Patient Preference and Adherence

Dovepress
open access to scientific and medical research

Open Access Full Text Article

REVIEW

The problem with health measurement

This article was published in the following Dove Press journal:
Patient Preference and Adherence
13 June 2011
Number of times this article has been viewed

Stefan J Cano
Jeremy C Hobart

Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, Tamar Science Park, Plymouth, UK

Abstract: In this review we discuss health measurement with a focus on psychometric methods and methodology. In particular, we examine some of the key issues currently facing the use of clinician and patient rating scales to measure the health outcomes of disease and treatment. We present three key facts and flag one crucial problem. First, the numbers generated by scales are increasingly used as the measurements of the central dependent variables upon which clinical decisions are frequently made. The rising profile of rating scales has significant implications for scale construction, evaluation, and selection, as well as for interpreting studies. Second, rating scale science is well established. Therefore, it is important to learn the lessons from those who have built and established the science over the last century. Finally, the goal of a rating scale is to measure. As such, over the last half century, developments in rating scale (psychometric) methods have caused a refocus in the way we should be measuring health. In particular, newer methods have significant clinical advantages over traditional approaches. These should be seriously considered for inclusion in everyday practice. This leads us to the central problem with health measurement, which is that we cannot currently be sure what most rating scales are measuring. This is because the methods we have in place to ensure the validity of rating scales fall short of what is actually required. We expand on this point, and provide some potential routes forward to help address this important problem.

Keywords: patient-reported outcome instruments, health-related quality of life, psychometrics, questionnaires, outcome assessment, health care

Introduction

Health measurement is increasingly at the heart of the agenda for high-stakes clinical research, trials, and practice,[1–3] which directly influences decisions about patient care and policy-making.[4] This rise in profile has been accompanied by an increased interest in rating scale science.[2,3] There are now growing numbers of clinical researchers who are either developing or using rating scales to quantify the effects of disease or treatment on abstract concepts, such as ability, emotional well-being, or memory. For example, the MAPI Trust, a nonprofit organization providing information on patient rating scales, houses over 3000 scales.[5]

Over the last 16 years we (SC, JH) have worked as health measurement researchers. We have been fortunate enough to have been involved in a wide range of clinical[6,7] and surgical[8,9] areas, have tested and developed a number of clinician-report[10,11] and patient-report rating scales,[12,13] and have used traditional and modern rating scale techniques.[14] Our main interest lies in the science that underpins health measurement, also known as psychometrics.[15] During our working careers, we have witnessed great progress relating

Correspondence: Stefan Cano
Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, Room N13 ITTC Building, Tamar Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK
Tel +44 017 5231 5245
Fax +44 017 5231 5254
Email stefan.cano@pms.ac.uk

to the application of psychometrics to the development of rating scales, and the development of documents containing key guidelines[16,17] and high-level requirements.[2,3]

However, we have also witnessed concerning problems in the field. Thus, despite the proliferation of rating scales in health measurement, many scales have not been psychometrically validated in an appropriate way.[18–22] This has wide-reaching effects. For example, despite the increased inclusion of rating scales in current "state-of-the-art" clinical research and trials, the same studies continue to use scales that have been proved to be scientifically wanting. This is demonstrated through even the most superficial of literature reviews, ie, a brief literature search in PubMed focusing on randomized controlled Phase III and IV trials in multiple sclerosis published in 2006–2011. This reveals that half of the 28 relevant articles used a rating scale, but only two articles include scales that have any supporting psychometric evidence. Parallels can be seen throughout neurology,[11,23] and our experience working in other clinical disciplines suggests that these problems are not uncommon.

