J Rehabil Med 2011; 43: 958–960

**COMMENTARY**

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

Lundgren Nilsson & Tennant (1) are to be congratulated for providing a succinct and focused summary of analytic issues that are important to consider when applying the Rasch model to the FIM instrument. Although the early published analyses appear to be crude in hindsight, the purpose of these early reports was to illustrate the benefits of the Rasch model in contrast to Classical Test Theory approaches. For example, Merbitz et al. (2) illustrated the misunderstanding of fundamental measurement issues that were prevalent at the time. The accumulation of literature over time led to a growing appreciation of Rasch model methods and subsequently "turned the tide" in favor of contemporary psychometric approaches in medical rehabilitation. Even recently, some authors continue to report and compare summed raw scores and Rasch measures (3, 4). On occasion, the variance explained using summed raw score is greater than that accounted for using Rasch measures, leading some investigators to believe that we need to better understand the Rasch model-derived measures (5).

After a nearly 20 year publication record of Rasch analyses of the FIM instrument, many scientists and clinicians now recognize the limitations of summed raw scores; but, they do not know how to convert FIM item scores into Rasch measures. Researchers who have the knowledge and skills to complete Rasch analyses can help their colleagues by publishing crosswalk tables of raw scores to Rasch measures such as was reported for stroke and other samples (6). Rasch measurement experts could also serve as consultants to clinicians and researchers who do not possess this specialized knowledge. We believe the time is right for journal editors to consider requesting use of contemporary psychometric approaches rather than summed scores, or routinely ask authors to highlight the limitations resulting from assuming summed scores are equal-interval.

While it is beyond the scope of their manuscript, the authors do not address the divergent perspectives of item response theory (IRT) and Rasch models. IRT adherents are apt to view this discussion as irrelevant. Rasch model users are challenged to distinguish models that fit data to models (Rasch) or fit models to data (IRT). Misconceptions about logical positivism as applied to measurement still abound in the 21st century.

*Rating Scale issues*

The results reported by Lundgren Nilsson & Tennant provide us with an opportunity to reflect on the utility of the rating scale used by the FIM instrument. The authors report a relatively large distance between ratings of 6 (independent with equipment) and 7 (complete independence) for several items, particularly eating. This distance does not parallel the experience of clinicians. Using burden of care as an external criteria to validate the rating scale, we would expect the distance between 6 and 7 to be relatively small (7–10). At the lower end of the rating scale range, medical complexity may account for the relatively large distance between ratings of 1 (total assist) and 2 (maximum assist). Their results highlight an important point: Rasch analysis is only part of an instrument validation process. Measures require validation using external criteria. External evidence will enhance the utility and validity of the rating scale.

After reading Lundgren Nilsson & Tennant's report, we are left wondering: What are the clinical implications of the disordered response thresholds? Why would a specific rating scale category be less likely to be used than its "neighbors?" Perhaps some rating scale definitions are unclear? Separating self-care and mobility subscales might help improve the psychometric properties of the rating scale. Perhaps the decision to rate "set-up" reflects nursing or therapist convenience rather than the underlying construct of patient functional status. Investigators who examined minutes of assistance provided to patients found that supervision assistance was not fully reflected in FIM scores (9). We are left wondering whether a 7-point rating scale exceeds nurses' and therapists' ability to reliably distinguish patient functional status on some items. Fewer rating scale options for these items may enhance reliability and reduce rater frustration. It is not clear from their report how rating scale categories were collapsed; information about collapsed categories would help inform these considerations.

*Local and time dependence*

The authors' application of "testlets" to deal with local dependency highlights the redundant content of the FIM instrument and provides an elegant psychometric solution to a vexing problem. They demonstrate convincingly that misfit of the FIM instrument motor items reflects local dependence. The relationship between rating scale use and item fit is illustrated well by their report. We agree that 7-point rating scale needs reconsideration. Developing robust rating scale categories that can be applied consistently across assessment intervals is a pressing need. The consistency of the rating scale structure across impairment groups deserves attention, too.

Lundgren Nilsson & Tennant do not describe the timing of the FIM instrument assessments and how they managed multiple assessments, if any. A primary goal of FIM instrument use is to monitor patient progress during medical rehabilitation and to anticipate resource needs. Thus, clinicians often complete multiple assessments during the course of a stay, and one is left wondering how to deal with dependency across assessments. Mallinson (11) provides an elegant approach to resolving this issue.

*Quibbles*

We have several minor concerns with the authors' data analysis and reporting that require clarification. The authors completed analyses on 340 stroke cases; we assume that there was one observation per patient. It is not clear when the assessments


 doi: 10.2340/16501977-0878
 ISSN 1650-1977

were completed. If all assessments were at admission or discharge, the distribution of responses across the rating scale could vary and lead to different conclusions. How were items with two modalities handled? It appears that the authors ignored distinctions between ratings of walking/wheelchair and tub/shower transfer. Use of modality-specific items may have led to different results.

How many raters were involved and what were their qualifications? It would be interesting to examine the magnitude of rater bias and fit. Some studies have found greater variability in item difficulty or location across raters than items. Given the multinational collaboration and multiple languages that were involved in the collection of these data, the opportunities for rater effects is magnified. Details about the four testlets would be helpful. How were items combined? How were rating scales collapsed?

The authors state that item deletion should be a "last resort." Why? Their admonition may reflect a desire to maintain the integrity of the item set for cross-study comparison or because of the perceived clinical utility of using all the items. But, if the goal is to estimate the location of patients on an underlying motor function continuum, there is nothing sacred about items. One should retain the best items. Often brevity of assessments is valued so that respondent burden, for patients and clinicians, is minimized. Thus, retaining as few items as possible while preserving adequate reliability is a desirable goal. We are reminded of Benjamin Wright's admonition (12) to keep in mind the purpose of measurement. Be clear whether one's purpose is to monitor functional recovery, plan resource use, or provide input to quality indicators. The purpose of data collection should guide one's item scoring and retention strategy.

