Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations

Jeremy C Hobart, Stefan J Cano, John P Zajicek, Alan J Thompson

Lancet Neurol 2007; 6: 1094–105

Neurological Outcome Measures Unit, Peninsula College of Medicine and Dentistry, Plymouth, Devon, UK (J C Hobart FRCP, S J Cano PhD, J P Zajicek FRCP); **Institute of Neurology, London, UK** (A J Thompson FRCP, J C Hobart, S J Cano)

Correspondence to: Jeremy Hobart, Peninsula College of Medicine and Dentistry, Room N16, ITTC Building, Tamar Science Park, Davy Road, Plymouth PL6 8BX, UK **jeremy.hobart@pms.ac.uk**

Have state-of-the-art clinical trials failed to deliver treatments for neurodegenerative diseases because of shortcomings in the rating scales used? This Review assesses two methodological limitations of rating scales that might help to answer this question. First, the numbers generated by most rating scales do not satisfy the criteria for rigorous measurements. Second, we do not really know which variables most rating scales measure. We use clinical examples to highlight concerns about the limitations of rating scales, examine their underlying rationales, clarify their implications, explore potential solutions, and make some recommendations for future research. We show that improvements in the scientific rigour of rating scales can improve the chances of reaching the correct conclusions about the effectiveness of treatments.

Introduction

A recent review of UK health research funding[1] emphasised the importance of translational research and highlighted an internationally recognised problem: success in basic science rarely leads to effective treatments. Why have state-of-the-art clinical trials failed to deliver treatments? Are all candidate molecules that work in controlled laboratory settings worthless when studied in human beings? Conversely, do some of the methods used to test the efficacy of treatments hinder advances in basic science?

In this Review we focus on the latter point and, in particular, the rating scales used to measure the health outcomes of trials for the treatment of neurological diseases, which are increasingly selected as primary or secondary outcome measures in clinical trials.[2–6] Rating scales are, therefore, the main dependent variables on which decisions are made that influence patient care and guide future research; the adequacy of these decisions depends directly on the scientific quality of the rating scales.

Two developments indicate an appreciation of this fact: the increased application of the science of rating scales (psychometrics) for the measurement of health outcomes in clinical neurology; and the impending US Food and Drug Administration's (FDA) scientific requirements for patient-reported rating scales in clinical trials.[7,8] The FDA requirements are likely to be emulated by the European Medicines Agency (EMEA)[9] and will be pertinent to all rating scales, not just those that are patient-reported.

Our opening remarks might suggest that we think that published data from clinical trials are littered with type-2 errors due to poor rating scales. We do not know whether this is the case; nor do we know the frequency of type-1 errors that arise from problems with rating scales. We do know, however, that the reliability, validity, and responsiveness of different scales will influence their ability to estimate accurately the effect of a disease, to detect clinical change, and will have implications for calculations of sample size.[10] As such, the differences among rating scales have the potential to influence the outcome of clinical trials (panel 1).

Therefore, clinicians need to ensure that rating scales are fit for purpose, and maximising the scientific rigour of rating scales improves the chances of coming to the correct conclusion about the efficacy of a treatment. On this basis, a fundamental requirement of rigorous clinical trials is that the numbers generated by rating scales satisfy established scientific criteria as measurements of explicit, clinically meaningful variables.

A review of the subject of rating scales as outcome measures is, therefore, timely. We introduce the basic principles of the mechanics of rating scales and the limitations of the data derived from them. We discuss the benefits of moving to new psychometric methods and make recommendations to bring rating scales into line with what they measure. We highlight two methodological limitations that require attention to ensure that state-of-the-art clinical trials are underpinned by state-of-the-art measurements: the first limitation is that the numbers generated by most rating scales do not satisfy criteria as rigorous measurements; the second limitation is that we do not really know what variables most rating scales are measuring. These facts have great potential to undermine clinical trials, patient care, and research. The extent to which the limitations of rating scales are to blame for the failure of clinical trials to deliver treatments is unknown. However, our review highlights the potential contribution of rating scales and the way their data are analysed.

Panel 1: The DATATOP study

The study of selegiline for Parkinson's disease in the DATATOP study[11] is an excellent example of how the quality of a rating scale might influence the results of a study. The problem with the DATATOP study was the determination of the mechanisms responsible for the apparent delay to needing treatment with levodopa in patients treated with selegiline: was selegiline neuroprotective or did the improvement in the symptoms of Parkinson's disease mask ongoing neurodegeneration?[12] Unfortunately, the unified Parkinson's disease rating scale (UPDRS), the primary outcome measure in the DATATOP study, confounds symptoms with disabilities. Thus, the validity of the UPDRS scores as measurements is likely to be more problematic than its sensitivity to change. Because the UPDRS was developed without established techniques of rating scale construction, and its evaluation to date is incomplete, this supports the argument for higher standards of scale construction and evaluation.

Basis of rating scales as outcome measures

Some variables (eg, height and weight) can be measured directly. Other variables (eg, disability, cognitive function, and quality of life) are measured indirectly by how they manifest; therefore, we need a method to transform the manifestations of these "latent" variables into numbers that can be taken as measurements.[13]

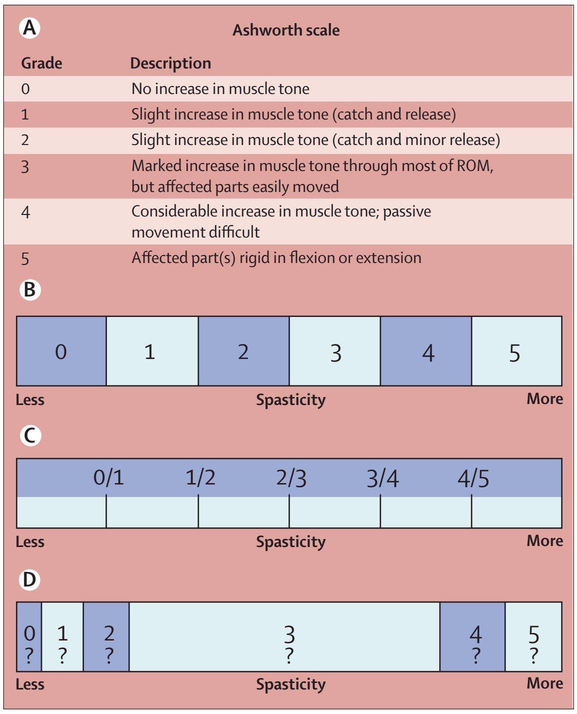
Rating scales are a means to measure latent variables, and two types of rating scale are commonly used in neurology: single item scales (eg, Ashworth scale [figure 1],[14] Kurtzke's expanded disability status scale [EDSS],[15] modified Rankin scale,[16] Hauser ambulation index,[17] and Hoehn and Yar scale[18]) and multiple item scales (eg, Rivermead mobility index [table 1, figure 2],[19] Barthel index,[20] and functional independence measure[21]).

Each type of scale has advantages and disadvantages. Single item scales generate scores that clinicians can easily identify with and communicate (eg, most neurologists recognise that someone with an EDSS of 6·5 can walk about 20 metres with two sticks). However, single item scales are scientifically weak because they have poor reliability, poor validity, and poor responsiveness. The low reliability is because single items are associated with substantial random error, and adequately high levels of reproducibility are hard to achieve.[22] Poor validity arises because it is difficult to represent a complex construct, such as spasticity, disability, cognitive function, or quality of life, with a single question.[22,23] As such, single items are often ambiguous. For example, the EuroQol[24] question "Rate you own health state today" does not provide a frame of reference for interpretation. Consequently, different people bring different frames of reference, and it is an ambiguous question.

