J Rehabil Med 2012; 44: 97–98

EDITORIAL

The use of raw scores from ordinal scales: Time to end malpractice?

Some 25 years ago, Forrest & Andersen (1) in the *British Medical Journal* reminded readers that descriptive statistical measures such as the mean and standard deviations, which require addition of observations, are invalid whenever data are on ordinal scales. It must be recalled that in any cumulative questionnaire, of the type predominant in many rehabilitation assessments and in Patient Reported Outcome Measures (PROM), the scores are bound to an artificial range (e.g. 0–100), thus causing distortion of intervals towards the margins. Consequently, movement across the margins of the range may understate the real metric increase, while movement across the centre may overstate the increase. Thus the increase in an observed raw score will reflect a different increase in the underlying metric range, depending upon the starting point. This is independent of any floor or ceiling effect, which may emerge, reflective of the lack of validity of the scale for the sample.

The practice of misusing ordinal scales as though they were interval measures was re-emphasized by Merbitz and colleagues (2) in their seminal paper "Ordinal scales and foundations of misinference" . They reminded readers that while ordinal numbers may be put on a number line with assumed equal intervals, in practice the intervals are, by definition, unknown. In practice they represent counts of observed events that are transfers from one category to the next and summed across different variables, yet the actual interval across the category thresholds remain to be determined (hence the distinction between "ordinal" and true "interval" measures). They went on to state that if ordinal scales are manipulated mathematically, the results are not logically valid, and conclusions may therefore be misleading. They concluded that readers should not permit the lack of a complete interval or ratio level functional outcome scale to make the practice of misinference socially acceptable. Subsequently, in this journal and elsewhere, the same message has been repeated on numerous occasions, including evidence of drawing the wrong conclusions in clinical trials where parametric analyses have been applied inappropriately (3, 4).

That ordinal scores are still widely subjected to mathematical operations such as the calculation of means, change scores or effect size, should be of increasing concern to the rehabilitation community. While it may sometimes be necessary to publish such data when comparison with previously reported data is needed, strong caveats should be expressed when doing so. However, today the science of measurement has rendered obsolete the "lack of a complete interval" scaled functional outcome measurement. Rasch modelling of ordinal data, given certain conditions and within a probabilistic framework, allows for the transformation of ordinal raw scores into interval scale measures (5, 6). This journal and many others have seen the growth of the application of the Rasch measurement model in a process widely known as Rasch analysis. When data from an assessment or PROM are shown to fit the model, they concord with the rules or general axioms of measurement laid down by Luce & Tukey (7) almost 50 years ago, and shared by measurement theory in the physical sciences. Thus the Rasch model properties of invariance of comparisons (a given difference meaning the same interval at whatever level of the variable) and sufficiency (the total score implies a predictable score on each of the constituent items, and is all that is required) comply with those axioms, and the latent estimate so derived is an interval scale. The Rasch model is the only model within the item response theory (IRT) family of models that so complies with these axioms (references in 8). We also refer readers to the Guidelines for reporting studies using Rasch analysis (www.medicaljournals.se/jrm; www.jampress.org) and the references listed therein.

Readers may be aware of the frequently expressed complaint that the "data never fit the Rasch model". Indeed, the Rasch model is especially demanding of data in its quest to satisfy the rules for constructing measurement, and this has led others to adopt alternative models within the IRT framework. As attractive as these models may appear (where the focus is on explaining the variance in the data, not on constructing measurement) it must be clear that they are incompatible with the construction of fundamental measurement, and that the estimates so derived are not interval scale measures (9, 10).

The measurement of individual change is the essence of outcome in rehabilitation (and in all behavioural sciences). Readers may be familiar with reports of minimal real (or detectable) difference (MRD), giving the minimal change beyond the level expected by chance (11). The MRD is the premise, but not a substitute, for the measurement of the minimal (clinically) important difference (MID), which is a more controversial concept (12). Once it is understood that ordinal scales are non-linear, it becomes clear that a MRD or an MID along an ordinal scale should be managed cautiously. In fact, the same numeric increment may mean a different substantial change depending on the baseline value. Consequently, misinference may follow. This understanding could be extended to classification and/or payment systems based on "points gained", and to the whole field of outcome assessment research. Raw scores can maximize the functional gain after rehabilitation if the baseline levels are far from the floor and the ceiling of the scale where score changes are inflated relative to those at the margins. Conversely, as patients approach the top of the scale, each raw score point represents an increasing metric distance, yet it appears that the patient is "slowing down" in his or her recovery, because it becomes increasingly difficult to gain further raw score points. It is at this time that it may be concluded that "the patient has plateaued". Unless this is consistent with clinical judgement, it is unlikely to be correct.

