Rating Scales for Clinical Studies in Neurology—Challenges and Opportunities

a report by

Jeremy Hobart, **PhD**, **FRCP** *and* **Stefan Cano**, **PhD**

Neurological Outcome Measures Unit, Peninsula College of Medicine and Dentistry, Plymouth, and Institute of Neurology, London

DOI: 10.17925/USN.2008.04.01.12

Rating scales are increasingly used as primary or secondary outcome measures in clinical studies in neurology.[1] They are therefore becoming the key dependent variables upon 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, which is reflected by the increased application of rating scale science (psychometrics) in health outcomes measurement in neuroscience and increasing regulatory involvement by governing bodies such as the US Food and Drug Administration (FDA).[2,3] However, the majority of clinical studies in neurology that use rating scales are currently inadequate. Two simple examples illustrate some of the key issues.

First, current 'state-of-the-art' clinical trials in neuroscience continue to use scales that have been proved to be scientifically poor. This is demonstrated through even the most superficial of literature reviews. For example, in a brief literature search in PubMed we identified randomized controlled trials (RCTs) in multiple sclerosis (MS) published over a 20-year period (1987–2007). Of the 68 relevant articles, we found that 59% had used a rating scale. However, only six (15%) of those articles had included scales that had any supporting psychometric evidence. This situation can be found throughout neurology and is further exemplified by the continued widespread use of the Rankin scale in stroke research, despite growing concerns,[4] the Ashworth scale, despite its inherent weakness as a single-item scale (see below), and the Alzheimer's Disease Assessment Scale Cognitive Behavior Section (ADAS-cog) in dementia, despite important limitations (further information available from authors).

Jeremy Hobart, PhD, FRCP, is a Consultant Neurologist at Derriford Hospital, Plymouth, and a Senior Lecturer at the Peninsula College of Medicine and Dentistry, Devon and the Institute of Neurology, London. His clinical sub-specialist interest is the diagnosis and management of people with multiple sclerosis. His research interest is rating scales for measuring health outcomes. He has led the development of a number of rating scales, published over 75 articles in this area, and is the recipient of numerous research grants. Dr Hobart completed his medical training at St Mary's Hospital Medical School, London, and the National Hospital for Neurology, London.

E: jeremy.hobart@pms.ac.uk

Stefan Cano, PhD, is a Chartered Psychologist and Lecturer in Neurological Outcomes Measurement at the Institute of Neurology, London, and at the Peninsula College of Medicine and Dentistry, Devon. He has published primary research, literature reviews, and case reports in clinical and surgical journals across a variety of health-related disciplines. He is a member of the British Psychological Society. Dr Cano initially trained in psychology at University College London, and later worked at the Chelsea and Westminster Health Authority, London.

Second, statistical adequacy does not automatically confirm clinical validity or interpretability. An example from our own research focused on probably the most widely used patient-reported fatigue rating scale (currently used in over 70 studies). We conducted two independent phases of research. In the first phase, we carried out qualitative evaluations of validity through expert opinion (n=30 neurologists, therapists, nurses, and clinical researchers). The second phase involved a standard quantitative psychometric evaluation (n=333 MS patients). The findings from the second phase implied that the fatigue measure in question was reliable and valid. However, the qualitative study in the first phase did not support either the content or face validity. In fact, expert opinion agreed with the scale placement of only 23 items (58%), and classified all of its 40 items as non-specific to fatigue (further information available from authors).

Our research findings support the need for stringent quantitative and qualitative requirements for rating scales used in neurology; such scales must also be proved to be clinically meaningful and scientifically rigorous for valid interpretations of clinical studies. So, why is this not happening right now? There are two key problems. First, the numbers generated by most rating scales do not satisfy the scientific definition for measurements. Second, we do not really know what variables most rating scales are measuring. This article addresses these two problems by introducing some of the key issues in current rating scale research methodology. For readers who would like to learn more, we expand on these ideas in a recent review[1] and forthcoming monograph.[5]

Rating Scales as Measurement Instruments—Some Basic Principles

Before anything can be measured, the variable along which the measurements are to be made must be identified and marked out.[6] Common examples are rulers and weighing scales, which mark out length in centimeters (or inches) and weight in grams (or ounces), respectively. They highlight three central features of all measurements, as illustrated in *Figure 1*: first, instruments are constructed to make measurements; second, the attribute being measured can be marked out as a line, or continuum, onto which the measurements can be located; and third, the markings on the continuum represent the units of measurement.