The limited responsiveness of single item scales is due to the division of wide variables into only a few levels (figure 1). As such, each level represents a thick band of the continuum. This also contributes to limited reliability because, by definition, people cannot be localised precisely on the continuum; rather, they are located somewhere within a band. The theoretical limitations of single item scales have been confirmed empirically.[25–31]

The problems with single item scales led to the increased use of multiple item scales, where the scores from a set of items are combined to give a single value. The theory is clinically sensible:[22] the combination of multiple items reduces random error; hence, reliability is improved. Multiple item scales enable complex variables to be broken down to their component parts. Thus, validity, responsiveness, and precision are improved because the continuum is divided into more parts. The theoretical advantages of multiple item scales are supported by empirical evidence.[26,32–34]

However, although multiple item scales are scientifically strong, they generate less clinically tangible scores. For example, what does a score of 50 mean in a disability scale that ranges from 0 to 100?



Grade	Description
0	No increase in muscle tone
1	Slight increase in muscle tone (catch and release)
2	Slight increase in muscle tone (catch and minor release)
3	Marked increase in muscle tone through most of ROM, but affected parts easily moved
4	Considerable increase in muscle tone; passive movement difficult
5	Affected part(s) rigid in flexion or extension

Figure 1: **Modified Ashworth scale for measuring spasticity**
A. The scale comprises six ordered categories of increasing spasticity that are, by convention, assigned sequential integer scores (0–5). B. The spasticity variable marked out by the Ashworth scale. Each category represents a range on the continuum. C. The spasticity ruler. The marks represent the points of transition between adjacent categories (ie, the points at which the degree of spasticity is such that they are equally likely to score in either of the categories [eg, 0 or 1]). The use of sequential integers implies that the categories represent equal amounts of spasticity and, therefore, that a change or difference of one point has the same meaning in terms of the underlying variable (spasticity) anywhere on the continuum. This implication, shown by giving each category the same size, is clearly an improbable assumption. D. The more probable scenario; however, one of the main limitations of all single item scales is the inability to locate accurately the points of transition relative to each other, so that the ranges of spasticity of each category are in appropriate proportions. ROM=range of movement.

Statements about the relative scientific adequacy of single item and multiple item scales can be supported by theoretical reasoning and empirical confirmation; however, although the scientific limitations of single item scales as measurement instruments have long been recognised,[22,23] their continued use as outcome measures in clinical trials[35–37] shows that clinicians do not fully appreciate the drawbacks. The relative value of single item and multiple item scales is, of course, not absolute. Some clinical scenarios lend themselves to single questions (eg, constipation),[33] and multiple item scales need to be constructed appropriately and their scientific validity rigorously proven. Because we are concerned with the measurement of clinical variables, such as disability, for clinical trials, and to our knowledge there are no published studies that show the scientific superiority of single item scales over multiple items scales, this Review focuses on multiple item rating scales.

Multiple item rating scales comprise a set of items, each of which has two or more ordered response categories that are assigned sequential integer scores. Typically, the item scores are summed to give a total score (also called the raw, summed, or scale score), which is a measure of the variable quantified by the set of items (figure 2). Therefore, the use of multiple item rating scales has two fundamental requirements: evidence that the values produced are actually rigorous measurements and not just numbers; and evidence that the set of items map out the variable they purport to measure.

The requirement for rating scales to generate rigorous measurements

Phase III clinical trials need rating scales that generate rigorous measurements. Unfortunately, this is rarely achieved because most rating scales generate ordered scores that are only suitable for group comparison studies, rather than precise measurements of an individual.

Please pick "Yes" or "No" for each question	No	Yes
1. Turning over in bed		+
Do you turn over from your back to your side without help?		
2. Laying to sitting		+
From laying in bed do you get up to sit in the edge of the bed on your own?		
3. Sitting balance		+
Do you sit on the edge of the bed without holding on for more than 10 seconds?		
4. Sitting to standing		+
Do you stand up from any chair in less than 15 seconds (using hands, and with an aid if necessary)?		
5. Standing unsupported		+
Observe standing for 10 seconds without any aid		
6. Transfer		+
Do you manage to move from the bed to a chair and back again without any help?		
7. Walking inside and with an aid if needed	+	
Do you walk 10 metres with an aid if necessary but with no standby help?		
8. Stairs		+
Do you manage a flight of stairs without help?		
9. Walking outside (even ground)	+	
Do you walk around outside on pavements without help?		
10. Walking inside with no aid	+	
Do you walk 10 metres inside with no calliper, splint, or aid or standby help?		
11. Picking item off the floor	+	
If you drop something on the floor, do you manage to walk 5 metres, pick it up and then walk back?		
12. Walking outside (uneven ground)	+	
Do you walk over uneven ground (grass, gravel, dirt, snow, ice, etc) without help?		
13. Bathing		+
Do you get in/out of bath or shower unsupervised and wash yourself?		
14. Up and down four steps	+	
Do you manage to go up and down four steps with no rail but using an aid if necessary?		
15. Running	+	
Do you run 10 metres without limping in four seconds (fast walk is acceptable)		
Score=total number of "Yes" responses		8

The Rivermead mobility index is a 15 item, clinician reported scale for the measurement of mobility. Each item has two response categories: No=I am unable to do this task (score 0); or Yes=I am able to do this task (score 1). The score (total number of "Yes" answers) is used to generate measurements with traditional and new psychometric methods (figure 2).

Table 1: **Example response for the Rivermead mobility index**

Ordered scores are not scientific measurements

The raw data generated by rating scales, both item scores and total scores, are ordinal level, which means that the values are rank ordered. For example, the Ashworth scale has six categories that are ordered in terms of increasing spasticity: from none to rigid (figure 1). Although each category represents more spasticity than the previous category, the difference between categories in terms of amount of spasticity is unknown, and by assigning the categories sequential integer scores, the implication is that the differences are equal.

Multiple item scales extrapolate this process. For example, the multiple sclerosis walking scale (MSWS-12)[34,38] has 12 items and five item response categories—1=not at all; 2=a little; 3=moderately; 4=quite a bit; and 5=extremely—which are summed to give a total. Scoring the items with sequential integers implies equal differences in walking ability at the item level (differences between each response category is implied to be equal) and the total score level (a change of one point implies an equal change in walking ability across the range of the scale). But does this mean that the ordinal level scores that are produced by scales are measurements? The "no" lobby argue that a constant unit is an absolute requirement for measurement,[39–44] whereas the "yes" lobby argue that ordinal scores are weaker forms of measurement[44,45] or adequately approximate interval level measurements.[22,45,46]

The careful consideration of the relation between the scores assigned to item response categories and generated by scales and the measurements they imply is required. We believe that state-of-the-art clinical trials should, whenever possible, use rating scales that generate interval level measurements. The analysis and interpretation of differences in scores and changes during time are most meaningful when the unit of measurement is constant, and the numerical meaning of the numbers is maintained when they are subjected to statistical analysis.[41,42] Thus, a change or difference of one point has the same meaning throughout the continuum, which is not the case for ordinal scores, where a change or difference of one point varies in meaning across the continuum (eg, the Rivermead mobility index [RMI]; table 1, figure 2).

Clearly, a linear relationship between scale scores and the measurements they imply is unlikely, and the relationship must be determined rather than assumed. In fact, the relationship is S-shaped (figure 3), and empirical studies show that the meaning of a one point change in ordinal score varies up to 15-fold across the scale range, and that the variation is scale dependent.[47,48] This has obvious and serious implications for clinical trials, in which the analytical cornerstones are the examination of change in people and differences among groups.

