RATING SCALES FOR NEUROLOGISTS

J Hobart

J Neurol Neurosurg Psychiatry 2003;**74**(Suppl IV):iv22–iv26

A neurologist once told me that he found the subject of rating scales ''exceedingly dull'', while another found the area ''abstruse''. I have therefore attempted to produce an overview that is helpful and conveys some of the basic principals underlying outcomes measurement and rating scales. Clinicians must realise that because this is an alien and somewhat ''dry'' area, they may need to invest some time to appreciate the issues. Instead of discussing specific scales or rating scales for rehabilitation, which will only be relevant to a limited audience, I have chosen to discuss the importance of rating scales and how to achieve high quality measurement. I hope this makes the text more widely applicable to the neurological community.

The take home message is simple; neurologists need to take their rating scales very seriously.

WHY ARE RATING SCALES IMPORTANT?

Rating scales are important because they are a method of measurement. Measurement is important because inferences are based on it.[1] For example, in clinical trials we measure variables (for example, disability), perform statistical tests on the numbers generated by scales, and base conclusions on the results. These conclusions influence patient care, prescribing, policy making, and the expenditure of public funds. Thus, the validity of inferences from clinical trials is directly dependent on the quality of the measurement instruments used. Some measurements are clear cut—for example, mortality rates. However, measurement becomes complex for more abstract, ill defined, ''soft'' outcomes such as patient's perspectives of the impact of disease and their quality of life. If we are serious about using these abstract variables to evaluate clinical practice we must be serious about our attempts to measure them as rigorously as possible.

Consider clinical trials of interferons and glatiramer acetate in multiple sclerosis (MS). These trials have produced interesting results: an incontrovertible reduction in relapse rate and accumulation of magnetic resonance image (MRI) lesions over time, but a debatable reduction in the progression of physical disability. These findings have prompted major developments including: research to understand the seemingly complex relation between pathology and disability; calls for definitive studies that are free from pharmaceutical conflicts of interest; a controversial review by the National Institute for Clinical Excellence; and the UK risk sharing scheme for prescribing disease modifying therapies.

Despite these major developments, few have questioned seriously how the choice of rating scale may have influenced this course of events. This is highly relevant because these major developments in MS are effectively based on the assumption that Kurtzke's expanded disability scale (EDSS), the rating scale used in most MS treatment trials, was considered adequate enough to handle the task of measuring disability and detecting clinically significant change when it occurs. Unfortunately, data concerning the measurement properties of the EDSS do not give us that confidence. First, the EDSS mixes the measurement of different health domains—that is, impairment in the early part of the scale, mobility in the mid range of the scale, and bulbar function in the upper part of the scale. As such, the EDSS is not a pure disability measure. While this may not seem such a big deal it is akin to having a scale that measures length at one end, weight in the middle, and volume at the end. Second, the EDSS generates ordinal scores rather than interval measures. More about that later. Third, the EDSS has been proven less able than other ''disability'' scales to differentiate between individuals at one point in time and detect change in disability over time. These facts undermine the validity of inferences made on the basis of the analysis of EDSS scores. Consequently, we are at risk of making inaccurate inferences about disability in MS every time we use the EDSS.

The use of different statistical methods to analyse results acquired from rating scales cannot overcome flaws within the rating scale itself.

Correspondence to:
Dr Jeremy Hobart, Department of Clinical Neurosciences, Peninsula Medical School, Derriford Hospital, Plymouth PL6 8DH, UK;
Jeremy.Hobart@phnt.swest.nhs.uk

DOES THE CHOICE OF RATING SCALE REALLY MAKE THAT MUCH DIFFERENCE?

The above discussion labours the point that inferences from studies are dependent on the quality of the rating scales used. However, surprisingly few studies have taken the next step to determine the implications for clinical trials of the choice of rating scale. This supports the suggestion that clinicians poorly appreciate the limitations of rating scales, perhaps because measurement in laboratory disciplines presents few inherent difficulties.

Treatment studies in MS provide illustrations that the validity of inferences made from all clinical studies is dependent on the quality of the measurement instruments used:

> Cohen *et al*[2] compared the EDSS with the MS Functional Composite in a pivotal study of interferons. We compared six ''disability'' measures, including the EDSS, in steroid treatment for MS relapses.[3] Both studies demonstrated that the statistical and clinical significance of the results and, therefore, inferences made about the treatment effectiveness, depended on which scale was used.

Some authors have played down the importance of rating scales in clinical trials suggesting that trial design, in particular randomisation and blinding, is more important. Maximising trial design will not overcome the problems cause by weak scales, and vice versa. Attention to rigor is needed in both arenas.