Variables such as height and weight can be measured directly. Other variables—such as disability, cognitive functioning, and quality of

life, which are particularly relevant to neurological disease—must be measured indirectly through their manifestations. These are often called latent variables in order to emphasize this fact. The implication is that instruments must be constructed to transform the manifestations of latent variables into numbers that can be taken as measurements.[6]

Rating scales are instruments constructed to measure latent variables. Two main types of rating scale are used in health measurement: single-item and multi-item scales.[7] *Figure 2* shows how single-item scales, such as the Kurtzke's Expanded Disability Status Scale (EDSS),[8] mark out the variable they purport to measure. Other widely used single-item scales include Ashworth's scale for spasticity,[9] the modified Rankin scale,[10] Hauser's Ambulation index,[11] and the Hoehn and Yar scale.[12]

Multi-item scales consist of a set of items, each of which has two or more ordered response categories assigned sequential integer scores (e.g. Barthel Index,[13] Functional Independence Measure,[14] Multiple Sclerosis Walking Scale).[15] *Figure 3* shows the Rivermead Mobility Index (RMI)[16] as an example of a multi-item scale and how it represents a mobility variable as a 'ruler' of a count up to 15 points. Typically, item scores are summed to give a single total score for each person (also called raw, summed, or scale score), which is taken to be a 'measure' of the variable quantified by the set of items.

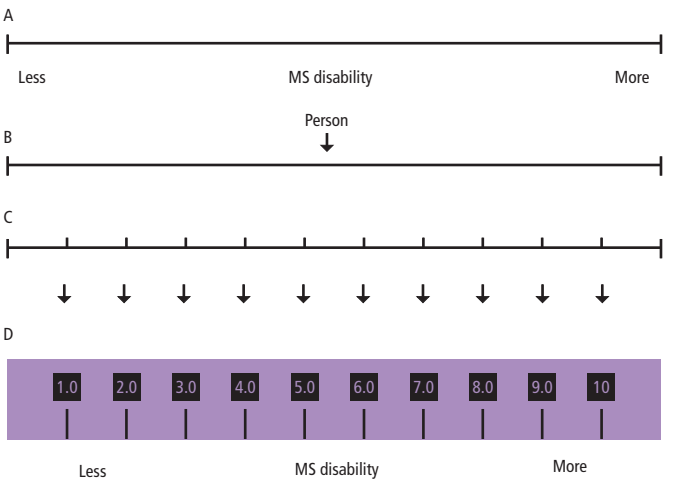
It is difficult to set up an argument against scale scores being ordinal in nature. However, a frequently asked question is: does this really matter in practice?

It has long been recognized that single-item scales are scientifically weak,[17] while multi-item scales can be scientifically strong. However, the fact that a single value, derived from summing the scores from a set of items, is taken to be a 'measurement' invokes two fundamental requirements of multi-item rating scales: evidence that the values produced satisfy the scientific definition of measurements rather than simply being numerals, and evidence that the set of items maps out the variable it purports to measure. In reality, these requirements are rarely met.

Problem 1—Scales Do Not Generate Measurement

The first problem with rating scales is that the numbers they generate are not measurements in the scientific sense of the word. To understand this statement we need to consider the definition of measurement and the extent to which the numbers generated by scales meet that definition. Measurement is defined as the quantitative comparison between two magnitudes of the same type, one of which is a standard unit, and in which the comparison is expressed as a numerical ratio.[18–21] An example makes this clinically intangible definition clear. Consider 10 meters in length. This is the comparison of two magnitudes (10 and 1) of the same type (meters) in which one magnitude is a standard unit (1 meter). The comparison is expressed as a numerical ratio (10/1 meters or 10 meters).

Figure 1: Central Features of All Measurements



A shows that a variable, here multiple sclerosis (MS) disability, can be represented as a line, or continuum, ranging from less disability to more disability. B shows a 'mark' that represents the location of a person on the variable and indicates the amount of disability that person has. C illustrates that to 'measure' a person's MS disability, the disability continuum must have marks that separate it into units. D shows a 'ruler' with equal interval units—the prototype of all measurements.

Thus, a fundamental requirement for making measurements, and meaningfully interpreting them, is the presence of a standard consistent unit. In this example the standard consistent unit is 1 meter.

