Application of Rasch Analysis in the Development and Application of Quality of Life Instruments

Alan Tennant, PhD,¹ Stephen P. McKenna, PhD,² Peter Hagell, PhD³

¹Academic Unit of Musculoskeletal & Rehabilitation Medicine, University of Leeds, Leeds, UK; ²Galen Research, Manchester, UK; ³Department of Nursing, Lund University, Lund, Sweden

ABSTRACT

This paper discusses recent advances that have been made in the field of psychometrics, specifically, the application of Rasch analysis to the instrument development process. It emphasizes the importance of assessing the fundamental scaling properties of an instrument prior to consideration of traditional psychometric indicators. The paper introduces Rasch analysis and shows how it has been applied in the development of needs-based measures in order to ensure that they provide unidimensional measurement. By ensuring that scales are based on the same measurement model and that they fit the Rasch model it is possible for QoL scores to be compared across diseases by means of cocalibration and item banking.

Keywords: classical test theory, differential item functioning, needs-based quality of life, Rasch analysis, unidimensionality.

As long as primitive counts and raw scores are routinely mistaken for measures by our colleagues in social, educational, and health research, there is no hope of their professional activities ever developing into a reliable or useful science [1].

A crucial aspect of the application of the needs model is that it has been allied with the most advanced psychometric methods. This paper argues the case for the use of Rasch analysis to ensure that scales are unidimensional [2], a fundamental requirement of construct validity [3]. The paper gives an overview of the nature of Rasch analysis and shows how it aids valid across-disease comparisons of quality of life (QoL) by means of cocalibration and the development of item banks.

Over the last two decades an analytical approach has been adopted that is pivotal to both judging the quality of existing outcome instruments and in developing new instruments. This approach is called Rasch analysis, after its originator, a Danish mathematician. He developed Poisson models for reading, intelligence, and achievement tests, the last becoming known as the Rasch model [2]. Rasch analysis has been employed in the development of most of the needs-based QoL instruments, to ensure that the resulting scales are unidimensional. Only the earliest developed needs-based measures did not benefit from this approach and studies are underway to ensure that earlier measures are updated to fit the Rasch model [4].

The quest for measurement is an important part of advancing science, and the type of measurement, which allows for arithmetic operations such as addition and subtraction, is known as fundamental measurement [5]. Most outcome measures used in health care are ordinal in nature, precluding such arithmetic operations [6]. Many such measures focus on attributes that are not directly measurable, such as pain, self-esteem, or quality of life. These measures give a “manifest score” of the construct being measured. Consequently, most outcomes are expressed as ordinal manifest scores, indicating some rank on a perceived underlying latent trait. Although there is a substantial body of nonparametric statistics to analyze such information, the importance of the calculation of change scores in clinical trial analysis (and attributes of measurement such as the “effect size” [7]), which require normally distributed interval-level measurement, gives urgency to achieving a quality of measurement that will sustain such arithmetic operations.

In order to achieve such fundamental measurement certain properties are required. These are reviewed in detail elsewhere [8,9] but essentially they are:

- the numerical properties of order (one mark on the ruler represents more or less of the construct than another);
- addition (points on rulers may be added together); and
specific objectivity (the calibration of the ruler (item set or questions) is independent of the persons used to calibrate and vice versa).

Where data fit the Rasch model these properties are confirmed and fundamental measurement follows. On a more formal level, the theory of simultaneous conjoint measurement [10] provides the mechanism for translating manifest to latent scores: Rasch analysis delivers conjoint measurement when data fit the model.

The Rasch model is a unidimensional model that has two main assertions:

1. that the easier the item is, the more likely it will be passed (affirmed); and
2. the more able the patient, the more likely they will pass (affirm) an item (or do a task) compared to a less able patient.

Unidimensionality is a prerequisite to the summation of any set of items [3,11,12]. The Rasch model assumes that the probability of a given patient “passing” an item or task is a logistic function of the relative distance between the item location parameter (the difficulty of the task) and the respondent location parameter (the ability of the patient), and only a function of that difference. Expressed formally, this gives:

\[ p_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \]

where \( p_i(\theta) \) is the probability that patients with ability \( \theta \) will be able to do item (task) \( i \), and \( b_i \) is the item (task) difficulty parameter. The model can be extended to cope with items with more than two response categories. From this, the expected pattern of responses to a set of items or tasks is determined given the estimated \( \theta \) and \( b \). When the observed response pattern coincides with or does not deviate greatly from the expected response pattern, the items fit the measurement model and constitute a true Rasch scale [13]. Various fit statistics determine whether or not the data do fit the model, and these tend to be software dependent, although all work on the principal of looking at the deviation of the observed data from the model expectation. Finally, where there is local independence of items (that is, no residual associations in the data after the Rasch trait has been removed), this, taken together with fit to the model, supports the contention that the scale is unidimensional [14].