WHAT TYPES OF SCALES ARE THERE?

Tables 1 and 2 show two rating scales, the Ashworth scale for measuring spasticity, and the MS walking scale for measuring patients' perceptions of the impact of MS on walking ability (MSWS-12). The two scales are very different. The Ashworth is an example of a single item scale. It considers spasticity as a continuum on which each of the five defined levels has a specific meaning (for example, 1 = ''catch''). Other examples of single item scales are the EDSS and Rankin scale. In contrast, the MSWS-12 is an example of a multi-item scale. It has 12 questions each with a range of response options, and scores are summed across items to generate a summed or total score. Walking ability is therefore measured on a continuum with 48 levels (12–60).

The theory underpinning multi-item scales is that when we are attempting to measure complex clinically relevant domains (for example, disability and quality of life) a single item is unlikely to represent well the broad scope of that domain. In addition, each level of the Ashworth scale is open to individual variation of interpretation (that is, random error). While each item of a multi-item scale contributes unique information, it is impractical clinically and analytically to allow each item to act as a rating scale. Consequently we seek to combine items to allow what they share in common to dominate the ways in which they differ. Furthermore, combining across items cancels out the unavoidable random error associated with each single item, hence reliability is often high. It goes without saying that we must prove it is appropriate to combine a set of items to generate a total score. This is rarely done. More about that later.

Single and multi-item scales contrast in their interpretability and scientific rigor. Single item measures are easy to interpret as each level determines a specific meaning. This is very meaningful clinically. For example, an EDSS of 6.5 means a person can walk about 20 m using bilateral assistance. In contrast, multi-item measures are less interpretable clinically as any person's score represents the sum of the item scores and any (except min and max) sum can be achieved by a variety of permutations. From a clinical perspective this creates problems with interpretation as a value of say 54 is somewhat intangible. It is, therefore, entirely understandable why clinicians find single item measures more meaningful and therefore lean towards them.

Single item scales are weak measures. They have poor reliability, validity, and limited ability to detect differences between individuals at one point in time and detect change

Table 1 Ashworth scale of spasticity

0	No increase in tone
1	Slight increase in tone giving a ''catch'' when the limb is moved in flexion or extension
[1+]	Slight increase in tone, manifested by a catch, followed by minimal resistance throughout the remainder (less than half) of the range of movement]
2	More pronounced increase in tone but the limb easily flexed
3	Considerable increase in tone, passive movement difficult

Table 2 The multiple sclerosis walking scale (MSWS-12)

- **These questions ask about limitations to your walking due to MS during the past two weeks**
- **For each statement, please circle the one number that best describes your degree of limitation**
- **Please answer all questions even if some seem rather similar to others, or seem irrelevant to you.**
- **If you cannot walk at all, please tick this box** □

In the past two weeks, how much has your MS:	Not at all	A little	Moderately	Quite a bit	Extremely
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
3. Limited your ability to climb up and down stairs?	1	2	3	4	5
4. Made standing when doing things more difficult?	1	2	3	4	5
5. Limited your balance when standing or walking?	1	2	3	4	5
6. Limited how far you are able to walk?	1	2	3	4	5
7. Increased the effort needed for you to walk?	1	2	3	4	5
8. Made it necessary for you to use support when walking indoors (e.g. holding on to furniture, using a stick, etc)?	1	2	3	4	5
9. Made it necessary for you to use support when walking outdoors (e.g. using a stick, a frame, etc)?	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

over time. A couple of analogies may help to explain this situation that some clinicians find paradoxical. Consider the introduction of a compulsory multiple choice examination for neurology trainees! The aim is to measure examinee level of neurological knowledge. If the exam has one question the results will be heavily influenced by the examinees' knowledge of that specific topic area rather than their overall neurology knowledge. The more questions asked, and aggregated, the better ''measure'' we get of that person's knowledge—provided the questions have a reasonable coverage of the subject mater. Another analogy is the Barclaycard Premiership. The league seeks to determine the best football team in the land, so it is a measure of ability. This season Manchester United (finished top) drew 1–1 with Sunderland (finished bottom). That single game was not a reliable or valid indicator of the relative difference in footballing ability of the two teams. From these analogies it is hopefully easier to appreciate why single item scales are likely to be unreliable (subject to random error) and poorly valid (a limited indicator of neurology knowledge). Can we afford these scientific weaknesses in our clinical trials?

HOW DO I CHOOSE THE BEST SCALE FOR MY PURPOSE?