Upon reflection, we are left wondering, "So what?" How much do the disordered thresholds, item misfit and deleted items matter? It would be helpful to compare person measures obtained from each of the analytic approaches. We suspect that the effects are relatively small.

*Concluding thoughts*

The contributions of Ben Wright and Mike Linacre to measurement in medical rehabilitation cannot be underestimated. Their early willingness to collaborate with Chicago-area colleagues at Marianjoy Rehabilitation Hospital and the Rehabilitation Institute of Chicago reflects their generosity of spirit and enthusiasm in bringing contemporary measurement approaches to healthcare generally and medical rehabilitation, in particular. Ben Wright's hosting of annual measurement conferences at the University of Chicago was critical in fostering communication and collegial relationships between clinicians and scientists in Europe, North America, Australia, and Asia. The National Institute on Disability and Rehabilitation Research and the Centers for Disease Control and Prevention's funding of early FIM instrument development and analysis were critical in propelling subsequent instrument developments. We are fortunate to have been part of the history of these developments and for the friendships we developed with many of the cited authors.

In summary, the authors provide a valuable historic review of Rasch measurement approaches as applied to the FIM instrument; these approaches parallel the developments of Rasch model applications generally. Rehabilitation clinicians and scientists are beneficiaries of their report. This review illustrates that the FIM instrument is a particularly challenging instrument to analyze given the seemingly simple – but actually quite complex – considerations in assigning a rating to an observed patient behavior. However, our challenge is no longer persuading clinicians and researchers to apply Rasch model methods to rehabilitation instruments rather than using an "add 'em up" Classical Test Theory approach. Instead, the imperative is to maintain a dialogue with IRT adherents so that the Rasch model is not relegated to special situations involving small samples.

The worldwide use of the FIM reflects the relevance of the component items to medical rehabilitation. Kudos to Dorothea Barthel (13) who described the motor item set on which the American Congress of Rehabilitation Medicine – American Academy of Physical Medicine and Rehabilitation Task Force developed the FIM items and rating scale. These approaches can be applied to new instruments under development for post-acute care in the United States (CARE Tool) (14) and elsewhere. The legacy of Georg Rasch and Ben Wright's collaboration and the enthusiasm of their students are reflected in the eloquent summary of Rasch measurement developments since the first publication in 1994 (15). We look forward to reading about and participating in additional developments and consequent insights into human performance with applications of the Rasch model.

## ACKNOWLEDGMENTS

Funding for this work was provided by the Rehabilitation Research and Training Center on Improving Measurement of Medical Rehabilitation Outcomes (H133B090024), awarded to the Rehabilitation Institute of Chicago by the National Institute on Disability and Rehabilitation Research.

## REFERENCES

1. 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.
2. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989; 70: 308–312.
3. Deutsch A, Granger CV, Fiedler RC, DeJong G, Kane RL, Ottenbacher KJ, et al. Outcomes and reimbursement of inpatient rehabilitation facilities and subacute rehabilitation programs for Medicare beneficiaries with hip fracture. Med Care 2005; 43: 892–901.
4. Stineman MG, Hamilton BB, Goin JE, Granger CV, Fiedler RC. Functional gain and length of stay for major rehabilitation impairment categories. Patterns revealed by function related groups. Am J Phys Med Rehabil 1996; 75: 68–78.
5. Linacre JM. Variance in data explained by Rasch measures. Rasch Measurement Transactions 2008; 22: 1162.
6. Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger CV. Measurement characteristics of the Functional Independence Measure. Topics in Stroke Rehabilitation 1994; 1: 1–15.
7. Granger CV, Cotter AC, Hamilton BB, Fiedler RC. Functional assessment scales: a study of persons after stroke. Arch Phys Med Rehabil 1993; 74: 133–138.

8. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: a study of persons with multiple sclerosis. Arch Phys Med Rehabil 1990; 71: 870–875.
9. Granger CV, Divan N, Fiedler RC. Functional assessment scales. A study of persons after traumatic brain injury. Am J Phys Med Rehabil 1995; 74: 107–113.
10. Hamilton BB, Deutsch A, Russell C, Fiedler RC, Granger CV. Relation of disability costs to function: spinal cord injury. Arch Phys Med Rehabil 1999; 80: 385–391.
11. Mallinson T. Rasch analysis of repeated measures. Rasch Measurement Transactions 2010; 24: 1317.
12. Wright BD. Measurement for social science and education: A history of social science measurement. MESA Memo 62 [http://www.rasch.org/memo62.htm. Accessed August 29, 2011.
13. Mahoney FI, Barthel DW. Functional Evaluation: The Barthel Index. Md State Med J 1965; 14: 61–65.
14. Gage B, Stineman M, Deutsch A, Mallinson T, Heinemann A, Bernard S, et al. Perspectives on the state-of-the-science in rehabilitation medicine and its implications for Medicare postacute care policies. Arch Phys Med Rehabil 2007; 88: 1737–1739.
15. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1994; 75: 127–132.

Submitted August 30, 2011; accepted August 31, 2011

*Allen W. Heinemann* and Anne Deutsch*
From the Center for Rehabilitation Outcomes Research, Rehabilitation Institute of Chicago and Department of Physical Medicine and Rehabilitation, Feinberg School of Medicine, Northwestern University, Chicago, USA
*E-mail: a-heinemann@northwestern.edu