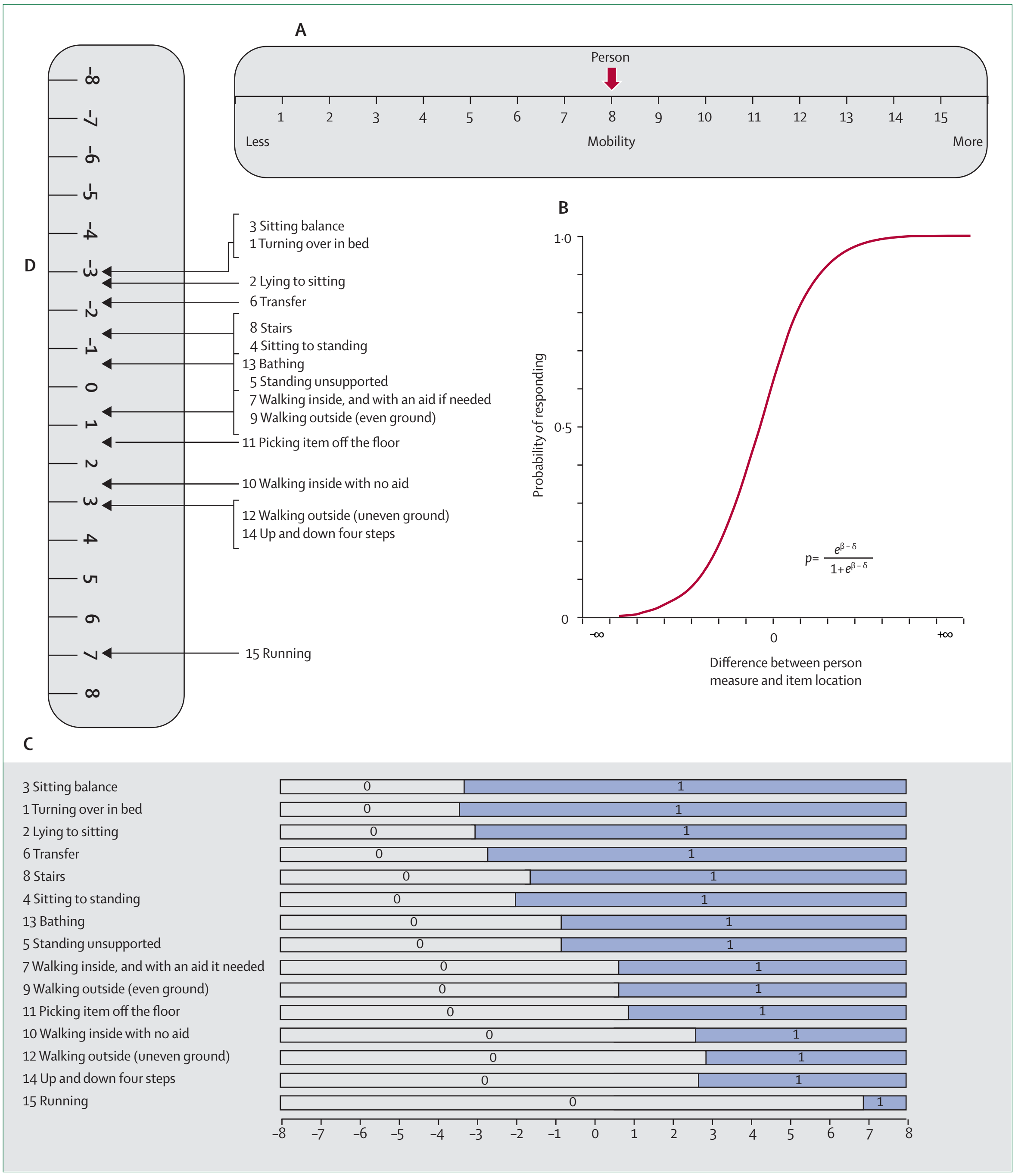


***Figure 2*: How measurements are generated from a multiple item rating scale with traditional and new psychometric methods**

A. Traditional psychometric methods, such as the Rivermead mobility index (RMI), generate mobility measurements from the answers given in table 1. Item scores are summed to give a total score that ranges from 0 (all “No” responses) to 15 (all “Yes” responses). The score (in this case 8) is the measure of mobility. This approach implies that the same change or difference in mobility is required to change a person’s score by one unit or that the items are spread equally across the continuum, which is highly unlikely. Such assumptions underpin the use of traditional psychometric methods but are untestable with those methods. B–D. An example of how new psychometric methods (Rasch measurement in this case) generate measurements from the answers to the RMI given in table 1. B. Rasch measurement theory states that the response of a person to any item is governed by the difference between the location of a person on the mobility ruler (β) and the location of the item on the mobility ruler (δ), shown in the graph. The relationship between the probability of responding “Yes” (y-axis) to an item and the difference (β – δ) between person and item locations (x-axis) is S-shaped (an ogive). The equation is the mathematic expression (model) that reproduces this ogive and represents explicitly Rasch’s measurement theory. Note: when the person location (β) is equal to the item location (δ), the probability of responding “Yes” or “No” is equal (50%) because β–δ=0. Thus, for any RMI item, the point at which the probability of responding “Yes” and “No” is equal is the transition point for that item and where it is located on the ruler. C. Computer software analyses of the dataset of responses from a sample of people to a set of items, to estimate the locations of people and items. The 2D plot shows the transition points between Yes and No for each of the 15 RMI items. Note the transition points differ across items. D. The ruler marked out by the RMI; the 15 items are shown relative to each other in equal interval units.

Panel 2: MRC spine stabilisation study

The results of the Medical Research Council spine stabilisation study[4] of people with lower back pain who were suitable for surgical stabilisation found no significant differences in outcome between rehabilitation and surgery. However, there was substantial variability in outcome in both treatment arms, which suggested that some people who were suitable for surgery had a poor outcome, whereas others had an excellent outcome with rehabilitation. Such variability makes it difficult to detect differences at the group comparison level, and it could be argued that this study should have compared the clinical features of responders and non-responders to treatment within and across treatment groups. Hence, the identification, revisitation, and re-evaluation of people who respond differently has the potential to enrich the inferences from trial data, build hypotheses for future research, and benefit studies that investigate clinical uncertainty.

Another related problem with ordinal scores is that they are only suitable for group-level comparisons; the confidence intervals around the ordinal score of an individual are wide. Two previous studies on the Barthel Index and MSWS-12 have shown that the confidence intervals around individual scores are +/–3·5 and +/–15·4 points, respectively, which equate to more than 30% of their respective total scale ranges.[49,38] Consequently, clinical trials cannot legitimately compare changes and differences among individuals,[50] which is important because treatment effects are typically variable, and group-based analyses only inform on the extent to which one treatment is statistically and generally better than another treatment. Understanding the complexities of why individuals undergo different levels and directions of change would be advantageous to interpret the results of clinical trials (panel 2).

The root of the problem: classical test theory

The root of these problems is the measurement theory that underpins the psychometric methods most widely used to analyse data from rating scales and determine the reliability and validity of a rating scale to the constructs they seek to estimate: classical test theory (CTT).[51–56]

A measurement theory is a theory of how the numbers generated by rating scales relate to measurements of the constructs they seek to estimate. CTT postulates that a person's rating scale score (the observed score [O]) is the sum of the unobservable measurement to be estimated (true score [T]) and the associated measurement error (E), where O=T+E. CTT assumes that measurement errors are randomly distributed and not correlated with the true score; furthermore, for any individual the measurement error associated with one scale is not correlated with the true score or measurement error of another scale.[57,58]

The simple theory of CTT and its associated assumptions expand to form the methods to test reliability and validity that are known as traditional psychometric methods.[57–59] However, because they are derived from CTT, their appropriateness requires that the theory and assumptions of CTT are supported by the data. If these requirements are not met, the conclusions of the data analysis might be incorrect. Therefore, CTT is a theory that cannot be tested, verified, or—more importantly—falsified in any dataset[60] because the parameters of the theory (T and E) cannot be determined in a way that enables the evaluation of their accuracy.[57,61]

This has four important implications. First, untestable measurement theories are, by definition, weak theories that lead to only weak inferences about the performance of a rating scale and what it measures. Second, theories that cannot be challenged are easily satisfied by datasets.[57,61] Third, because the parameters can not be estimated with confidence, only the ordinal raw scores (O) can be analysed. Finally, the equation derived from CTT for calculating the confidence intervals around the scores for individuals (95% CI=observed score +/–1·96 SEM) gives large values that indicate a lack of confidence when comparing changes and differences among individuals. Therefore, CTT has been called weak true score theory,[57,61] a tautology[43] and a theory that has no theory.[60]

Approaches to overcome the limitations of ordinal scores

The fact that rating scales generate ordinal scores is well known, and several potential solutions to manage this problem have been suggested, including dichotomising of scale scores and the use of parametric statistics to analyse the data.