Now consider rating scales. These assign numbers to rank-ordered clinically distinct magnitudes of unknown interval size. For example, the Rankin scale assigns sequential integer scores (0, 1, 2, 3, 4, 5) to a set of ordered clinical descriptions of worsening 'disability.' Likewise, multi-item scales assign sequential integer scores to progressive (ordered) item response categories (e.g: no/yes; not at all/a little/a lot; mild/moderate/severe), and these values are summed across items to give a total score. Undeniably, therefore, rating scale scores are ordinal-level data. More specifically, they are counts of the numbers of item response categories achieved. This tells us nothing about the distances between response categories or total scores (see *Figure 2*). Although counting observations is the beginning of measurement, as all observations begin as ordinal if not nominal data, something must be done to turn counts into measurements.[22] This is because a fundamental requirement of the definition of 'measurement' is a constant unit.[22–25]

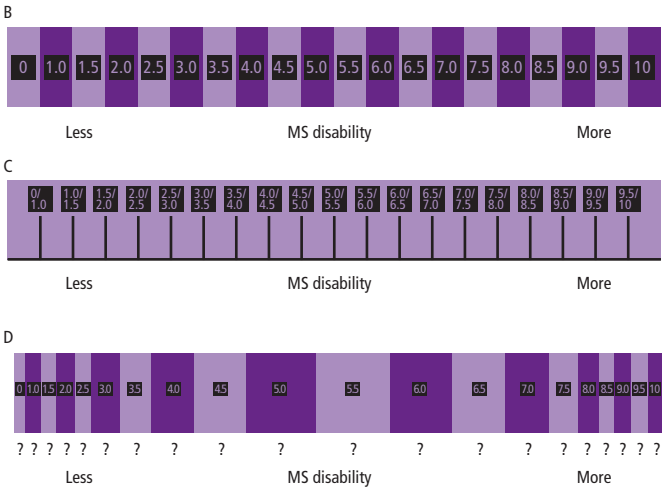
It is difficult to set up an argument against scale scores being ordinal in nature. However, a frequently asked question is: does this really matter in practice? This question arises from the logic that the clinical descriptors of the different levels of the Ashworth scale, for example, are ordered to map out progressive spasticity, and the logic that producing clinical descriptors representing near-equal intervals would be unrealistic. Therefore, why not simply assign sequential scores? The problem arises when the data are analyzed. The importance of a constant unit is that the numerical meaning of numbers is maintained when they are added, subtracted, divided, or multiplied (i.e. subjected to statistical analysis).[22,25] By simply assigning sequential integer scores we are implying that there is a constant unit, and by analyzing the data statistically and making clinical inferences we are believing it. This is a potentially very dangerous practice.

Figure 2: The Expanded Disability Status Scale and How It Maps Out a 'Ruler' for Measuring the Impact of Multiple Sclerosis

A

Grade	Descriptions
0	Normal neurological examination (FS = 0; cerebral grade 1 acceptable)
1.0	No disability (FS ≤1 excluding cerebral grade 1)
1.5	No disability (FS >1 excluding cerebral grade 1)
2.0	Minimal disability (1 x FS = 2, others 0/1)
2.5	Minimal disability (2 x FS = 2, others 0/1)
3.0	Moderate disability (1 x FS = 3, others 0/1) or mild disability in three or four FS (3/4 x FS = 2, others 0/1)
3.5	Fully ambulatory + moderate disability (1 x FS = 3 + 1/2 x FS = 2) or 2 x FS = 3, others 0/11, or 5 x FS = 2, others 0/1)
4.0	Fully ambulatory without aid, self-sufficient, up and about some 12 hours a day despite relatively severe disability (1 x FS = 4, others 0 or 1), able to walk without aid or rest some 500 meters (0.3 miles)
4.5	Fully ambulatory without aid, up and about much of the day, able to work a full day, may otherwise have some limitation of full activity or require minimal assistance, characterized by relatively severe disability (1 x FS = 4, others 0 or 1), able to walk without aid or rest for some 300 meters (975ft)
5.0	Ambulatory without aid or rest for about 200 meters (650ft), disability severe enough to impair full daily activities (e.g. able to work full day without special provisions; 1 x FS = 5, others 0 or 1)
5.5	Ambulatory without aid or rest for about 100 meters (325ft), disability severe enough to impair full daily activities (1 x FS = 5, others 0 or 1)
6.0	Intermittent or constant unilateral assistance (cane, crutch, brace) required to walk about 100 meters (325ft) with or without resting (>2 x FS = 3+)
6.5	Constant bilateral assistance (canes, crutches, braces) required to walk about 20 meters (65ft; >2 x FS = 3+)
7.0	Unable to walk beyond about 5 meters (16ft) even with aid, essentially restricted to wheelchair, wheels self in standard wheelchair a full day and transfers alone, up and about in wheelchair some 12 hours a day (>1 FS = 4+, very rarely pyramidal grade 5 alone)
7.5	Unable to take more than a few steps, restricted to wheelchair, may need aid in transfers, wheels self but cannot carry on in standard wheelchair a full day, may require motorized wheelchair (>1 x FS = 4+)
8.0	Essentially restricted to bed or chair or perambulated in wheelchair, but may be out of bed much of the day, retains many self-care functions, generally has effective use of arms (FS = 4+ in several systems)
8.5	Essentially restricted to bed for much of the day, has some effective use of arm(s), retains some self-care functions (FS = 4+ in several systems)
9.0	Helpless bed patient, can communicate and eat (usual FS ≥4)
9.5	Totally helpless bed patient, unable to communicate effectively or eat/swallow (FS = 4+)
10	Death due to MS

