent 'Item response' models, of which the Rasch model is a special case, are actually very similar, differing only in terms of how the observed item response (e.g. correct/incorrect) is related to terms describing individual 'ability' and each item's 'difficulty'. Goldstein and Wood(1989) describe this in detail. In particular, all claims about item characteristics being group-independent and abilities being test-independent, can be applied, to both types of model. By failing to point this out, the authors claim that the Rasch model was a 'revolutionary' innovation, becomes very thin.

Secondly, Payanides et al do not seem to appreciate the importance of the unidimensionality assumption made by the Rasch model. In my 1980 paper (not referenced by Payanides et al.) I showed how a 2-dimensional set of items (representing different aspects of mathematics) could actually appear to conform to a (unidimensional) Rasch model , so that fitting the latter would be misleading. Payanides et al also seem to be unaware of more recent generalisations of Rasch and other item response models to include multidimensionality, especially within a multilevel structure (see e.g. Goldstein et al, 2007).

Thirdly, the authors claim that there are no sample distributional assumptions associated with the Rasch model. This cannot be true, however, since all the procedures used to estimate the model 'parameters', such as maximum likelihood, and in common with all statistical models, necessarily make distributional assumptions.

Fourthly, In their discussion of item 'invariance' the authors make it fairly clear why they favour the Rasch model. They claim that a 'fundamental requirement' for measurement is that for every possible individual the 'difficulty' order of all items is the same. This is, of course, a position that one can take, but is extremely restrictive. It is also one that can be tested on any given assessment, and as Goldstein et al (2007) demonstrate, can be shown not to hold, at least in some cases, where the Rasch model has been used. I also find it difficult to see any theoretical justification for such invariance to be a desirable property of a measuring instrument.

Fifthly, the authors do not seem to appreciate the problem of item dependency. The example they give of items *designed* to be dependent is irrelevant. There are all kinds of subtle ways in which later responses can be influenced by earlier ones, over and above an individual's 'ability' and this is extremely difficult to detect, and as far as I am aware, almost never is studied.

Sixthly, the authors elaborate their stance in their response to criticism 7 by stating that 'the aim of measurement should not be to accommodate the test data, but to satisfy the requirements of measurement'. This comes dangerously close to saying that the data have to fit the preconceived model rather than finding a model that fits the data. It is quite opposed to the usual statistical procedure whereby models (of increasing complexity) are developed to describe data structures. Indeed, the authors are quite clear that the idea of 'blaming the data rather than the model' is an important shift from standard statistical approaches. In my view that is precisely the weakness of the authors' approach.

Conclusion

Finally, perhaps the most depressing aspect of this paper is that it appears to be stuck in a time warp. Since the original work in the 1970s and 1980s, item response modelling has moved on. The old Rasch formulation is just one, oversimple, special case. All of these models are in fact special kinds of factor analysis, or structural equation, models which have binary or ordered responses rather than continuous ones. As such they can be elaborated to describe complex data structures, including the study of individual covariates that may be related to the responses, multiple factors or dimensions, and can be embedded within multilevel structures.

The original issues of how to obtain comparability of test measures over time and place remain important ones for debate. Attempting to resurrect the Rasch model contributes nothing new.

References

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology **33**: 234-246.

Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. British Journal of Mathematical and Statistical Psychology **42**: 139-167.

Goldstein, H., Bonnet, G. and Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. Journal of Educational and Behavioural Statistics **32**: 252-286.

Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P., Eds. (2008). Techniques for monitoring the comparability of examination standards. London, Qualifications and Curriculum Authority.

Payanides, P., Robinson, C. and Tymms, P. (2010). The assessment revolution that has pased England by: Rasch measurement. British Educational Research Journal **36**: 611-626.