Given the increasing importance of rating scale data, we strongly believe that rating scales should provide scientifically robust results. However, the problem with health measurement runs deeper than psychometric "validation". In order to understand why, we need to step back initially and provide some background and context. So, in this review, we explore health measurement, beginning with key concepts, followed by some important historical landmarks, then move on to the development and application of psychometric methods, finishing with some of the pressing issues of the current time. Health measurement covers a lot of ground. Of course it would be impossible to discuss all aspects of the area. So, before we get started, it is important to clarify what we will not be discussing here, but, given the omissions, why we believe our title is appropriate.

First, we do not include discussions on health economics, clinimetrics, or specific aspects of psychometric testing. In relation to health economics, the extent to which this falls under the remit of health measurement per se is debatable, but more importantly, this in itself is a large area that deserves its own review. For those interested in our views, we discuss health econometrics more fully elsewhere.[9]

In relation to clinimetrics, we would point readers to another of our publications, in which we provide a perspective on Feinstein's contribution to the health measurement debate.[23] For now, we would say that in this review we focus on the "measurement" part of health measurement. In particular, we discuss rating scales when they are used as measurement instruments to quantify variables of interest (eg, ability, depression, short-term memory) via patient self-report or clinician report. We do not discuss rating scales when they are used for other purposes, such as checklists, clinical assessment tools, methods of predicting outcome, structured interviews, or other methods for gathering information (eg, surveys). This is because terms such as evaluation, assessment, and measurement are often used interchangeably. However, measurement has a very specific meaning with respect to quantifying attributes (ie, a characteristic, or property belonging to a person).[24] In contrast, evaluation and assessment are often qualitative processes.

Finally, we do not include a review (or appraisal) of specific psychometric tests, because once again this deserves its own review, given the size of the area and the issues. For those readers who would like to learn more, we have previously published a monograph that examines, in detail, the key tests used in traditional and modern psychometric methods.[14]

Why then, given that health measurement encompasses such a wide area, and has potentially many good and bad points, do we believe that our title is appropriate? In order to answer this question we must anticipate the punch line of our review. Thus, we believe that the cornerstones of health measurement are the instruments used to measure the target variables of interest. For these instruments to be fit for purpose they must provide clinically useful, meaningful, and interpretable data. We argue that, at the present time, the extent to which the vast majority of currently available scales achieve these vital criteria is unclear at best. This presents a "house of cards" situation, ie, if we are unclear as to the exact variables that our scales are measuring, what exactly can we do with the information they provide? We would suggest this fundamental issue has serious repercussions for the whole of health measurement. However, before we expand on this, we first need to revisit some key concepts to set the scene.

Key concepts

Rating scales are used to measure unobservable (latent) variables known as theoretical constructs, which are abstract (as opposed to concrete).[25] Latent variables can be measured indirectly by asking questions intended to capture, empirically, the essential meaning of a construct. The simplest way to do this is to ask a single straightforward question, or item. However, single items are limited because they are: unlikely to represent the broad scope of a complex theoretical construct; likely to be interpreted in many different ways by respondents; imprecise because they cannot discriminate, to a fine degree, between different levels of an attribute;

and unreliable (prone to random error) because they do not produce consistent answers over time.[26] As such, rating scales are usually made up of multiple items, in which each item addresses a different aspect of the same underlying construct. Using multiple items overcomes the scientific limitations of single items because: more items increase the scope of a scale; are less open to variable interpretation; enable better precision; and improve reliability by allowing random errors of measurement to average out.[26] In this review, we use the term "rating scale" as the umbrella term to cover any instrument that conforms to a questionnaire-style structure, and is used to obtain scores, from a person's responses to statements or questions, which in turn are considered to be measurements of a given variable.

There are many methods, termed scaling models, for combining multiple items into scales, depending on the purpose the resulting scale is to serve.[27–31] The most widely used scaling model in health measurement is the method of summated ratings proposed by Likert.[32,33] Four characteristics constitute a summated rating scale. First, there are multiple items whose scores are summed, without weighting, to generate a total score. Second, each item measures a property that can vary quantitatively. Third, each item has no right answer. Fourth, each item in the scale can be rated independently. Examples of Likert scales used in health measurement include the Medical Outcomes Study 36-item Short Form Health Survey (SF-36),[34,35] General Health Questionnaire (GHQ),[36] and the Hospital Anxiety and Depression Scale (HADS).[37] The way in which developers propose that items should be combined to form a scale is called a measurement model. These models are the focus of a psychometric evaluation.