FS = functional system; MS = multiple sclerosis.



A (table to the left) shows the Expanded Disability Staus Scale (EDSS). The 20 ordered categories reflect increasing multiple sclerosis (MS) disability, and are assigned, by the author, sequential half-point scores (0–20, except there is no 0.5). B shows how the EDSS marks out the MS disability variable. Each of the 20 categories represents a range on the MS disability continuum. C represents the MS disability ruler and the marks represent the points of transition between adjacent categories; that is, the points at which a person's MS disability is such that they are equally likely to score either of the two categories (e.g. 0 or 1). The fact that the categories have sequential integer fractions implies that the categories represent equal amounts of MS disability and, therefore, that a 0.5-point change or difference has the same meaning in terms of underlying variable (MS disability) anywhere on the continuum. This is also represented by giving each category the same 'size' in B. Clearly, this is a very unlikely assumption; D represents a more likely scenario.

Research using the new psychometric methods discussed later in the article has confirmed what we inherently know: that a one-point change in scale score varies in its meaning in terms of the health variable being measured (e.g. disability or spasticity). Worryingly, research has also shown that this variation can be dramatic: we have demonstrated variability of up to 29 times.[5] Also, the relationship between ordinal scale scores and the interval measurements they imply varies within a scale both across the range of that particular scale and between scales.

Given the above discussion, why is it common practice to analyze scale scores as if they were measurements? This can be attributed to the measurement theory underpinning the most widely used 'traditional' psychometric methods for analyzing rating scale data and determining rating scale reliability and validity. This theory, known as Classical Test Theory (CTT), stems from Spearman's work in the early 1900s[26] and postulates that the number a person scores on a rating scale (their 'observed score,' or O) is the sum of that person's unobservable measurement that we are trying to estimate ('true score,' or T) and some associated measurement 'error' (E).

This simple theory with its associated assumptions expanded to form the methods for testing reliability and validity known as traditional psychometric methods.[27] However, the fact that these methods are derived from CTT means that their appropriateness requires that the theory and assumptions of CTT are supported by the data. If these requirements are not met, the conclusions arising from the data analysis may be incorrect. This is where the problems lie: CTT is a theory that cannot be tested, verified, or—more importantly—falsified in any data set,[28] as T and E cannot be determined in a way that enables evaluation of their accuracy.[29,30]

This has four important implications. First, untestable measurement theories are, by definition, weak theories enabling only weak inferences about rating scale performance and the measurements of people. Second, theories that cannot be challenged are easily satisfied by data sets.[29,30] Third, as T scores cannot be estimated from O scores in a way that enables their accuracy to be checked, only the observed data (ordinal scores) are available for analysis. Finally, the equation derived from CTT for computing confidence intervals around individual person scores gives large values, indicating a lack of confidence in comparing changes and differences at the individual person level. As such,

therefore, CTT has been called Weak-True Score Theory,[29,30] a tautology,[31] and a theory that has no theory.[28]

Solution 1—New Psychometric Methods

The solution to the first problem is to use new psychometric methods when constructing and evaluating rating scales and when analyzing rating scale data. These methods, known as Item Response Theory (IRT)[30,32–34] and Rasch measurement,[11,35–37] constitute the 'something (that) must be done to turn counts into measurements' we mentioned above. Essentially, new psychometric methods are mathematical models that articulate the conditions (measurement theories) under which equal interval measurements can be estimated from rating scale data. Thus, when rating scale data satisfy (fit) the conditions required by these mathematical models, the estimates derived from the models are considered robust because the measurement theory is supported by the data. When data do not fit the chosen model, two directions of inquiry are possible. Essentially, albeit simply stated, when the data do not fit the chosen model, the IRT approach is to find a mathematical model that best fits the observed item response data; in contrast, the Rasch measurement approach is to find data that better fit one model (the Rasch model). Thus, it follows that proponents of IRT use a family of item-response models, while proponents of Rasch measurement use only one model (Rasch model).