Clinicians often have to choose one scale from among many potential candidates. Unfortunately, no one scale exhibits all desirable qualities, different scales have different virtues, and scales that are useful for one situation may not be useful for others. Therefore, a scale must be selected for a particular purpose. To do this scale users must be able to choose measures intelligently based on their needs.

Rating scales must be clinically useful and scientifically sound. Clinical usefulness refers to the successful incorporation of an instrument into clinical practice and its appropriateness to the study sample. Scientific soundness refers to the demonstration of reliable, valid, and responsive measurement of the outcome of interest. Clinical usefulness does not guarantee scientific soundness, and vice versa.

I will concede that diatribes on reliability and validity testing are dull. There are also many publications on evaluating psychometric properties and these are regularly updated as the field moves forward.[4] Here, then, I simply make a few key statements.

- Explore beyond the title of a scale. For example, consider the Rankin scale which is called a handicap measure. It seems curious that the six levels mention symptoms (0 = no symptoms) and disability (1 = slight disability; 2 = mild disability; 3 = moderate disability; 4 = moderately severe disability; 5 = severe disability) but not handicap.
- Be very clear about what you want to measure. There is a current vogue to use ''quality of life'' as a primary outcome for clinical trials. But there are many definitions of quality of life. Also, quality of life may not be the most appropriate variable to measure. The more distant the outcome chosen is in relation to the aim of the intervention, the greater the chance of confounding. For example, hip replacement is often undertaken to relieve pain. Should we be disappointed, or critical, if the effect on psychosocial functioning is far less dramatic
- Studying the distribution of scale scores in samples is simple and a very valuable basic test to determine whether a scale will be useful in that sample. Although this does not provide evidence for reliability and validity per se, targeting of scales to samples is important as ceiling and floor effects (percent scoring maximum and minimum possible scores) represent sub-samples whose scores cannot and may not change regardless of the effects of the intervention. This simplest of analyses is rarely undertaken.
- Reliability, validity, and responsiveness are, to a large extent, independent psychometric properties. Therefore, they must all be undertaken. There is little value in studying a single property alone even though this is more common than full psychometric evaluations.
- Reliability, validity, and responsiveness are sample dependent properties. Hence it is important to study scales in different samples. This is particularly important for generic scales; these are scales that can be used in a wide range of disorders. For example, the medical outcomes study short form 36-item health survey (SF-36) is the most widely used health status measure across the world. It is therefore tempting to use it. However, evidence demonstrates important limitations as an outcome measure for clinical trials in MS, stroke, and amyotrophic lateral sclerosis/motor neurone disease.
- One of the best tests of validity is the development method of a scale. If recognised techniques of rating scale construction were used the chances of good reliability and validity are high.
- Using a scale in clinical practice or a study will usually provide enough information to make statements about its reliability and validity even though this may not be, or was not intended to be, a psychometric study. Although, obviously, psychometric properties should be tested and demonstrated before a scale is used, this retrospective approach, which enables clinicians to support or refute some of the inferences they make, is better than nothing.

HOW ARE SCALES DEVELOPED?

Developing rating scales is a labour intensive process requiring considerable expertise in health measurement. Therefore, it is advisable to carefully evaluate existing measures before abandoning them. The psychometric properties of available measures can be determined more quickly. Here is an overview of instrument development. Fuller accounts can be found elsewhere.

Multi-item scale development can be considered to have four stages. First, define what you want to measure, which in measurement speak is the construct, and any potential subdivisions of it (the sub-constructs). Second, generate a pool of items so that all important issues are considered for inclusion in the final scale. Third, administer the item pool to a sample of patients and, from the analysis of the resulting data, develop a scale(s) that are reliable and valid representations of the construct. Finally, examine the full properties of the scales in independent samples.

ARE TOTAL SCORES GENERATED BY SUMMING ITEM SCORES REALLY GOOD MEASURES?

The answer here is yes and no. It depends on the definition of measurement being used, and the goals that we are trying to achieve. This issue is becoming very important, and therefore it is appropriate to consider it. However, things do start to get a bit more complex from here on in.

If we make the assumption that measurement is quantification of a variable, and that variables can go from ''less of'' to ''more of'', then it is reasonable to consider the total score generated by adding up a set of items is a measure of that variable *provided* that we have some way of demonstrating that the items address the same underlying construct. This is the basic theory that underpins multi-item rating scales and was discussed earlier.

