OPEN ACCESS

University of BRISTOL

Goldstein, H. (2015). Rasch measurement: a response to Payanides, Robinson and Tymms. *British Educational Research Journal*, *41*(1), 176-179. https://doi.org/10.1002/berj.3170

Peer reviewed version

Link to published version (if available):
10.1002/berj.3170

Link to publication record on the Bristol Research Portal
PDF-document

This is the peer reviewed version of the following article: Goldstein, H. (2015), Rasch measurement: a response to Payanides, Robinson and Tymms. British Educational Research Journal, 41: 176–179. doi: 10.1002/berj.3170, which has been published in final form at dx.doi.org/10.1002/berj.3170. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

University of Bristol – Bristol Research Portal

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/red/research-policy/pure/user-guides/brp-terms/

Rasch measurement: a response to Payanides, Robinson and Tymms

Abstract

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

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. Some of my own writings in this area feature in their critique, and there are several issues that I would like to respond to. First, however, it will be useful to place the Rasch model in context.

Using the notation adopted by Payanides et al. the model specifies a relationship between an individual's observed responses to a set of dichotomous (correct/incorrect) test items and an assumed 1-dimensional individual 'ability'. It has the form

$$logit\big(p_i(\theta)\big) = \log\left(\frac{p_i(\theta)}{1-p_i(\theta)}\right) = \theta - b_j \qquad (1)$$

where $p_i(\theta)$ is the probability that an individual with ability θ gives a correct response to item i and b_i is interpreted as the 'difficulty of the item. This specifies a particularly simple mathematical relationship between an individual testee's ability and the probability of a correct response to each item in an educational or psychological test. It has the useful property that an estimate of an individual testee's ability, given a set of correct/incorrect test item responses, is a simple (nonlinear) function of the number of correct responses, the so-called raw score.

More traditionally assessment practitioners have used a basic version of the 'classical' test model that the authors refer to, in order to derive an ability estimate and this model can be written as

$$p_i(\theta) = \theta - b_j \qquad (2)$$

with a similar interpretation for the θ and b_i. In this case the ability estimate is simply the raw score itself and in fact both (1) and (2) ability estimates will rank individuals in exactly the same order. From this perspective the Rasch model does not represent such a 'breakthrough' that its proponents have tended to claim.

In both formulations the model can be elaborated , for example by including a 'discrimination' parameter so that (1) and (2) become respectively

$$logit\big(p_i(\theta)\big) = \mathrm{a_i}\theta - b_j$$

and

$$p_i(\theta) = a_i\theta - b_j.$$

Such models are often referred to as two-parameter models. Goldstein and Wood (1989) provide further details and examples.

History

Payanides et al. deal very briefly with the period around 1980 when the utility of using the Rasch model was debated within the Department for Education and Science (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 was largely responsible for the dropping of the use of this model for the APU and also in other areas.

The technical weaknesses of the Rasch model for national assessment were discussed by myself 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. This is of course one of the Rasch model's attractions, but In fact, this all remains very much an area for debate (see for example, Newton et al., 2008)

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

Misconceptions and inaccuracies

First, as pointed out above, the 'classical' test score model and the more recent 'Item response' models, of which the Rasch model is a special case, are actually very similar. 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 essence this states that, while items themselves may differ in ability, there is only a single ability that characterises an individual that determines that

individual's response to each item. In my 1980 paper (not referenced by Payanides et al.) I showed how an actual 2-dimensional set of items (representing separate algebra and geometry abilities) could 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, and also to incorporate predictors such as social background, 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 the procedures used to estimate the model parameters, such as maximum likelihood, necessarily make distributional assumptions. Indeed, they themselves describe the Rasch model as a probabilistic one.

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 envisage any convincing 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 studied.

Sixthly, the authors state 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 the Payanides et al. 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 Rasch formulation they describe is just one, simple, special case. All of these models

are in fact particular 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 they can be embedded within the multilevel data structures that are ubiquitous in educational research.

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 passed England by: Rasch measurement. British Educational Research Journal **36**: 611-626.