> There is no doubt that both Item Response Theory and Rasch measurement offer substantial advantages over Classical Test Theory for neurology research.

There is no doubt that both IRT and Rasch measurement offer substantial advantages over CTT for neurology research. Other advantages, beyond the scope of this article, include item banking, scale equating, computerized scale administration, and the handling of missing data. As such, clinicians should be actively looking to apply these methods in the future. However, which approach is better, and does it matter which approach is used?

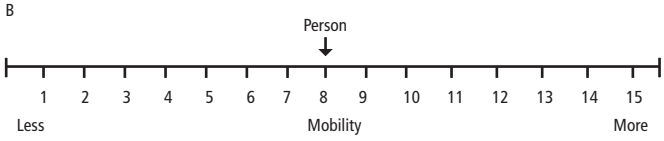
The answer to both questions depends on which central philosophy is followed, as this divides proponents of IRT and Rasch measurement. As IRT prioritizes the observed data, it sees the Rasch perspective of using only one model as too restrictive and the 'selection' of data to meet that model as threatening to content validity.[38,39] As Rasch measurement prioritizes the mathematical model, it sees the process of modeling data as precluding the ability to achieve core requirements of measurement, too accepting of poor quality data, and threatening to construct validity. Not surprisingly, it has been suggested that IRT and Rasch measurement have irreconcilable differences, and the two groups have come into conflict regarding which approach is preferable.[40]

Figure 3: The Rivermead Mobility Index Also Represented As a 'Ruler'

A

Please tick 'No' or 'Yes' for each question	No	Yes
1. Turning over in bed		
Do you turn over from your back to your side without help?	☐	✔
2. Lying to sitting		
From lying in bed, do you get up to sit on the edge of bed on your own?	☐	✔
3. Sitting balance		
Do you sit on the edge of the bed without holding on for 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 bed to chair and back 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 meters inside with no calliper, splint, or aid, and no standby help?	☐	✔
11. Picking off the floor		
If you drop something on the floor, do you manage to walk 5 meters, 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 meters without limping in four seconds (fast walk is acceptable)?	✔	☐
Score (total number of 'Yes' responses) = 8		

B



A shows the Rivermead Mobility Index (RMI), a 15-item scale for measuring mobility. Each item, a mobility-related task, has two response categories: 'No' = I am unable to do this task (scores 0) or 'Yes' = I am able to do this task (scores 1). The RMI is clinician-scored by interview and observation. Item scores are summed to give a total score that ranges from 0 (all 'No' responses) to 15 (all 'Yes' responses). This number (here = 8) is taken to be a 'measure' of the variable quantified by the set of items (mobility). B illustrates the RMI mobility variable as a 'ruler' of a count up to 15 points, which represents the number of questions that can be ticked 'Yes.' For convenience, the items are ordered sequentially at roughly equal intervals. The mark represents the location of a person who scored 8 on the mobility variable.

Problem 2—Exactly What Do Scales Measure?

Pivotal clinical trials obviously require rating scales that measure the health constructs they purport to measure (i.e. are valid) and health constructs that are clinically meaningful and interpretable. Unfortunately, current methods of establishing rating scale validity rarely enable these goals to be confirmed. To appreciate this opinion, some scale basics must

be recapped. When a set of items is used as a scale, a claim is being made that a construct is being measured.[41] Implicit to this claim is some theory of the construct being measured (a construct theory).[42] For example, the RMI (see *Figure 3*) uses a set of 15 items. It makes a claim that mobility is being measured. As such, there must be some theory of mobility underpinning the use of these specific 15 items. It follows that the aim of validity testing is to establish the extent to which a specific construct is being measured and, by implication, the extent to which the construct theory is supported.

> Statistical tests of scale validity are more formal than their non-statistical counterparts, but remain weak evaluations of the extent to which a set of items measures a construct.