Consider the MSWS-12. Our aim was to measure the impact of MS on walking ability. The variable (construct) we wished to measure was walking ability. By interviewing patients and clinicians we got a set of statements on how MS affected walking ability. When redundant statements were removed we were left with n = 12. A response option was written so that the impact of MS on each item could be graded. This potential scale was sent to a large group of people with MS and the resulting data analysed to determine if it was appropriate to combine the scores of the 12 items to generate a total score and if the total score was reliable (reproducible) and valid (evidence that it was an indicator of walking ability). Evidence for this is presented in the development paper.

While this all seems reasonable, we cannot get away from the fact that summed scores make a series of assumptions that do not hold. First, the response categories for each item are given sequential integers (1, 2, 3, 4, 5). This assumes equal differences between the different levels. This is not the case, logically or empirically. Second, ''quite a bit'' is assumed to be more than ''moderately''. There is evidence that a substantial proportion of the population think ''moderately'' is more than ''quite a bit''. Third, the use of total scores assumes that given differences have equal meaning. That is, a score difference of 10 points has the same meaning across the scale range. There are clear demonstrations that this is not true.

Over the last few hundred years mathematicians, physicists, psychologists, measurement theorists, philosophers, and others have articulated what they mean by measurement. It has been defined that measurement in the physical sciences, termed fundamental measurement by the physicist Norman Campbell, has five main characteristics: unidimensionality, linearity, sample independence, scale independence, and invariance. Consider a ruler for measuring height. The ruler describes only one attribute (unidimensionality), which it measures on a linear continuum (that is, the differences between the calibrations are equal). The ability of a ruler to measure height is not seriously affected by the people being measured (sample independent). It does not matter which ruler is used to measure height (scale independent). The process of measurement remains the same at different areas of the continuum (invariance). Campbell suggested that measurement in the social sciences (effectively anybody using rating scales as measurement instruments) could not be called a science until it achieved these characteristics.

Now when we use a rating scale, and sum the item scores to get a total score, it is difficult to be certain that we are measuring a single construct. We have not proven that the distance between units is stable. We know the properties of scales are sample dependent and the measurement of people is scale dependent. In short, we have not and cannot achieve measurement as defined by others, and certainly not the type of measurement achieved in the physical sciences. When we think about it further, rating scales are merely counts of discrete events. But this is the only format in which we can get such data and thus it is what we have to work with. It is clear then that something must be done to rating scale data before we can consider total scores as measures that satisfy the characteristics stated by measurement theorists across the years.

HOW DO WE ACHIEVE LINEAR MEASURES FROM SUMMED SCORES?

This brings us into the domain of new psychometric methods—Rasch analysis (RA)[5] and item response theory (IRT).[6] There are statistical techniques that can be applied to rating scale data. They attempt to transform ordinal scores, that are scale dependent and of limited accuracy, into interval measures that are scale independent and suitably accurate for individual patient assessment. In essence, these methods model the probability of an individual's response to an item. They are based on a logical assumption: individuals with high levels of whatever is being measured (for example, physical function) should have an increased probability, relative to individuals with low levels, of getting a better score on any item (for example, dressing). Technically this gets very complex but it is important to consider the clinical benefits.

There is a huge potential for new psychometric methods to change the face of health outcomes measurement. Using linear measures instead of non-linear raw scores would give a true reflection of disease impact, differences between individuals and groups, and treatment effects. The value of this is highlighted by studies demonstrating that raw score changes underestimate interval level change by up to ninefold.[7] Improved accuracy would enable individual patient assessment. The ability to generate interval measures that are independent of the rating scale used enables scales measuring the same health construct to be equated on the same linear ruler. This is the basis for comparisons of studies, meta-analyses, and systematic reviews. Moreover, the process of scale equating generates a pool of commonly calibrated items, an item bank. Item banks are flexible measurement methods because any subset of items can be selected from the bank to generate an accurate score. Therefore, investigators are no longer wedded to defined scales and can simply select the most appropriate group of items for their study. Alternatively, of course, they could choose a defined scale if they wish.

The availability of item banks opens the way for the most exciting development in health measurement, computerised administration of rating scales (computer adaptive testing). Here, a computer uses the response to an item to determine the next item presented to the respondent. As a result, the optimum items for any individual are identified thus providing rapid, efficient, user friendly, and precise individual person measurement. Computer adaptive testing offers the opportunity to bring patient based outcome measurement into routine clinical practice and influence decision making for individual patients. Currently this does not happen.