Thus we consider it of importance to encourage researchers to use Rasch analysis and Rasch-derived instruments, both

in the development and evaluation of instruments, and in the analysis of data from ordinal scales in outcome research. Increased use of Rasch-derived instruments could be achieved both by training in scientific methodology and by efforts during the review process by Editors and Editorial Boards of scientific journals in the field. In achieving these standards, it is important that Rasch-derived instruments should be perceived to be user-friendly by all concerned. To promote this, the raw-score to linear measure (and surrounding error) conversion tables should be provided, thus allowing use of the original raw scores in everyday clinical practice. This can then be converted to interval scaling whenever required simply by consulting the conversion table (or, for example, by creating a look-up routine in Excel). Where there are missing data, internet-based algorithms should be made available to provide interval scale estimates based on the information available (e.g. see www.rehab-scales.org). In part this reflects some of the most recent developments, as applied to health outcomes by modern test theory, which is the selection of a sub-set of items, where the choice of items are targeted to the subject's level of ability, out of a larger set spanning a much wider range of difficulty levels. This method, called computerized adaptive testing (CAT), allows shorter tests and greater precision (13).

In many respects this is an exciting time for the development of outcomes in rehabilitation, as new techniques have become available that enhance our understanding of how assessments and PROMs work. They facilitate the construction of fundamental measurement from such scales, a type of measurement previously found largely only in the natural sciences. Perhaps it is time for the rehabilitation community to unite in stating enough is enough; that it is time to end the ordinal misrule. Journal Editors and reviewers should now consider requesting full justification from authors for performing mathematical operations on any ordinal scales, be they change scores or effect sizes, or for subjecting such scales to parametric statistics. However, we are aware that such a change in practice concerning ordinal scales is likely to be incremental but, even so, these efforts have to be encouraged. Authors who report IRT-based latent estimates should fully justify that their estimates are at the interval scale level (if such claims are made, or mathematical operations performed), referencing the mathematical proofs to support their assertion. In this way, rehabilitation can lead the way in improving the science of outcome measurement, much as it did over 20 years ago when these issues were first highlighted in the *Archives of Physical Medicine and Rehabilitation* (2).

REFERENCES

1. Forrest M, Andersen B. Ordinal scale and statistics in medical research. BMJ 1986; 292: 537–538.
2. Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989; 70: 308–312.
3. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. J Rehabil Med 2001; 33: 47–48.
4. Kahler E, Rogausch A, Brunner E, Himmel W. A parametric analysis of ordinal quality-of-life data can lead to erroneous results. J Clinical Epidemiol 2008; 61: 475–480.
5. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
6. Van Newby A, Conner GR, Bunderson CV. The Rasch model and additive conjoint measurement. J Appl Meas 2009; 10: 348–354.
7. Luce RD, Tukey JW. Simultaneous conjoint measurement: a new type of fundamental measurement. J Math Psychol 1964; 1: 1–27.
8. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004; 42 Suppl 1: I7–16.
9. Lord FM. The 'ability' scale in Item characteristic curve theory. Psychometrika 1975; 40: 205–217.
10. Karabatos, G. The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. J Applied Meas 2001; 2: 389–423.
11. Tesio L. Outcome measurement in behavioural sciences: a view on how to shift attention from means to individuals and why. Int J Rehabil Res 2012 (in press).
12. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, et al. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol 2010; 63: 28–36.
13. Jette AM, Haley SM. Contemporary measurement techniques for rehabilitation outcomes assessment. J Rehabil Med 2005; 37: 339–435.

Göteborg, Leeds and Milan, December 2011

Gunnar Grimby	*Alan Tennant*	*Luigi Tesio*
Editor-in-Chief	Associate Editor	Member of Editorial Board