Rating scales in health measurement: a brief history

We have come a long way since Ernest Amory Codman's "end result" idea.[38] Codman was an orthopedic surgeon at the Massachusetts General Hospital, Boston, MA, during the first three decades of the 20th century.[39] His "end result idea" entailed long-term follow-up of patients to determine treatment success, and taking steps to prevent new failures if outcomes were undesirable. Although Codman has been described as one of the most important figures in the history of clinical outcomes research, the conception and development of his "idea" have been largely neglected in the history of health measurement.[38,39] It was not until after the Second World War that clinical researchers began to develop scales to measure the outcomes of procedures.

One of the first surgeons to do this was Visick, who attempted to measure the functional results of gastric surgery, focusing particularly on postprocedural complications.[40] In 1949, Karnofsky, an oncologist, developed the first "performance" measure,[41] ie, a 10-point observer-rated scale spanning the extremes of physical dependency defined by nursing burden. For many years, this scale was used widely, but often, it has been argued, inappropriately.[42] It was improved 20 years later with Katz's Activities of Daily Living Scale, which broadened the focus to wider aspects of quality of life.[43] The same period saw an increase in the development and use of new scales across medicine, with the most noticeable increase in neurology.[44] The decades following the 1960s witnessed increasing recognition of the importance of assessing a broader array of outcomes when measuring the impact of disease or evaluating the effectiveness of procedures.

During the 1970s, the focus of health care evaluation moved from traditional clinical outcomes (ie, mortality and morbidity) to the measurement of function (ie, the ability of patients to perform activities of daily living).[25] The shift from traditional outcome measures to the wider encompassing measurement of health occurred for a number reasons. First, the narrow definition of health in terms of morbidity and mortality was replaced by a broader definition of health as a "complete state of physical, mental and social well-being and not merely the absence of disease or infirmity".[45] Second, public health campaigns, rising standards of living, ageing populations, and development of health technology led to a shift in attention from the cure of acute diseases to the management of more complex, chronic conditions (eg, asthma, rheumatoid arthritis, multiple sclerosis). This led to increased interest in measuring more complex and subjective aspects of outcomes pertaining to the health impact of disease and/or treatment (for which we use the shorthand term "health outcomes" in this review). Third, there was increased demand for clinicians to demonstrate evidence of cost-effectiveness, in which the benefits of a particular health service or intervention are weighed against the costs of that service or intervention.[46]

The 1980s witnessed patient report rating scales (now known as Patient Reported Outcome [PRO] instruments) being increasingly used in clinical research, and as a result, phrases such as "quality of life" became buzz words.[47] Scales for use across different clinical populations (generic measures) were developed and became widely used, including the Sickness Impact Profile,[48] Nottingham Health Profile,[49] and SF-36.[50] The 1990s saw a proliferation of more targeted patient rating scales, including dimension-specific

(eg, mood[37]), disease-specific (eg, cancer[51]), site-specific (eg, orthopedic[52]), and individualized scales.[53] The gradual but important shift from clinical research to practice and policy[2–4] over the last decade has witnessed the proposal of even more sophisticated measuring instruments in the form of item banks.[54–56]

Rating scales in health measurement: type and kind

Philosophically, the different types of rating scales can be classified into two distinct approaches.[57,58] First, the standard needs approach describes measuring health outcomes as the extent to which certain universal needs are met. This approach advocates that there is a standard set of life circumstances that are required for optimal functioning. Although subjective phenomena, health outcomes are objective characteristics of an individual. Second, and in contrast, the psychological processes approach views health outcomes as being constructed from individual evaluations of personally salient aspects of life. This approach sees health outcomes as being made up of perception of life circumstances, dependent on the psychological makeup of an individual, rather than on their life circumstances alone. The central assumption of this approach is that each person is the best source of judgments about health outcomes, and one cannot assume that all people will value different circumstances in the same way.