The last few years has seen the application of new psychometric methods. Most studies have analysed existing scales. However, there is evidence that health measures can be successfully equated. Computer adaptive administration of a calibrated item pool for the impact of headache has been shown to generate rapid (five items or less) person measurement. These measurements are as precise as those generated by the entire item pool (54 items) and suitable for individual patient clinical decision making. Given the clinical potential of new psychometric methods it is curious that they are not more widely available. There are a number of possible explanations for this. First, the area is complex which naturally attracts scepticism and is off-putting. Complexity can lead to confusion and misunderstandings (see later). Second, PC based software for

undertaking Rasch and IRT analyses have only become available in the last few years.

Perhaps the most important fact impeding progress in the field of new psychometric methods is misunderstandings about the similarities and differences between RA and IRT. The two statistical methods are consistently considered as members of the same family, and usually termed IRT. This is probably because of their theoretical and mathematical similarities. However, RA and IRT differ at the most fundamental level—the philosophy underpinning their development.[8] The Rasch model is a definition of measurement, a mathematical derivation from the requirement that stable linear measures be constructed from the ordered qualities of rating scale data. Therefore, the aim of a Rasch analysis is to determine the extent to which observed rating scale data satisfy this stringent definition. Stable linear measures can be constructed only when the data satisfy the model. Therefore, we seek data that fit the model. In stark contrast, IRT models were developed to explain data. Therefore, the aim of an IRT analysis is to seek a model that fits the data. In his recent article, Massof compares and contrasts RA and IRT, explaining this fundamental difference in detail, demonstrating the limitations of IRT, and the importance of Rasch.[9] Massof demonstrates, empirically, that only the Rasch model enables investigators to achieve measurement, as described by measurement theorists, from rating scale data. He demonstrates that IRT models are not valid measurement models.

It seems surprising that this fundamental difference between RA and IRT is not highlighted in any of the articles in a recent supplement of *Medical Care* devoted to ''IRT''. The two methods represent different paradigms with different research agendas. They are, therefore, incompatible.[8]

CONCLUSION

Developments in basic neuroscience are generating new treatments that need to be evaluated and compared. The emphasis is that these evaluations be done from the patient's perspectives. Unless high quality rating scales are available we run the risk of making inaccurate inferences from clinical trials. However, this challenge is not as daunting as it may appear because techniques are available to achieve, from rating scale data, the type of measurement taken for granted in the basic sciences. It is time that clinicians recognised that fact, insisted on better measures, and encourage investment in measurement research.

ACKNOWLEDGEMENTS

I am grateful to Professors Alan Thompson (Institute of Neurology) and Benjamin Wright (University of Chicago) for their input to my work in this area. Some of this work was supported by the NHS Health Technology Assessment Programme, but the views and opinions expressed do not necessarily reflect those of the NHS Executive.

REFERENCES

1 **Bond T**, Fox C. *Applying the Rasch model: fundamental measurement for the human sciences*. New York: Lawrence Erlbaum Associates, 2001.
► **An important text for those interested in Rasch technology, although it is from the perspective of educational and psychological measurement.**
2 **Cohen JA**, Cutter GR, Fischer JS, *et al* for the IMPACT Investigators. Use of the multiple sclerosis functional composite as an outcome measure in a phase 3 trial. *Clinical Trial Arch Neurol* 2001;**58**:961–7.
3 **Hobart JC**, Riazi A, Lamping DL, *et al*. Measuring the impact of MS on walking ability: the 12-item MS walking scale (MSWS-12). *Neurology* 2003;**60**:31–6.
► **The paper outlining the development of the MSWS-12.**
4 **Scientific Advisory Committee of the Medical Outcomes Trust**. Assessing health status and quality of life instruments: attributes and review criteria. *Quality of Life Research* 2002;**11**:193–205.
► **The most up to date guidelines for evaluating rating scales.**
5 **Rasch G**. *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press, 1960.
6 **Lord FM**, Novick MR. *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley, 1968.
► **References 38 and 39 in this paper are the primary texts for Rasch analysis and item response theory—not for the faint hearted.**
7 **Wright BD**, Linacre JM. Observations are always ordinal: measurements, however, must be interval. *Arch Phys Med Rehabil* 1989;**70**:857–60.
► **A clear explanation of why summed scores will not do.**
8 **Andrich D**. Controversy and the Rasch model: a clash of two paradigms. *Medical Care* (in press).
► **Essential reading for anyone interested in Rasch analysis and item response theory. It explains the similarities and differences and the basis of a long standing ongoing controversy.**
9 **Massof R**. The measurement of vision disability. *Optometry and Vision Science* 2002;**79**:516–52.
► **An exceptional documentation of the history of rating scales and a fine demonstration of the limitations of traditional psychometric methods, advantages of new psychometric methods, and differences between Rasch technology and item response theory.**