J Rehabil Med 2011; 43: 1045–1046

LETTER TO THE EDITOR

COMMENT ON "COMMENTARY ON: PAST AND PRESENT ISSUES IN RASCH ANALYSIS: THE FIM REVISITED"

We would like to thank Allen W. Heinemann and Anne Deutsch for their commentary (1) on our paper (2). We entirely concur with their reminder of the debt owed by many to the work of Benjamin Wright and Mike Linacre at the Measurement, Evaluation, Statistics and Assessment (MESA) Psychometric Laboratory in Chicago, USA. It was through their efforts that, by-and-large, the Rasch model was disseminated into the area of health. We also concur with their reminder that the earlier work had an overriding purpose of introducing the Rasch model and, particularly, to emphasize that ordinal raw scores are not interval measures. While the former is now well established and, with over 1500 Rasch papers indexed in MEDLINE, stands as a testament to their efforts, the latter remains a challenge, given the inertia in the health outcomes community, and the frequent application of mathematical operations to ordinal scales.

It is interesting to note that in their Commentary Heinemann & Deutsch (1) raise important questions as to the magnitude of the difference made to person estimates by the various modifications that can be made through collapsing of categories, and deletion of items that do not fit model expectations. It may be that the difference is, in fact, small, but that difference may reflect a solution that does satisfy the model assumptions and expectations, as opposed to one that does not. This is akin to a *p*-value of 0.06, which may reflect, in a clinical trial, very little difference in the magnitude of effect compared with that associated with a *p*-value of 0.05, but nevertheless, by convention, would fail to show a significant difference. Should we be concerned about this, and report that the study did show an effect? With respect to the Rasch model, either the data fit the model, or they do not, and person estimates under the latter scenario, are not valid, irrespective of the magnitude of difference from a fitting solution. Good science dictates that we specify in advance the acceptable parameters of fit for our analysis, and make our judgements about fit accordingly. Of more concern is the wide range of "acceptable" fit parameters to be found in the literature. Relaxing fit requirements and ignoring, for example, scientific evidence with regard to the appropriate range of fit statistics for a given sample size, consistent with a 0.05 Type I error rate, is a continuing threat to the integrity of Rasch analysis (3).

On two occasions Heinemann & Deutsch (1) allude to item response theory (IRT). They are particularly concerned about the potential relegation of the Rasch model to small samples, as a special case of the wider IRT approach. However, it is important to note here that some argue that the Rasch approach is so fundamentally different from an epistemological perspective to the rest of IRT, that it is incompatible with those other approaches (4). Thus, fitting data to the Rasch model is concerned with constructing measurement and, consistent with this, the model has special properties of sufficiency, specific objectivity, freedom from distributional properties, and has fully testable assumptions. Mathematical proofs of the compliance with a probabilistic version of additive conjoint measurement have been published, thus confirming the interval scale latent estimate of the Rasch model (5). Other IRT models simply do not have these attributes, despite the propensity of their proponents to suggest otherwise. Consequently, the IRT approach is concerned with statistical modelling of data of the sort we are familiar with in techniques such as regression. As Lord (6) stated long ago, these techniques deliver ordinal estimates. They also have peculiar interpretation for existing legacy scales, where two people with the same raw score can be given different latent estimates of the attribute being "measured".

There are now a substantial number of scales that have deliberately been built to Rasch model standards, and so benefit from all the advantages listed above. Looking from the opposite perspective, IRT-based scales do not have sufficiency; that is, the clinician cannot simply add up the raw score and obtain an interval scale estimate from a simple exchange table; IRT models are also sample-dependent and their estimates cannot pass from one sample to another. Thus, despite the amount of investment in some quarters into IRT applications in the area of health, they lead the health outcome community away from a simple patient-centred perspective where a paper-and-pencil test can be added up to provide, with an exchange table, an interval scale estimate of the attribute being measured. One of the many advantages of the Rasch model is that it supports this type of simple application, as well as the more complex computer-adaptive testing solutions that are beginning to emerge (7). Consequently it offers the opportunity for a fully integrated person-centred approach to outcome measurement, providing comparable interval scale estimates from all possible platforms, in all settings, and with varying levels of resource.