Current methods for establishing scale validity cannot achieve these aims because they do not include formal methods for defining and testing construct theories.[42] While scales (e.g. the RMI) and the constructs they purport to measure (e.g. mobility) always have names, they are rarely underpinned by a theory of the construct being measured that has been deduced. Thus, there are rarely construct theories to test formally. History has proved that proposing and challenging theories is central to scientific development.[43,44]

This situation seems surprising as explicit definitions of constructs would seem to be pre-requisites for establishing scale validity. It has arisen, in part, because the constructs measured by many scales are determined during their development. Typically, scale developers generate a large pool of items, group them into potential scales, either statistically or thematically, decide what construct each group seems to measure, and then remove unwanted or irrelevant items. The main limitation of this approach is that the scale content, rather than the construct intended for measurement, defines what the scale measures. Neither grouping items statistically nor thematically ensures that the items in a group measure the same construct, but this does explain why items such as 'having trouble meeting the needs of my family' and 'few social contacts outside the home' appear in scales purporting to measure mobility and fatigue, respectively. Furthermore, both methods of grouping items avoid the process of defining, conceptualizing, and operationalizing variables, which is central to valid measurement.[45–48]

Even if the circumstances were different, and scales were underpinned by explicit construct theories, current methods of validity testing would not enable those theories to be tested adequately. Why? Because current methods, which integrate evidence from non-statistical and statistical tests, provide circumstantial evidence at best that a set of items is measuring a specific construct.

Non-statistical tests of validity typically consist of assessments of content and face validity. Content validation assesses whether scale development sampled all the relevant or important content or domains[49] and used 'sensible methods of scale construction' and a 'representative collection of items.'[50] Face validation assesses whether the final scale looks, on the face of it,[49] like it measures what is intended.[50] In the middle of the last century, Guilford named these evaluations 'validity by assumption' and 'faith validity,'[51] yet they remain essentially unchallenged.

Statistical tests of scale validity are more formal than their non-statistical counterparts, but remain weak evaluations of the extent to which a set of items measures a construct. For example, examinations of internal construct validity (e.g. factorial validity, internal consistency)[52] test the extent to which the items of a scale are related statistically. This does not confirm that a set of items marks out a clinically meaningful variable of interest, let alone tell us what a scale measures.

Examinations of external construct validity (e.g. correlations with other measures,[53,54] testing known group differences,[55] hypothesis testing[52,53]) assess the extent to which scale scores 'behave' as predicted and seek to determine whether a scale 'does what it is intended to do.'[21] These tests, which focus on person scores and between-person variation in those scores, are weak because there is no independent means of assessing the extent to which the intention of the scale is attained.[56] Consequently, these validation techniques entail circular reasoning,[56] generate only circumstantial evidence of validity,[31] enable limited development of construct theories, and result in 'primitive' understandings of exactly what is being measured.[42] Like their non-statistical counterparts, they have remained essentially unchallenged for decades.

Solution 2—Theory-referenced Measurement

Two things are needed to advance our understanding of precisely what scales measure: explicit theories of the constructs being measured, and explicit methods of testing those theories. Over the last 25 years, a number of groups have addressed these issues.[42,56–59,60,61] One group in particular has developed their ideas to an advanced level.[42,56,59] However, their work is largely inaccessible to clinicians as it concerns the measurement of reading ability. A review of that work is illuminating.

> Two things are needed to advance our understanding of precisely what scales measure: explicit theories of the constructs being measured, and explicit methods of testing those theories.

The central premise of this group's approach is a change in focus from studying people to studying items.[42] An example helps to make this idea tangible. The Lexile system is a scale for measuring people's reading ability. The items of the scale are passages of text with different levels of readability (reading difficulty). Responses to the items are scored to give a measure of reading ability. Theories suggest that the reading difficulty of a passage of text is determined by the frequency of its words as they are used in everyday communications and sentence length. Empirical studies support this construct theory by showing that these two item characteristics (word frequency and sentence length) combine to form a construct specification equation consistently explaining >80% of the variation in item location (text difficulty).[59]

New Harris poll reveals groundbreaking data...

64%

of people with MS reported losing balance, trouble walking, or inability to walk at least twice a week.[1]

Yet...

39%

of people with MS who were surveyed reported that they rarely or never discuss mobility issues with a physician.[1]

Visit www.msmobility.org to learn more.