Dichotomising of scale scores

Dichotomising is the assignment of clinically meaningful cut-off points (eg, Rankin scale scores are frequently dichotomised into disabled [scores 3–5] and not disabled [scores 0–2]). Although the simplification of outcomes is clinically appealing, there are three important concerns. The first concern is whether it is meaningful to interpret ordinal rating scale data at the level of the individual, which, as we have discussed, is not legitimate because the confidence intervals around the scores of individuals are wide. The second concern is that the dichotomising of scale scores reduces a spectrum of outcomes into two crude categories (eg, disabled or not disabled; normal or abnormal). The limitations of this have been discussed recently.[62] The third problem is that the dichotomising of scale scores does not deal with unequal scale increments; rather, by forming binary categories, the dichotomising of scale scores moves us further away from the goal of accurate outcomes measurement.

Parametric statistics

The use of parametric statistics with multiple item rating scale data has been the source of a long-standing debate. One side, which advocates the classification of scales as categorical, ordinal, interval, or ratio,[63] argues that the nature of the scale dictates which statistical tests can be used. Parametric statistics (those based on addition, subtraction, multiplication, or division) are only meaningful when the data are interval level (interval or ratio scales); therefore, ordinal scales must be analysed with non-parametric statistics. The other side argues that the nature of a rating scale should not influence the choice of statistics.[64–66] Two different justifications are given for this: first, statistical tests merely report a fact about a set of measurements and, as such, attempts to ban such reports are unreasonable;[46,67] second, the scores produced when items are summed approximate interval level measurements adequately enough to warrant analysis with parametric statistics. The evidence to support this argument is the high correlation between summed scores and the interval measurements they imply (figure 3), which was first reported by Likert,[68] and that parametric statistics, such as *t*-tests, can deal with the weaknesses of ordinal measurements.[66] Consequently, most rating scale data is analysed using parametric statistics. Neither justification, however, accounts for the real issue: ordered scores have unequal intervals. To solve this problem, a method is required that constructs equal interval measurements from ordinal rating scales data; when this is achieved, the debate about what statistics are permissible becomes redundant.

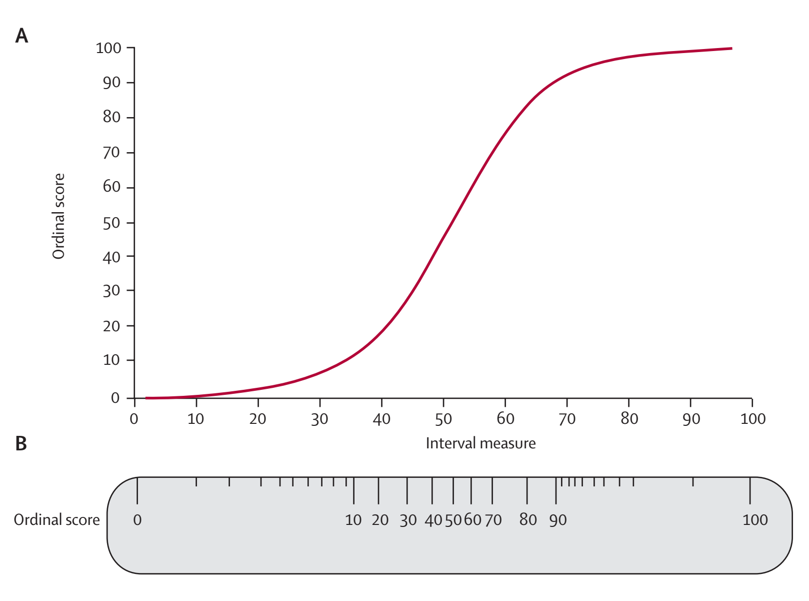


***Figure 3*: Relationship between the ordinal scores produced by summing items and the interval measurements that can be constructed from them with new psychometric methods**

A. In the graph, in which both axes have been transformed to have a range of 0–100, the correlation between the ordinal scores and the interval measurements is 0·95; however the relationship is an ogive (S-shaped curve) rather than linear. Thus, a change of 1 point in ordinal score corresponds to a change in interval measurement that varies across the range of the scale. For example, a 10 point change in ordinal score, from 50–60, corresponds to a change from 52 to 55 interval units. By contrast, 10 point changes in raw score from 0–10 and 90–100 correspond to changes of 0–35 and 68–100 interval units, respectively. Thus, in this example, the meaning in equal interval units of a one point change in raw score varies 12-fold across the scale. B. The implications of this are shown by the 'ruler' produced from this graph.

Latent trait theories

The value of being able to construct measurements with equal intervals from ordinal rating scale data,[39] and the need to develop strong measurement theories,[69] were stated in the early 1900s. However, it was not until the 1960s that two related but different solutions were proposed: item response theory (IRT)[61,70–72] and Rasch measurement.[13,73–77] Together, these solutions are sometimes thought of as latent trait theories (LTTs), new, or modern psychometric methods.[78]

LTTs, like CTT, are measurement theories that are presented as equations (mathematical models); from these models, statistical methods are derived to analyse rating scale data and test the reliability and validity of the scale. However, unlike CTT, LTTs are thought of as strong theories because they can be tested, verified, or falsified. LTTs also differ from CTT because they focus on the relationship between a person's measurement and the probability of them responding to an item, rather than the relationship between a person's measurement and their observed total score on the scale. This is exemplified by Rasch measurement theory, which postulates that the probability of a person's response to each of the categories of a rating scale item is governed by the difference between where the person is on the scale and where the item is on the continuum measured by the item set (figure 2). In essence, a Rasch analysis, typical of any analysis of LTTs, assesses the extent to which the responses of the observed item accord with the responses predicted by the mathematical model.

When the data fit the LTT model, the estimates derived from the model are deemed robust because the measurement theory is supported by the data. When the data do not fit the model, two lines of inquiry are possible: one questions the suitability of the mathematical model as a representation of the theory embodied in the data; the other questions the suitability of the data as a representation of the measurement theory embodied in the model. These divergent lines of inquiry are the fundamental difference between IRT and Rasch measurement.

Essentially, albeit an oversimplification, when the data do not fit the chosen LTT model, the IRT approach is to find a mathematical model that best fits the observed item response data. By contrast, the Rasch measurement approach is to explore why the data do not fit the Rasch model. Thus, proponents of IRT use a range of item response models that differ in the number of parameters (components of the mathematical model that can be estimated), whereas proponents of Rasch measurement use only one (Rasch) model.

The IRT approach is consistent with data modelling and is easy to comprehend. The Rasch measurement approach is less common, less easy to understand, and warrants explanation because it has inherent mathematical properties that are not found in any other person–item response model.[43,47,78] The main property is the ability to test for, and thus achieve, invariance (stability).[78] This is achieved because the central tenet of measurement is that the instrument of measurement is stable—not sample dependent—and the property being measured is stable at one point in time—not instrument dependent. Only Rasch measurement can test stability of instruments and people; other parameters in IRT models render these estimates sample dependent.[43,78]

IRT and Rasch measurement have substantial advantages over CTT for clinical trials and, as such, clinicians should use them. However, the above discussion raises two questions. Which approach is better? And does it matter which approach is used? The answer to both questions depends on which central philosophy is followed: IRT or Rasch measurement. Because IRT prioritises the observed data, it sees the one-model Rasch perspective as too restrictive, and the selection of data to meet that model as a threat to the validity of the content.[79,80] However, because Rasch measurement prioritises the mathematical model, proponents of Rasch measurement see the process of modelling data as a bar to the core requirements of measurement, too accepting of poor quality data, and a threat to the validity of the construct.[80] Not surprisingly, IRT and Rasch measurement are suggested to have irreconcilable differences,[78] and the two groups have come into conflict about which approach is preferable.[78,81–83]

The requirement to know precisely what variables are measured

See Online for webappendix 1

Clinical trials require rating scales that actually measure the health constructs that they claim to (ie, the scales are valid) and health constructs that are clinically meaningful and can be interpreted. Unfortunately, current methods to establish the validity of a rating scale rarely meet these goals.

