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

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 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.2,4 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.2,4 The FDA requirements are likely to be emulated by the European Medicines Agency (EMEA) 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.37 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.
**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.\(^{19}\)

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]),\(^{19}\) Kurtzke’s expanded disability status scale [EDSS],\(^{20}\) modified Rankin scale,\(^{21}\) Hauser ambulation index,\(^{22}\) and Hoehn and Yar scale\(^{23}\) and multiple item scales (eg, Rivermead mobility index [table 1, figure 2])\(^{19}\) Barthel index,\(^{24}\) and functional independence measure\(^{25}\).

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.\(^{22,23}\) 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\(^{26}\) 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.\(^{25–26}\)

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: 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.\(^{24,25–26}\)

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?

**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 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,\(^{25–26}\) their continued use as outcome measures in clinical trials\(^{26–27}\) 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),\(^{26}\) 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.

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.\(^{22,46–48}\)

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 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.\(^{44,46}\) 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.\(^{44,46}\) 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.

---

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

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

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).

---
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.
Another related problem with ordinal scales 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 $+\sim 3.5$ and $+\sim 15.4$ points, respectively, which equate to more than 30% of their respective total scale ranges. Consequently, clinical trials cannot legitimately compare changes and differences among individuals, 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). A measurement theory is a theory of how the numbers generated by rating scales relate to measurements of the constructs they seek to estimate; classical test theory (CTT). 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.

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. 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—validated in any dataset because the parameters of the theory (T and E) cannot be determined in a way that enables the evaluation of their accuracy.

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. 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, a tautology and a theory that has no theory.

Approaches to overcome the limitations of ordinal scales

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. 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, 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;64,65 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; when this is achieved, the debate about what statistics are permissible becomes redundant.

Latent trait theories

The value of being able to construct measurements with equal intervals from ordinal rating scale data,19 and the need to develop strong measurement theories,49 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.64,65,78 The main property is the ability to test for, and thus achieve, invariance (stability).65 This is achieved because the central tenet of measurement is that the instrument of measurement is

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.
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.47,78

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.81 Not surprisingly, IRT and Rasch measurement are suggested to have irreconcilable differences,82 and the two groups have come into conflict about which approach is preferable.83-85

The requirement to know precisely what variables are measured

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.91,92 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 validity93 (eg, factorial validity94 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.98,99 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.97

Testing convergent and discriminant construct validity96 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.\textsuperscript{a} Consequently, these validation techniques are based on circular reasoning,\textsuperscript{32} generate circumstantial evidence,\textsuperscript{43} enable only limited development of construct theories, and result in only a basic understanding of what is being measured.\textsuperscript{41} 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,\textsuperscript{85,98–103} one group has developed their ideas to an advanced level.\textsuperscript{85,98} 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.\textsuperscript{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.\textsuperscript{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).\textsuperscript{46}

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.\textsuperscript{66} 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;\textsuperscript{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).

**Recommendations**

The FDA draft recommendations for patient-reported rating scales in clinical trials highlight the importance of “conceptually sound, reliable, and valid measures”.\textsuperscript{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.\textsuperscript{30–34} 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.

**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\textsuperscript{10} or report scores for scales that are invalid.\textsuperscript{10–13} Furthermore, scales continue to be developed without adequate recourse to recognised methods for scale construction.\textsuperscript{10} 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
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,43,56,73,115,116 many reports104–109 and standard measurement texts79,82,83,118 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,133 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
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. SJG 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. SJG 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 for educational and consulting activities from Eisai, Bayer, and Oxford Outcomes. JPZ has received honoraria for lectures. This review was not supported by a corporate sponsor.

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

References