Many types of rating scales can be classed as following the standard needs approach, ranging from generic scales that provide comprehensive, general evaluations of health outcomes, to those that concentrate on a specific aspect of health (eg, symptoms). The former is illustrated by the SF-36,[50] which focuses on activities of daily living (eg, personal care, domestic roles, mobility) and on role functioning (eg, work, finance, family, friends, and social). Generic measures permit direct comparisons of different patient populations, thereby providing the opportunity to make policy decisions across a variety of diseases.[59] The use of generic measures may enhance the generalizability of a study or help interpret results in a wider context. In addition, it can be argued that generic measures are likely to be robust because they are used and tested in many different settings. However, generic measures may be limited because they are may be unable to address important aspects of outcome that are affected by a particular disease, and may not be sensitive enough to detect changes in outcome which occur in response to treatment or over time.[60]

There are three types of standard needs rating scales that concentrate on a more specific aspect of health, ie, disease/condition-specific, site-specific, and dimension-specific. The most commonly used of these scales are disease/condition-specific scales, which are developed for use in a specific disease or condition. These include items that are directly relevant to the condition and, therefore, are likely to be shorter and apparently more appropriate,[59] which can help to reduce patient burden and increase acceptability.[61] Disease-specific scales ensure more comprehensive assessment of important outcome domains, and are generally more sensitive in detecting the effects of treatment on outcome and changes in outcome over time.[59]

A site-specific scale focuses on health problems in a specific part of the body, such as the Oxford Hip Score.[52] As with disease/condition-specific scales, these include fewer items and appear to be more appropriate, reducing patient burden and increasing acceptability.

A dimension-specific scale provides a comprehensive, general evaluation of one specific aspect of health, which may be applicable across different patient groups and treatments. Examples of these types of scale include the GHQ[62] and HADS[37] which focus on aspects of psychological well-being. The advantage of such measures is that they provide a more detailed assessment in the area of concern.

The main drawback of specific measures is that they do not allow comparisons between different patient groups. Therefore, it is argued that comprehensive assessment of outcome should include a combination of generic and specific measures.[59,60] Generic measures allow comparisons across studies, thus enhancing the generalizability of findings, and specific measures provide better content validity, so are generally more responsive to measuring change due to greater relevance to the specific population.

In contrast to using generic or specific rating scales with predetermined content, proponents of the psychological processes approach argue that listing items in rating scales do not capture the subjectivity of human beings and the individual structure of values. In short, prescribing items using a preordained definition of health outcome (eg, quality of life) and matching the person to the definition (ie, "goodness of fit"), does not let us know whether all the domains, pertinent and meaningful to each respondent, are included. This viewpoint prompted the development of "individualized" measures, such as the Schedule for the Evaluation of Individual Quality Of Life (SEIQoL).[53] The SEIQoL allows individuals to nominate important domains of quality of life and weight those domains in order of importance. Another, the Patient Generated Index (PGI), asks individuals to identify those aspects of life that are personally affected by health.[63] The main advantage of these measures includes a claim

for validity, given that the areas of importance are selected by the individuals involved in completing the measures. The main disadvantages are that some of these measures require trained interviewers, which translates into a need for greater resources and lower practicality. Also, it is less easy to compare data from individualized measures between patients due to the variation in each individual completed measure.[64]