The interpretation and use of the response categories is also a point for further discussion. With 7 response options for each item in the Functional Independence Measure (FIM™), there is always the chance that some options are not used, or that categories (or specifically the thresholds between categories) become disordered. There are technical issues that may affect interpretation at this level; for example, the pair-wise conditional maximum likelihood estimate that underpins the Rasch Unidimensional Measurement Models programme is more robust to null categories. However, when disordering categories occur (a score of 3 on the item represents more independence than does a score of 4) it is also a problem for the clinical utility. Categories are used not only for sum scores in the clinic, but also for setting treatment goals, following progress and

making discharge decisions. The gain of a scale point over time is increasingly used as a measure of effectiveness at the clinic, and reported in national quality registers, or even on the clinics' web homepage, so that patients can compare different clinics when choosing where to be treated.

The appropriateness of the partial credit parameterization of the Rasch model, over that of the rating scale model, asserts the lack of equidistance between categories across items. Within items, we acknowledge that the distance between 6 and 7 can be small and, for example, between 1 and 2 much larger, which is characteristic of the ordinal scale. But do clinicians or other health professionals also have this in mind when making decisions on who to discharge? A reliable and effective rating scale is one of the fundamental requirements for clinical utility and decision-making. It is possible that there is a limit as to how many categories any clinician can realistically distinguish, emphasizing that definitions need to be extremely clear. Further research is needed into this aspect.

Heinemann & Deutsch (1) also raise a number of technical points about the article. The sample was a single sample on admission to a rehabilitation hospital, where raters had all received formal training in FIM assignment. While we acknowledge that assessments at discharge may have given a different distribution, other than the magnitude of error associated with item estimates, we do not agree that the results would have been different. Where data fit the Rasch model, the item hierarchy should be invariant, and this has been shown for the FIM™ in other settings where admission and discharge data have been compared (8).

We agree that if an item simply misfits the model expectation, after all attempts to rectify problems, for example through splitting for differential item functioning, then it compromises the scale validity and should be removed. However, our position on item deletion as a last resort reflects the frequently expressed concern of the clinical team that "this item is important to our clinical management", and we would want every effort made to retain the item set where possible, even if there is obvious redundancy from a measurement perspective. The article on the FIM™, upon which the Commentary is based, has shown that we must be careful to consider the impact of local response dependency upon fit before taking decisions about deletion. One wonders how many times the claim that "the data never fit the Rasch model" is in fact a reflection of the failure to take account of such factors.

For the results of the FIM™ in our paper, the really positive outcome is the synergy between clinical usefulness (in retaining all the items) and the psychometric requirements, such that application of the Rasch approach can support the identification of both the psychometrically sound, and the clinically useful, scale.

REFERENCES

1. Heinemann AW, Deutsch A. Commentary on "Past and present issues in Rasch analysis: the FIM revisited". J Rehabil Med 2011; 43: 958–960.
2. Lundgren Nilsson Å, Tennant A. Past and present issues in Rasch analysis: the Functional Independence Measure (FIM™) revisited. J Rehabil Med 2011 43: 884–891.
3. Smith R. M. Fit analysis in latent trait measurement models. J Appl Meas 2000; 1: 199–218.
4. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004; 42 Suppl 1: S17–S16.
5. Van Newby A, Conner GR, Bunderson CV. The Rasch model and additive conjoint measurement. J Appl Meas 2009; 10: 348–354.
6. Lord FM. The "ability" scale in item characteristic curve theory. Psychometrika 1975; 40: 205–217.
7. Elhan AH, Oztuna D, Kutlay S, Kucukdeveci AA, Tennant A. An initial application of computerized adaptive testing (CAT) for measuring disability in patients with low back pain. BMC Musculoskel Dis 2008; 9: 166.
8. Küçükdeveci AA, Yavuzer G, Ehan AH, Sonel B, Tennant A. Adaptation of the Functional Independence Measure for use in Turkey. Clin Rehabil 2001; 15: 311–319.

Submitted September 26, 2011; accepted September 26, 2011

Åsa Lundgren Nilsson, PhD[1]** *and Alan Tennant, PhD*[2]
From the [1]Institute of Neuroscience and Physiology, Department of Clinical Neuroscience and Rehabilitation, University of Gothenburg, Gothenburg, Sweden and [2]Department of Rehabilitation Medicine, Faculty of Medicine and Health, The University of Leeds, Leeds, UK.
*E-mail: asa.lundgren-nilsson@neuro.gu.se