The current methods to establish validity are weak

When a set of items is used as a scale, a claim is made that a construct is being measured,[84] and some theory of that construct (a construct theory) is implicit.[85] Thus, by implication, the aim of validity testing is to establish the extent to which the construct theory is supported. Current methods to establish the validity of a scale are weak because they lack formal methods that define and test construct theories.[85] Although scales and the constructs they claim to measure always have names, they are rarely underpinned by a deduced theory of the construct being measured. This situation is surprising; explicit definitions of constructs would seem to be a prerequisite for the development and validation of a scale. The consternation is, in part, because the constructs measured by many scales are determined during their development. Typically, scale developers generate a large pool of items that they group—either statistically or thematically—into potential scales; they then decide what construct each group seems to measure, and remove unwanted or irrelevant items. The main limitation of this approach is that the content of the scale, rather than the construct intended for measurement, defines what the scale measures. Grouping items statistically or thematically does not ensure that the items in a group measure the same construct but does explain why items such as "having trouble meeting the needs of my family" and "few social contacts outside the home" are in widely used scales purporting to measure mobility and fatigue, respectively. Furthermore, both methods to group items avoid the process of defining and conceptualising variables, which is central to valid measurement.[86–89]

However, even if scales are underpinned by explicit construct theories, standard methods of validity testing would not enable those theories to be tested adequately. Why? Because current methods, which integrate evidence from statistical and non-statistical tests, provide at best circumstantial evidence that a set of items measures a specific construct.

Non-statistical tests of validity typically assess content validation and face validation. Content validation assesses whether all the relevant or important content is sampled during scale development,[90] sensible methods were used to construct the scale, and a representative collection of items were assessed.[91] Face validation assesses whether the final scale measures what it is supposed to.[90,91] More than 50 years ago, Guilford named these evaluations "validity by assumption" and "faith validity"[92] and they are essentially unchallenged, with the exception of Alvan Feinstein's contribution of "clinimetrics" (webappendix 1).

Statistical tests of scale validity are more formal than their non-statistical counterparts but are still weak evaluations of the extent to which a set of items measures a construct. For example, statistical examinations of internal construct validity[93] (eg, factorial validity[94] or internal consistency) test the extent to which the items of a scale are related statistically. This does not confirm that a set of items mark out a clinically meaningful variable nor tell us what a scale measures.

Statistical tests of external construct validity consist of a range of examinations, including correlations with other measures,[95,96] tests of known group differences,[97] and hypothesis testing.[93,95] The tests assess the extent to which scale scores behave as predicted and seek to determine if a scale does what it is intended to do.[22]

Testing convergent and discriminant construct validity[96] is deemed the strongest statistical evidence of scale validity. Here, multiple scales, which measure similar and dissimilar constructs, are applied to a sample. The scores are correlated, and the pattern and magnitude of the correlations determine if the scale being validated correlates higher with scales that measure similar constructs rather than scales that measure dissimilar constructs. The limitation of this approach is that to show that a scale does not correlate highly with measures of a dissimilar construct tells us nothing about what the scale actually measures. Similarly, to show that a scale correlates highly with measures of similar constructs tells us only that the two are related.

A key problem with all statistical tests of validity is their focus on people scores and how these scores vary among people. There is no independent means to assess the extent to which the aim of the scale is satisfied.[98] Consequently, these validation techniques are based on circular reasoning,[98] generate circumstantial evidence,[43] enable only limited development of construct theories, and result in only a basic understanding of what is being measured.[85] However, in keeping with their non-statistical counterparts, they have been, essentially, unchallenged for decades.

Theory-referenced measurement

Two requirements are needed to advance our understanding of precisely what scales measure: explicit theories of the constructs being measured; and explicit methods to test those theories. Although several researchers have investigated these requirements,[85,98–103] one group has developed their ideas to an advanced level.[85,98,101] However, their work is largely inaccessible to clinicians because it concerns the measurement of reading ability. The central premise of this group's approach is to change from studying people to studying items.[85] A logical rationale underpins this approach. Multiple item scales consist of a set of items that aim to measure a single construct. Thus, the aim of these items is to mark out the construct as a continuum on which people can be located, which implies that the individual items of a scale are located across the continuum, analogous to the way that the locations of individual people are spread out across the continuum. If a scale developer can explain why items are located at different points on the continuum (ie, if the scale developer can define the characteristics that determine the location of an item) they are justified in saying that they know what construct is being measured.[85,98] To do this, scale developers need to propose and test explicit construct theories. The validity of a construct theory is, then, the extent to which the theory predicts variation in the locations of items. This process is made explicit if the construct theory can be articulated as a mathematical or construct specification equation (figure 4).[85]

Construct specification equations are developed by regression analysis of item locations on selected item characteristics. They afford a test of fit between scale-generated observations and theory.[98] In essence, the greater the proportion of variation in item location explained by the selected item characteristics, the greater the support for the proposed construct theory, the greater the evidence for scale validity, and the more clinically meaningful the interpretation of person locations. Moreover, construct specification equations enable different construct theories to be articulated and challenged; thus, enabling dynamic interplay between theory and scale,[85] and a thorough investigation of individual items to aid item development and selection. An example from educational measurement of theory-referenced measurement—the Lexile system for measuring reading ability—is provided (webappendix 2).

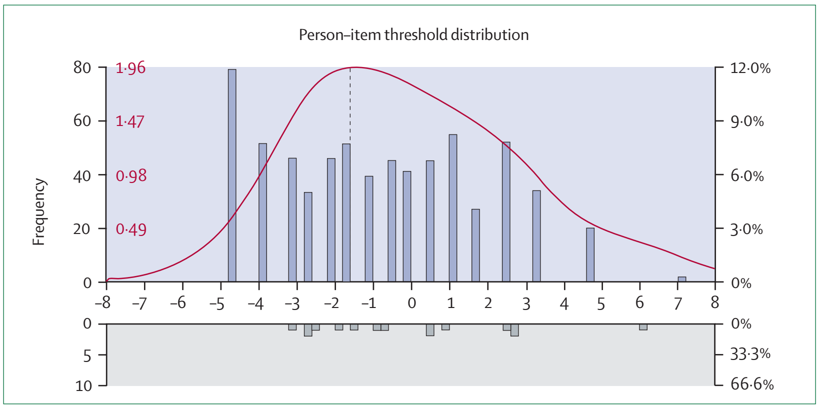


Figure 4: **Results from the Rasch analysis of RMI data from the CAMS study**[2]
The columns in the top half are the person locations; the columns in the lower half are the item locations (n=15). The locations of both items and persons are on the same, equal-interval metric. The metric is unbounded and, therefore, runs (theoretically) from –∞ to + ∞ and is centred around 0 because the analysis always centres the mean of the item locations at zero. Clearly, people located at different places on the continuum are assumed to have different levels of mobility. Likewise, RMI items (mobility tasks) located at different places on the continuum need different amounts of mobility.[85] Theory-referenced measurement is based on the latter statement, and suggests that investigators propose and test, "construct" theories about the factors that determine the locations of the RMI items. This process is made explicit by articulating the construct theory as a mathematical equation or construct specification equation.[85] The validity of any construct theory is the extent to which the construct specification equation predicts variation in item locations. When the equation predicts the vast majority of the variance in the item locations, a scale developer can explain why items are located at different points on the continuum: that is, they can define the characteristics that determine item locations. At this point they are justified in saying that they know what construct is being measured.[85,98] This graph was produced by the Rasch measurement software program RUMM 2020.

Recommendations

The FDA draft recommendations for patient-reported rating scales in clinical trials highlight the importance of "conceptually sound, reliable, and valid measures".[8] Such an acknowledgment is a vital, albeit first, step. Surprisingly, the document barely mentions new psychometric methods, despite their clear advantages and increased use;[104–109] furthermore, despite the emphasis on the improvement of methods to establish validity, they do not provide detailed guidance on how this can be achieved. Here we offer some recommendations to build on the FDA draft guidelines for clinical trials and rating scale development and evaluation.