ACORDA® THERAPEUTICS

[1] **Source:** Harris Interactive. *Experiences with Multiple Sclerosis (MS): Perspectives of People with MS and MS Care Partners* [poll]. Poll commissioned by: Acorda Therapeutics, Inc. and the National MS Society. March 25, 2008.

Construct specification equations are developed by regression analysis of item locations (here text difficulty) on selected item characteristics (here word frequency and sentence length). They afford a test of fit between scale-generated observations and theory.[56] 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 allow different construct theories to be articulated and challenged, thus enabling dynamic interplay between theory and scale[42] and a thorough investigation of individual items to aid item development and selection.

So What Next?

There are three key steps neurologists can take right now to help improve the rating scales used in neurology. First, more neurologists need to be formally trained in rating scale methods to ensure that health measurement develops clinically meaningful scales. Second, awareness of the critical role played by rating scales must increase, thus neurologists who are also journal editors, reviewers, and involved with grant-giving bodies should build links, or have direct access to, people with expertise in rating scale development and evaluation. Third, neurologists already involved in rating scale research should begin to aspire to new methodologies, such as Rasch measurement and theory-referenced measurement.

We hope the arguments in this article have helped to illustrate some of the current problems and potential solutions in using rating scales in clinical studies of neurology. Although we have only touched upon the value of new psychometric methods and theory-referenced measurement, we feel that these new avenues have much to offer all neurological outcome measurement, state-of-the-art clinical trials, and, most importantly, the individual patients that neurologists treat. We hope that neurologists interested in conducting rating scale research will use this article as a springboard to finding out more about new developments in this rapidly growing area. ■