Assuming that the data fit, the Rasch model transforms them from ordinal scores into interval level measurement with the logit (log odds unit) as the unit of measurement. A logit is the distance along the line of the variable that increases the odds of observing the event by a factor of 2.718. There is a clear relation between the ability–difficulty difference, and the probability of affirming an item or undertaking a task. For example, where the difference between a patients’ ability and the item or task difficulty is zero, the probability is 0.5. Where the difference is +1 logit (that is, the patient has greater ability—or more of the trait—than expressed by the item) the probability is 0.73, or 0.27 if the difference is −1.0. Where the difference is ±3 logits then the probabilities are 0.95 and 0.05, respectively.

Differential Item Functioning (DIF) can also be examined by fitting data to the Rasch model [15]. Essentially, the scale should work in the same way, irrespective of the group assessed. Thus, the probability of being able to do a task, or affirming an item, for patients at the same level of ability (or, for example, with the same QoL) should remain the same across groups. Assessment of DIF can yield crucial information about the measurement equivalence of an instrument between various cultural groups [16] but should also be applied across gender and age groups within those cultures.

In comparison with classical test theory, the Rasch model provides a means of assessing a range of additional measurement properties, increasing the information available about a scale’s performance [17–19]. The model is one of many used in this way, which are generally subsumed under the rubric of Item Response Theory (IRT) [20,21]. The Rasch model is known as the one-parameter model within this framework, but it has unique properties, which are crucial to attaining conjoint measurement [22], a prerequisite for the calculation of change scores [7].

The Rasch model was readily adopted in rehabilitation in the late 1980s [23], as the language of ability and difficulty easily transferred from education. Patients undergoing rehabilitation have a given level of ability. In order to assess this level they can be presented with a range of tasks requiring differing degrees of ability. Since then the approach has become used with a wide range of clinical and diagnostic groups [24,25]. All recent needs-based quality of life instruments are developed using this approach [26–32].

Given that both patients and items are calibrated on the same underlying metric trait, the potential for innovation in measurement is considerable. Consider for example the current debate about disease-specific and generic QoL measures. Where scales are based on the same theoretical unidimen-
sional construct, items from different diseases can be calibrated on the same scale, given that some items (that are free of DIF by diagnosis) common to both scales are employed. This provides disease-specific and comparative QoL measures by “item banking” items or questions onto the same underlying metric [33–35]. Currently this approach, based on the needs-based model of QoL, is being used to establish an item bank for disease-specific QoL measures in the rheumatic diseases [36–38]. A similar exercise is planned for dermatology and links between these two disease areas could be made possible by means of the Psoriatic Arthritis Quality of Life (PSAQoL) measure [37].

The needs-based QoL measures all have the same theoretical basis, are unidimensional (insofar as their items fit the Rasch model) and have good traditional psychometric properties. Not only do they work as effective outcome measures in clinical trials but they also offer the potential for allowing valid comparisons of QoL to be made across diseases [39] and between healthy and diseased populations [40]. While it has been common practice to use generic health status measures such as the SF-36 to make such comparisons, the results have been both misleading and invalid [39]. This is because, although a question is expressed in the same way for all respondents, different types of patients who have had different experiences interpret it differently. For example, a “yes” response to a question about feeling tired can represent a very different response for a healthy person and one with rheumatoid arthritis. This explains why surprising results are frequently obtained for cross-disease comparisons. For example, data collected with the SF-36 suggest both that individuals with psoriasis have worse scores than patients with arthritis, cancer, and myocardial infarction [41] and that such patients have comparable or even better scores than those experienced by an average population [42].

Summary

Only occasionally do we see concerns raised about inappropriate analysis of data that are erroneously assumed to be at the interval level [6]. The extent to which analyses of reliability, validity, and responsiveness are compromised by ignoring such assumptions is unknown. It is also unknown at present to what extent the misuse of ordinal manifest scores compromises the results of clinical trial analyses when these scores are used to calculate changes across experimental and control groups. However, the potential implications should not be underestimated [43–45].

The ability of a scale to provide fundamental measurement should be established prior to the more commonly reported psychometric attributes. Rasch analysis offers a method of ensuring that key measurement assumptions are tested and, where data fit the model, arithmetic operations may be undertaken. It has particular value in the development of new measures, specifically in guiding item reduction. Traditional methods of item reduction that rely on item–total correlations and/or indices of internal consistency can have unfortunate effects on the sensitivity of measures and their ability to provide valid scores at the extremes of the construct range. This is because items at the extreme of the measurement range are generally discarded because too many or too few respondents affirm them. In reality, these “extreme” items may be the most important in a scale—extending its range of coverage of the construct.

References


44 Wright BJ, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989;70: 857–60.