For more information on **RUMM 2020** see http://www.rummlab.com

Clinical trials

Increased awareness of the crucial role of rating scales is needed. State-of-the-art clinical trials continue to use rating scales that are scientifically weak[4,5] or report scores for scales that are invalid.[4,110–113] Furthermore, scales continue to be developed without adequate recourse to recognised methods for scale construction.[114] More clinicians need to be formally trained in rating scale methods, to ensure that health measurement develops clinically meaningful scales; furthermore, journal editors, reviewers, and grant-giving bodies should include or have direct access to people who are trained in the development and evaluation of rating scales.

The clinically meaningful advantages of new psychometric methods mean that future (and present) clinical trials rating scale data ought to be reanalysed with

See **Online** for webappendix 2

new psychometric methods. In addition, the differences between Rasch measurement and IRT need to be better publicised because these methods answer fundamentally different questions about rating scales. Although this is acknowledged by some,[43,58,78,115,116] many reports[104–109] and standard measurement texts[57,90,117] imply that there is no difference and, as such, inaccurate representations are common.[79,82,83,118]

Development and evaluation of rating scales

Scale development would benefit from being bottom up (a construct definition), rather than top down (a method of grouping items), to ensure that construct theory determines scale content and validation tests construct theories. This requires robust guidelines to define constructs and explicit definitions for content and face validity. Evaluation of rating scales should fully acknowledge the equally important and complementary roles of qualitative and quantitative evaluations. Scale evaluation could be rethought under these two headings, rather than the more traditional headings of reliability and validity. The aim of qualitative evaluation could be defined as the determination of the extent to which the items of a scale map a construct as a clinically meaningful continuum and, when available, the extent to which construct theory is supported. The aim of quantitative evaluation could be defined as the determination of the extent to which the numbers generated by scales are measurements rather than numerals.

Clinicians should aspire to theory-referenced measurement. Although construct specification equations are some way off, the measurement of neurological outcomes would benefit from the development of consensus guidelines to strengthen the theory that underpins new and existing scales. We recommend greater use of qualitative assessments, including the adoption of inductive and deductive approaches to construct theory development, evaluation of the extent to which the items of a scale map out the construct to be measured, the application of the most appropriate item phrasing, structuring, and context, and cognitive debriefing to ensure consistency in meaning.

Conclusions

In this Review we posed a question: why have state-of-the-art clinical trials in neurology failed to deliver treatments? Our aim was to highlight the potential contribution to this failure of the currently available rating scales and the way their data are analysed. However, rating scales are not always to blame. Indeed, the extent to which rating scales undermine inferences from clinical trials is difficult to determine. Our message is simple: when rating scales are used, they must be fit for purpose. We believe strongly that there can be no compromise in the efforts made to advance this area because rating scales will have an increasingly crucial role in the determination of patient care, the guidance of research directions, and the evaluation of advances in basic science.

We have not discussed the ability of scales to detect change, nor the relationship between clinically and statistically significant change. Although this is an area of great importance, research,[33,109] and debate, we believe that this relationship cannot be studied rigorously until we have rating scales that generate numbers that satisfy scientific criteria as measurements of explicit, clinically meaningful variables. We do not suggest that traditional methods of analysing rating scale data are valueless. We do, however, advocate the advantages of moving to methods of analysis that articulate quantities in linear units because these enable clinicians to compare and study meaningfully the differences between people and how they change with time. Moreover, the ability to measure an individual's score accurately and estimate the confidence intervals around the score at any time point offers the legitimate and meaningful study of differences and changes at the level of the individual, which is the unit of clinical practice.

A comment we commonly hear is that more sensitive scales will simply lead to type-1 errors. This concern is often used as a justification for blunt scales, which must be clinically meaningful if they detect statistically significant change. Certainly, the greater the ability of a scale to detect change, the greater the possibility that a clinical trial will detect change that is not clinically significant. But the reverse is also true. To avoid type-1 or type-2 errors, we must rely on the clarification of which changes in scores on rating scales are clinically significant, which is a matter of interpretation of the scale score. With respect to this, two things are noteworthy: logically, for a scale to detect clinically

Search strategy and selection criteria

Our Review is a focused critique of the literature on the basis of articles, reports, and book chapters that span more than a century of research in three areas: psychometrics, health measurement, and neurological clinical trials. These were collected as part of the general strategy in our unit (Neurological Outcome Measures Unit) during the past 15 years to develop a clear and detailed understanding of the science behind rating scales. Our search strategy included searches of electronic databases (including Medline, Cochrane library, PubMed, Psychoinfo, and Embase) from 1950 to September, 2007, or whenever the database started if it was later than 1950). The key search terms included: "neuro", "multiple sclerosis", "Parkinson's disease", "spine", "treatment trial", "questionnaire", "scale", "measure", "instrument", "validity", "reliability", "responsiveness", and "psychometric". In addition, we checked reference lists of selected articles and hand-searched relevant journals, reports, and books. The remaining articles were identified from the authors' collections and by consultations with experts.

significant change and distinguish the change from clinically unimportant change, the threshold to detect change must be greater than the threshold of clinically significant change. Also, the assignment of meaning to measurements and how they change or differ is typically done after a rigorous method of quantification has been developed.

Measurement of health outcomes is a new and developing field. Our perspectives come from a critical re-evaluation of our own work and experiences during the past 15 years. During this time, we have made many of the mistakes we identify, perhaps because the field continues to move quickly, lacks consensus, is often inaccessible to clinicians, is frequently complex, and is often abstract. The solutions to the problems discussed in this Review are a challenge for clinicians. The mathematics of new psychometric methods and the development of construct theories and specification equations for health variables require considerable intellectual investment, which makes it far easier for neurologists and other clinicians to use current methods than to meet those challenges. However, the patients we treat, whose interests we profess to advocate and whom we will ultimately become, have much to lose from that action.

Contributors

JCH conceptualised and wrote this Review. SJC contributed to the writing of this Review, the preparation of tables and figures, the search for references, and formatting. JPZ and AJT reviewed and commented on drafts of this Review and guided the document through its development.

Conflicts of interest

JCH has received honoraria for educational and consulting activities from Eisai, Biogenidec, GSK, and Pfizer, and support to attend international meetings from Biogenidec, Schering, and Teva. SJC has received honoraria for educational and consulting activities from Eisai, Bayer, and Oxford Outcomes. JPZ has received honoraria for educational and consulting activities from IKF, and support to attend international meetings from Biogenidec. AJT serves on advisory boards for Novartis and Genentech, chairs Teva's data, safety, and monitoring committee for the glatiramer acetate for amyotrophic lateral sclerosis trial, and has received honoraria from Schering and Serono for lectures. This review was not supported by a corporate sponsor.

Acknowledgments

The authors wish to acknowledge the people who have contributed to their development in this area.

References

1 Cooksey D. A review of health research funding. Norwich: HM Treasury, 2006.

2 Zajicek J, Fox P, Sanders H, et al. Cannabinoids for treatment of spasticity and other symptoms related to multiple sclerosis (CAMS study): multi-centre randomised placebo-controlled trial. *Lancet* 2003; **362:** 1517–26.

3 Aisen P, Schafer K, Grundman M, et al. Effects of rofecoxib or naproxen *vs* placebo on Alzheimer's disease progression: a randomized controlled trial. *JAMA* 2003; **289:** 2819–26.

4 Fairbank J, Frost H, Wilson-MacDonald J, Yu L, Barker K, Collins R. Randomised controlled trial to compare surgical stabilisation of the lumbar spine with an intensive rehabilitation programme for patients with chronic low back pain: the MRC spine stabilisation trial. *BMJ* 2005; **330:** 1233.

