Rating Scales for Clinical Studies in Neurology—Challenges and Opportunities

Jeremy Hobart, PhD, FRCP and Stefan Cano, PhD

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

Rating scales are increasingly used as primary or secondary outcome measures in clinical studies in neurology. 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).1,2 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 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 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).1,2 However, the majority of clinical studies in neurology that use rating scales are currently inadequate. Two simple examples illustrate some of the key issues.

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 review1 and forthcoming monograph.3

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.4 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.1

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.2 Figure 2 shows how single-item scales, such as the Kurtzke’s Expanded Disability Status Scale (EDSS),3 mark out the variable they purport to measure. Other widely used single-item scales include Ashworth’s scale for spasticity,4 the modified Rankin scale,5 Hauser’s Ambulation index,6 and the Hoehn and Yar scale.7

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,8 Functional Independence Measure,9 Multiple Sclerosis Walking Scale).10 Figure 3 shows the Rivermead Mobility Index (RMI)11 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.

![Image of rating scales](image_url)

**Figure 2: Central Features of All Measurements**

<table>
<thead>
<tr>
<th>A</th>
<th>Less</th>
<th>MS disability</th>
<th>More</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td>Person</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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. This is because a fundamental requirement of the definition of ‘measurement’ is a constant unit.12

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,13 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.14-21 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).

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.

US NEUROLOGY
Multiple Sclerosis

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

<table>
<thead>
<tr>
<th>Grade</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Normal neurological examination (FS = 0; cerebral grade 1 acceptable)</td>
</tr>
<tr>
<td>1.0</td>
<td>No disability (FS &gt; 1 excluding cerebral grade 1)</td>
</tr>
<tr>
<td>1.5</td>
<td>No disability (FS &gt; 1 excluding cerebral grade 1)</td>
</tr>
<tr>
<td>2.0</td>
<td>Minimal disability (1 x FS = 2, others 0/1)</td>
</tr>
<tr>
<td>2.5</td>
<td>Minimal disability (2 x FS = 2, others 0/1)</td>
</tr>
<tr>
<td>3.0</td>
<td>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)</td>
</tr>
<tr>
<td>3.5</td>
<td>Fully ambulatory = moderate disability (1 x FS = 3 + 1/2 x FS = 2) or 2 x FS = 3, others 0 or 1, able to walk without aid or rest some 500 meters (0.3 miles)</td>
</tr>
<tr>
<td>4.0</td>
<td>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 for some 300 meters (0.2 miles)</td>
</tr>
<tr>
<td>4.5</td>
<td>Fully ambulatory without aid, 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 (0.2 miles)</td>
</tr>
<tr>
<td>5.0</td>
<td>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)</td>
</tr>
<tr>
<td>5.5</td>
<td>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)</td>
</tr>
<tr>
<td>6.0</td>
<td>Intermittent or constant unilateral assistance (cane, crutch, brace) required to walk about 100 meters (325ft) with or without resting (&gt;2 x FS = 3+)</td>
</tr>
<tr>
<td>6.5</td>
<td>Constant bilateral assistance (cane, crutches, braces) required to walk about 20 meters (65ft; &gt;2 x FS = 3+)</td>
</tr>
<tr>
<td>7.0</td>
<td>Unable to walk beyond about 5 meters (16ft) even with aid, essentially restricted to wheelchair, wheelchair self in standard wheelchair a full day and transfers alone, up and about in wheelchair some 12 hours a day (&gt;1 FS = 4+, very severe pyramidal grade 5 alone)</td>
</tr>
<tr>
<td>7.5</td>
<td>Unable to take more than a few steps, restricted to wheelchair, may need aid in transfers, wheelchair self but cannot carry on in standard wheelchair a full day, may require motorized wheelchair (&gt;1 x FS = 4+)</td>
</tr>
<tr>
<td>8.0</td>
<td>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)</td>
</tr>
<tr>
<td>8.5</td>
<td>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)</td>
</tr>
<tr>
<td>9.0</td>
<td>Helpless bed patient, can communicate and eat (usual FS = 4+)</td>
</tr>
<tr>
<td>9.5</td>
<td>Totally helpless bed patient, unable to communicate effectively or eat/swallow (FS = 4+)</td>
</tr>
<tr>
<td>10</td>
<td>Death due to MS</td>
</tr>
</tbody>
</table>

FS = functional system; MS = multiple sclerosis.

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 that the variation can be as much as 29 times.5 Essentially restricted to bed for much of the day, has some effective use of arms (FS = 4+), retains some self-care functions (FS = 4+ in several systems) or 5 x FS = 2, others 0/1) or mild disability in three or four FS (3/4) x FS = 2, others 0/1) or mild disability in three or four FS (3/4) x FS = 2, others 0/1).

Given the above discussion, why is it common practice to analyze scale scores and the interval measurements they imply varies within a scale both across the range of that particular scale and between scales. As such,

rating scale reliability and validity. This theory, known as Classical Test Theory (CTT), stems from Spearman’s work in the early 1900s26 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.31,32 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, and a theory that has no theory. 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) and Rasch measurement, 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 IRT and Rasch measurement offer substantial advantages over Classical Test Theory for neurology research. 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 (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 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. 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 irrevocable differences, and the two groups have come into conflict regarding which approach is preferable.

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. Implicit to this claim is some theory of the construct being measured (a construct theory). 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. 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.

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.

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 and used ‘sensible methods of scale construction’ and a ‘representative collection of items’. Face validation assesses whether the final scale looks, on the face of it, like it measures what is intended. In the middle of the last century, Guilford named these evaluations ‘validity by assumption’ and ‘faith validity’, 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) 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, testing known group differences, hypothesis testing) assess the extent to which scale scores ‘behave’ as predicted and seek to determine whether a scale ‘does what it is intended to do’. 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. Consequently, these validation techniques entail circular reasoning, generate only circumstantial evidence of validity, enable limited development of construct theories, and result in ‘primitive’ understandings of exactly what is being measured. 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.

The central premise of this group’s approach is a change in focus from studying people to studying items. 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).
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.

¹Source: 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.

©2008 Acorda Therapeutics, Inc. All rights reserved. May 2008 M80001
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. 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 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.

---

5. Hobart JC, C, Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods, Monograph for the UK Health Technology Assessment Programme, in press.
12. Hoehn MM, Yahr MD, Parkinsonism: onset, progression, and evaluation. Third, neurologists already involved in rating scale research the 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.