1. Hobart J, Cano S, Zajicek J, Thompson A, Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations, *Lancet Neurol*, 2007;6:1094–1105.
2. Food and Drug Administration, Patient reported outcome measures: use in medical product development to support labelling claims, 2006.
3. Revicki D, FDA draft guidance and health-outcomes research, *Lancet*, 2007;369:540–42.
4. Kasner SE, Clinical interpretation and use of stroke scales, *Lancet Neurol*, 2006;5(7):603–12.
5. Hobart J, Cano S, Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods, Monograph for the UK Health Technology Assessment Programme, in press.
6. Wright BD, Masters G, *Rating scale analysis: Rasch measurement*, Chicago: MESA, 1982.
7. Hobart JC, Rating scales for neurologists, *J Neurol Neurosurg Psychiatry*, 2003;74(Suppl. IV):iv22–iv26.
8. Kurtzke JF, Rating neurological impairment in multiple sclerosis: an expanded disability status scale (EDSS), *Neurology*, 1983;33: 1444–52.
9. Ashworth B, Preliminary trial of carisoprodol in multiple sclerosis, *Practitioner*, 1964;192:540–42.
10. Rankin J, Cerebral vascular accidents in patients over the age of 60: II. Prognosis, *Scott Med J*, 1957;2:200–15.
11. 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.
12. Hoehn MM, Yahr MD, Parkinsonism: onset, progression, and mortality, *Neurology*, 1967;17:427–42.
13. Mahoney FI, Barthel DW, Functional evaluation: the Barthel Index, *Md State Med J*, 1965;14:61–5.
14. Granger CV, Hamilton BB, Sherwin FS, *Guide for the use of the uniform data set for medical rehabilitation*, Buffalo: Research Foundation of the State University of New York, 1986.
15. 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.
16. Collen FM, Wade DT, Robb GF, Bradshaw CM, Rivermead Mobility Index: a further development of the Rivermead Motor Assessment, *Int Disabil Stud*, 1991;13:50–54.
17. Nunnally JC, *Psychometric theory. 1st ed.*, New York: McGraw-Hill, 1967.
18. Bridgeman P, *The logic of modern physics*, New York: Macmillan, 1927.
19. Michell J, Measurement: a beginner's guide, *J Appl Meas*, 2003;4(4):298–308.
20. Michell J, *An introduction to the logical of psychological measurement*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1990.
21. Michell J, Measurement scales and statistics: A clash of paradigms, *Psychol Bull*, 1986;100(3):398–407.
22. Wright BD, Linacre JM, Observations are always ordinal: measurements, however must be interval, *Arch Phys Med Rehabil*, 1989;70:857–60.
23. Thorndike EL, *An introduction to the theory of mental and social measurements*, New York: The Science Press, 1904.
24. Thurstone LL, Theory of attitude measurement, *Psychol Rev*, 1929;36:222–41.
25. Merbitz C, Morris J, Grip J, Ordinal scales and foundations of misinference, *Arch Phys Med Rehabil*, 1989;70:380–12.
26. Traub R, Classical Test Theory in historical perspective, *Educational Measurement: Issues and Practice*, 1997(winter):8–14.
27. Novick MR, The axioms and principal results of classical test theory, *J Math Psychol*, 1966;3:1–18.
28. Lord FM, *Applications of item response theory to practical testing problems*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.
29. Allen MJ, Yen WM, *Introduction to measurement theory*, Monterey, California: Brooks/Cole, 1979.
30. Lord FM, Novick MR, *Statistical theories of mental test scores*, Reading, Massachusetts: Addison-Wesley, 1968.
31. Massof R, The measurement of vision disability, *Optom Vis Sci*, 2002;79:516–52.
32. Hambleton RK, Swaminathan H, *Item response theory: principles and applications*, Boston, Massachussets: Kluwer-Nijhoff, 1985.
33. Lord F, A theory of test scores, *Psychometric Monographs*, 1952; no. 7.
34. Lord FM, The relation of the reliability of multiple-choice tests to the distribution of item difficulties, *Psychometrika*, 1952;17(2): 181–94.
35. Wright BD, Solving measurement problems with the Rasch model, *Journal of Educational Measurement*, 1977;14(2):97–116.
36. Andrich D, A rating formulation for ordered response categories, *Psychometrika*, 1978;43:561–73.
37. Wright BD, Stone MH. Best test design: Rasch measurement. Chicago: MESA, 1979.
38. Cook K, Monahan P, McHorney C, Delicate balance between theory and practice, *Med Care*, 2003;41(5):571–4.
39. Fisher W, The Rasch debate: Validity and revolution in education measurement. In: Wilson M (ed.), *Objective measurement: Theory into practice*, Norwood, NJ: Ablex, 1992.
40. Andrich D, Controversy and the Rasch model: a characteristic of incompatible paradigms?, *Med Care*, 2004;42(1):I7–I16.
41. Cronbach LJ, The two disciplines of scientific psychology, *Am Psychol*, 1957;12:671–84.
42. Stenner AJ, Smith M, Testing Construct theories. *Percept Mot Skills*, 1982;55:415–26.
43. Popper K, *The Logic of Scientific Discovery*, London: Routledge, 1992.
44. Kuhn TS, *The structure of scientific revolutions*, Chicago: University of Chicago Press, 1962.
45. 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.
46. Andrich D, A framework relating outcomes based education and the taxonomy of educational objectives, *Studies in Educational Evaluation*, 2002;28:35–59.
47. Andrich D, Implication and applications of modern test theory in the context of outcomes based research, *Studies in Educational Evaluation*, 2002;28:103–21.
48. 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(1):224–34.
49. Streiner DL, Norman GR, *Health measurement scales: a practical guide to their development and use. 2nd ed.*, Oxford: Oxford University Press, 1995.
50. Nunnally JCJ, *Introduction to psychological measurement*, New York: McGraw-Hill, 1970.
51. Guilford JP, *Psychometric methods. 2nd ed.*, New York: McGraw-Hill, 1954.
52. Bohrnstedt GW, Measurement. In: Rossi PH, Wright JD, Anderson AB (eds), *Handbook of survey research*, New York: Academic Press, 1983:69–121.
53. Cronbach LJ, Meehl PE, Construct validity in psychological tests, *Psychol Bull*, 1955;52(4):281–302.
54. Campbell DT, Fiske DW, Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychol Bull*, 1959;56(2): 81–105.
55. Kerlinger FN, *Foundations of behavioural research. 2nd ed.*, New York: Holt, Rinehart, and Winston, 1973.
56. Stenner AJ, Smith M, Burdick D, Towards a theory of construct definition, *Journal of Educational Measurement*, 1983;20(4): 305–16.
57. 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*, Mahwah, NJ: Lawrence Erlbaum Associates, 2002.
58. Embretson SE, A cognitive design system appraoch to generating valid tests: application to abstract reasoning, *Psychol Methods*, 1998;3(3):380–96.
59. Stenner AJ, Burdick H, Sandford EE, Burdick DS, How accurate are lexile text measures?, *J Appl Meas*, 2006;7(3):307–22.
60. Stone MH, *Knox cube test – revised*, Itasca, IL: Stoelting, 2002.
61. Stone MH, Wright BD, Stenner AJ, Mapping variables, *J Outcome Meas*, 1999;3(4):308–22.