5 Lees K, Zivin J, Ashwood T, et al. NXY-059 for acute ischemic stroke. *N Engl J Med* 2006; **354:** 588–600.

6 Olanow C, Schapira A, Lewitt P, et al. TCH346 as a neuroprotective drug in Parkinson's disease: a double-blind, randomised, controlled trial. *Lancet Neurol* 2006; **5:** 1013–20.

7 United States Food and Drug Administration. Patient reported outcome measures: use in medical product development to support labelling claims, 2006. www.fda.gov/cber/gdlns/prolbl.pdf.

8 Revicki D. FDA draft guidance and health-outcomes research. *Lancet* 2007; **369:** 540–42.

9 European Medicines Agency. Reflection paper on the regulatory guidance for the use of the health-related quality of life (HRQL) measures in the evaluation of medicinal products. London: European Medicines Agency, 2006.

10 Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. How responsive is the MSIS-29? A comparison with other self report scales. *J Neurol Neurosurg Psychiatr* 2005; **76:** 1539–43.

11 DATATOP group. DATATOP: a multicenter controlled clinical trial in early Parkinson's disease. *Arch Neurol* 1989; **46:** 1052–60.

12 Stocchi F, Olanow C. Neuroprotection in Parkinson's disease: clinical trials. *Ann Neurol* 2003; **53:** S87–S99.

13 Wright BD, Masters G. Rating scale analysis: Rasch measurement. Chicago: MESA, 1982.

14 Ashworth B. Preliminary trial of carisoprodol in multiple sclerosis. *Practitioner* 1964; **192:** 540–42.

15 Kurtzke JF. Rating neurological impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology* 1983; **33:** 1444–52.

16 Rankin J. Cerebral vascular accidents in patients over the age of 60: II. Prognosis. *Scott Med J* 1957; **2:** 200-215.

17 Hauser S, Dawson D, Lehrich J. Intensive immunosuppression in progressive multiple sclerosis: a randomised three-arm study of high dose intravenous cyclophosphamide, plasma exchange, and ACTH. *N Engl J Med* 1983; **308:** 173–80.

18 Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. *Neurology* 1967; **17:** 427–42.

19 Collen FM, Wade DT, Robb GF, Bradshaw CM. The Rivermead Mobility Index: a further development of the Rivermead Motor Assessment. *Int Disabil Stud* 1991; **13:** 50–54.

20 Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Maryland State Med J* 1965; **14:** 61–65.

21 Granger CV, Hamilton B, Keith R, Zielezny M, Sherwin F. Advances in functional assessment for medical rehabilitation. *Topics Geriatr Rehab* 1986; **1:** 59–74.

22 Nunnally JC. Psychometric theory, 1st edn. New York: McGraw-Hill, 1967.

23 Manning W, Newhouse J, Ware J. The status of health in demand estimation: or beyond excellent, good, fair, and poor. In: Fuchs V, ed. Economic aspects of health. Chicago: The University of Chicago Press, 1982: 143–84.

24 EuroQol Group. EuroQoL: a new facility for the measurement of health-related quality of life. *Health Policy* 1990; **16:** 199–208.

25 Haas B, Bergstrom E, Jamous A, Bennie A. The inter rater reliability of the original and of the modified Ashworth scale for the assessment of spasticity in patients with spinal cord injury. *Spinal Cord* 1996; **34:** 560–64.

26 Hobart JC, Freeman JA, Thompson AJ. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. *Brain* 2000; **123:** 1027–40.

27 Blackburn M, van Vliet P, Mockett S. Reliability of measurements obtained with the modified Ashworth scale in the lower extremities of people with stroke. *Phys Ther* 2002; **82:** 25–34.

28 Clopton N, Dutton J, Featherston T, Grigsby A, Mobley J, Melvin J. Interrater and intrarater reliability of the Modified Ashworth Scale in children with hypertonia. *Pediatric Physical Therapy* 2005; **17:** 268–74.

29 Wilson J, Hareendran A, Hendry A, Potter J, Bone I, Muir K. Reliability of the modified Rankin Scale across multiple raters: benefits of a structured interview. *Stroke* 2005; **36:** 777–81.

30 Yam W, Leung M. Interrater reliability of Modified Ashworth Scale and Modified Tardieu Scale in children with spastic cerebral palsy. *J Child Neurol* 2006; **21:** 1031–35.

31 New P, Buchbinder R. Critical appraisal and review of the Rankin scale and its derivatives. *Neuroepidemiology* 2006; **26:** 4–15.

32 McHorney CA, Ware JE Jr, Rogers W, Raczek AE, Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts. *Med Care* 1992; **30:** MS253–MS265.

33 Sloan JA, Aaronson N, Cappelleri JC, Fairclough DL, Varricchio C, and the Clinical Significance Consensus Meeting Group. Assessing the clinical significance of single items relative to summated scores. *Mayo Clin Proc* 2002; **77:** 479–87.

34 Hobart JC. Rating scales for neurologists. *J Neurol Neursurg Psychiatr* 2003; **74:** iv22–iv26.

35 Vaney C, Heinzel-Gutenbrunner M, Jobin P, et al. Efficacy, safety and tolerability of an orally administered cannabis extract in the treatment of spasticity in patients with multiple sclerosis: a randomized, double-blind, placebo-controlled, crossover study. *Mult Scler* 2004; **10:** 417–24.

36 Kappos L, Freedman M, Polman C, et al. Effect of early versus delayed interferon beta-1b treatment on disability after a first clinical event suggestive of multiple sclerosis: a 3-year follow-up analysis of the BENEFIT study. *Lancet* 2007; **370:** 389–97.

37 Uyttenboogaart M, Luijckx G, Vroomen P, Stewart R, De Keyser J. Measuring disability in stroke: relationship between the modified Rankin scale and the Barthel index. *J Neurol* 2007; **254:** 1113–17.

38 Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. Measuring the impact of MS on walking ability: the 12-item MS Walking Scale (MSWS-12). *Neurology* 2003; **60:** 31–36.

39 Thorndike EL. An introduction to the theory of mental and social measurements. New York: The Science Press, 1904.

40 Thurstone LL. Theory of attitude measurement. *Psychol Rev* 1929; **36:** 222–41.

41 Merbitz C, Morris J, Grip J. Ordinal scales and foundations of misinference. *Arch Phys Med Rehabil* 1989; **70:** 308–12.

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

43 Massof R. The measurement of vision disability. *Optom Vis Sci* 2002; **79:** 516–52.

44 Michell J. Measurement: a beginner's guide. *J Appl Meas* 2003; **4:** 298–308.

45 Michell J. An introduction to the logical of psychological measurement. New Jersey: Lawrence Erlbaum Associates, 1990.

46 Michell J. Measurement scales and statistics: a clash of paradigms. *Psychol Bull* 1986; **100:** 398–407.

47 Wright B. A history of social science and measurement. *Educ Meas* 1997; **52:** 33–52.

48 Katzenschlager R, Schrag A, Evans A, et al. Quantifying the impact of dyskinesias in Parkinson's disease: the PDYS-26. *Neurology* 2007; **69:** 555–63.

49 Hobart J, Thompson A. The five-item Barthel Index. *J Neurol Neurosurg Psychiatr* 2001; **71:** 225–30.

50 McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 1995; **4:** 293–307.

51 Spearman CE. The proof and measurement of association between two things. *Am J Psychol* 1904; **15:** 72–101.

52 Spearman CE. Correlations of sums and differences. *Br J Psych* 1913; **5:** 417–26.

53 Spearman CE. Correlation calculated from faulty data. *Br J Psychol* 1910; **3:** 271–95.

54 Spearman CE. Demonstration for true formulae of true measurement of correlation. *Am J Psychol* 1907; **18:** 161–69.

55 Spearman CE. "General intelligence" objectively determined and measured. *Am J Psychol* 1904; **15:** 201–92.

56 Traub R. Classical Test Theory in historical perspective. *Educ Meas* 1997; **16:** 8–14.

57 Allen MJ, Yen WM. Introduction to measurement theory. California: Brooks/Cole, 1979.

58 Novick MR. The axioms and principal results of classical test theory. *J Math Psychol* 1966; **3:** 1–18.

59 Crocker L, Algina J. Introduction to classical and modern test theory. Forth Worth, Texas: Harcourt, Brace, Jovanovich, 1986.

60 Lord FM. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

61 Lord FM, Novick MR. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

62 Kasner S. Clinical interpretation and use of stroke scales. *Lancet Neurol* 2006; **5:** 603–12.

63 Stevens SS. On the theory of scales of measurement. *Science* 1946; **103:** 677–80.

64 Burke CJ. Additive scales and statistics. *Psychol Rev* 1953; **60:** 73–75.

65 Boneau CA. A note on measurement scales and statistical tests. *Am Psychol* 1961; **16:** 260–61.

66 Baker B, Hardyck C, Petronovich L. Weak measurement vs. strong statistics: an empirical critique of S.S. Stevens proscriptions on statistics. *Educ Psychol Meas* 1966; **26:** 291–309.

67 Lord FM. On the statistical treatment of football numbers. *Am Psychol* 1953; **8:** 750–51.

68 Likert RA. A technique for the measurement of attitudes. *Arch Psychol* 1932; **140:** 5–55.

69 Thurstone LL. A method for scaling psychological and educational tests. *J Educ Psychol* 1925; **16:** 433–51.

70 Hambleton RK, Swaminathan H. Item response theory: principles and applications. Boston, Massachussets: Kluwer-Nijhoff, 1985.

71 Lord F. A theory of test scores. *Psychometric Monogr* 1952; **7.**

72 Lord FM. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika* 1952; **17:** 181–94.

73 Wright BD. Solving measurement problems with the Rasch model. *J Educ Meas* 1977; **14:** 97–116.

74 Andrich D. A rating formulation for ordered response categories. *Psychometrika* 1978; **43:** 561–73.

75 Wright BD, Stone MH. Best test design: Rasch measurement. Chicago: MESA, 1979.

76 Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research, 1960.

77 Andrich D. Rasch models for measurement. Beverley Hills, California: Sage, 1988.

78 Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med Care* 2004; **42:** I7–I16.

79 Cook K, Monahan P, McHorney C. Delicate balance between theory and practice. *Med Care* 2003; **41:** 571–74.

80 Fisher W. The Rasch debate: validity and revolution in education measurement. In: Wilson M, ed. Objective measurement: theory into practice. Norwood, New Jersey: Ablex, 1992.

81 Wright BD. Misunderstanding the Rasch model. *J Educ Meas* 1977; **14:** 219–25.

82 Divgi D. Does the Rasch model really work for multiple choice items? Not if you look closely. *J Educ Meas* 1986; **23:** 283–98.

83 Goldstein H, Wood R. Five decades of item response modelling. *British Journal of Math Stat Psychol* 1989; **42:** 139–67.

84 Cronbach LJ. The two disciplines of scientific psychology. *Am Psychol* 1957; **12:** 671–84.

85 Stenner AJ, Smith M. Testing Construct theories. *Percept Mot Skills* 1982; **55:** 415–26.

86 Nicholl L, Hobart JC, Cramp AFL, Lowe-strong AS. Measuring quality of life in multiple sclerosis: not as simple as it sounds. *Mult Scler* 2005; **11:** 708–12.

87 Andrich D. A framework relating outcomes based education and the taxonomy of educational objectives. *Stud Educ Eval* 2002; **28:** 35–59.

88 Andrich D. Implication and applications of modern test theory in the context of outcomes based research. *Stud Educ Eval* 2002; **28:** 103–21.

89 Hobart JC, Riazi A, Thompson AJ, et al. Getting the measure of spasticity in multiple sclerosis: the Multiple Sclerosis Spasticity Scale (MSSS-88). *Brain* 2006; **129:** 224–34.

90 Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use, 2nd edn. Oxford: Oxford University Press, 1995.

91 Nunnally JC. Introduction to psychological measurement. New York: McGraw-Hill, 1970: 197.

92 Guilford JP. Psychometric methods, 2nd edn. New York: McGraw-Hill, 1954: 399–400.

93 Bohrnstedt GW. Measurement. In: Rossi PH, Wright JD, Anderson AB, eds. Handbook of survey research. New York: Academic Press, 1983: 69–121.

94 Maurischat C, Ehlebracht-Konig I, Kuhn A, Bullinger M. Factorial validity and norm data comparison of the Short Form 12 in patients with inflammatory-rheumatic disease. *Rheumatol Int* 2006; **26:** 614–21.

95 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955; **52**: 281–302.

96 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959; **56:** 81–105.

97 Kerlinger FN. Foundations of behavioural research, 2nd edn. New York: Holt, Rinehart, and Winston, 1973.

98 Stenner AJ, Smith M, Burdick D. Towards a theory of construct definition. *J Educ Meas* 1983; **20:** 305–16.

99 Enright MK, Sheehan KM. Modelling the difficulty of quantitative reasoning items: implications for item generation. In: Irvine SH, Kyllonen PC, eds. Item generation for test development. New Jersey: Lawrence Erlbaum Associates, 2002.

100 Embretson SE. A cognitive design system approach to generating valid tests: application to abstract reasoning. *Psychol Meth* 1998; **3:** 380–96.

101 Stenner AJ, Burdick H, Sandford EE, Burdick DS. How accurate are lexile text measures? *J App Meas* 2006; **7:** 307–22.

102 Stone MH. Knox cube test—revised. Itasca: Stoelting, 2002.

103 Stone MH, Wright BD, Stenner AJ. Mapping variables. *J Outcomes Meas* 1999; **3:** 308–22.

104 Ware JE Jr, Bjorner JB, Kosinski M. Practical implications of item response theory and computer adaptive testing. A brief summary of ongoing studies of widely used headache impact scales. *Med Care* 2000; **38** (suppl 11)**:** 73–82.

105 McHorney CA, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000; **38** (suppl 11)**:** 43–59.

106 Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000; **38** (suppl 11)**:** 28–42.

107 Cella D, Chang C-H. A discussion of item response theory and its application in health status measurement. *Med Care* 2000; **38** (suppl 11)**:** 66–72.

108 Hambleton RK. Item response theory modelling in instrument development and data analysis. *Med Care* 2000; **38** (suppl 11)**:** 60–65.

109 Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. *J Eval Clin Pract* 2000; **6:** 39–49.

110 Hobart JC, Freeman JA, Lamping DL, Fitzpatrick R, Thompson AJ. The SF-36 in multiple sclerosis (MS): why basic assumptions must be tested. *J Neurol Neurosurg Psychiatr* 2001; **71:** 363–70.

111 Hobart JC, Williams L, Moran K, Thompson AJ. Quality of life measurement after stroke: uses and abuses of the SF-36. *Stroke* 2002; **33:** 1348–56.

112 Jenkinson C, Hobart JC, Chandola T, Fitzpatrick R, Peto V, Swash M. Use of the short form health survey (SF-36) in patients with amyotrophic lateral sclerosis: tests of data quality, score reliability, response rate and scaling assumptions. *J Neurol* 2002; **249:** 178–83.

113 Cano SJ, Thompson A, Fitzpatrick R, et al. Evidence-based guidelines for using the Short Form 36 in cervical dystonia. *Movement Disorders* 2006; **22:** 122–26.

114 Cano SJ, Hobart JC, Fitzpatrick R, et al. Patient-based outcomes of cervical dystonia: a review of rating scales. *Mov Disord* 2004; **19:** 1054–59.

115 Andrich D. The Rasch model explained. In: Alagumalai S, Curtis, DD, Hungi N, eds. Applied Rasch measurement: a book of exemplars. Dordrecht: Springer–Kluver, 2005.

116 Thissen D, Steinberg L. A taxonomy of item response models. *Psychometrika* 1986; **51:** 567–77.

117 Nunnally JC, Bernstein IH. Psychometric theory, 3rd edn. New York: McGraw-Hill, 1994.

118 Reeve B. An introduction to modern measurement theory. Bethesda: National Cancer Inst, 2002.