Item banks can be viewed as very large "rating scales", in which patients only complete a subset of targeted items. These banks capitalize on modern psychometric methods (which we describe more fully in the next section). In essence, modern methods provide rich information about item performance not available using traditional psychometric methods, that can be used to create banks of items (up to many hundreds or thousands of items) with known characteristics. New items can then be calibrated against the best available measures to obtain scales of higher quality and better precision.[65] Item banking also makes it possible to carry out computer adaptive testing.[66] In this technique, rather than giving the same set of items to each individual, the items are selected based on ability level or other characteristics. Computer adaptive testing has already been developed in many areas including migraine, combining datasets using different outcome measures.[67]

As alluded to in this last paragraph, the increased application of rating scales in health measurement has required the introduction of more advance psychometric methods. To elaborate on this, we first need to place these "newer" methods in context.

Psychometrics in health measurement: a brief history

Psychometrics was adopted as part of health measurement in the early 1980s.[68–70] However, its scientific foundations are deeply rooted in education and psychology. In fact, its origins can be traced to the mid 1800s when psychophysicists were demonstrating that subjective judgment can be used as a valid approach to measurement.[71,72] Through the advent of the mental test movement (circa 1925–1960),[30] these ideas were taken further and, as such, Thurstone proposed the "law of comparative judgment", an approach with close connections to the psychophysical theory developed by Weber and Fechner. This demonstrated that psychophysical scaling methods could be used to measure psychological attributes accurately[27,73] and prompted the development of psychological (or psychometric) scaling methods, which are defined as procedures for constructing scales for the measurement of psychological attributes.[71] The mental test movement led to the widespread use of standardized tests (eg, educational achievement, attitudes and personality, personnel) and, at the same time, scientific interest in methods of testing led to the development of psychometrics as a prominent discipline in psychology, within which were established the cornerstones of the scientific evaluation of measures.[71,74]

As explained above, since the 1970s health care evaluation has moved towards the measurement of physical, psychological, and social functioning.[25] The importance of psychometric methods for measuring health variables was demonstrated by two related key studies conducted in the US. First, the Health Insurance Experiment[75] showed that psychometric methods could be used to generate reliable and valid measures for assessing changes in health status for both adults and children in the general population. Second, the Medical Outcomes Study[25,76] showed that psychometric methods of scale construction and data collection were successful for measuring health status in samples of sick and elderly people. Since then, the use of psychometrics has proliferated throughout health measurement.

Psychometric methods

The main psychometric approaches as related to health measurement have been classical test theory and, more recently, Rasch measurement models and item response theory. Of all three approaches, classical test theory is currently the dominant paradigm.

Classical test theory

Spearman laid down the foundations of classical test theory in 1904, when he introduced the decomposition of an observed score into a true score and an error, and showed how to estimate the reliability of observed scores.[77] It took a further 50 years before the role of classical test theory analyses became clearer[78] as an accumulation of statistical evidence to establish the scientific robustness of measures (eg, Kuder-Richardson's coefficients for internal inconsistency, Cronbach's alpha, correlations between replicated measurements). Classical test theory is grounded in the definition of measurement as proposed by Stevens (ie, "the assignment of numerals to objects or events according to some rule").[79] It is important to note that this definition differs in important respects from the more classical definition of measurement adopted throughout the physical sciences, which is that measurement is the numerical estimation and expression of the magnitude of one quantity relative to another.[80] Classical test theory is based upon analyses of raw scores that are used to test the assumptions underlying a

given measurement model, ie, that the items can be summed (without weighting or standardization) to produce a score. The key traditional measurement properties that should be considered are data quality, scaling assumptions, targeting, reliability, validity, and responsiveness. We and others describe these tests in more detail elsewhere.[2,14]

Rasch measurement methods

Georg Rasch, a Danish mathematician, was principally concerned with the measurement of individuals rather than distribution of levels of a trait in a population. He argued that the core requirement of social measurement should be the same as that in physical measurement (ie, "invariant comparison"). With this in mind, he developed the simple logistic model (now known as the Rasch model) and through applications in education and psychology, he was able to demonstrate that his approach met the stringent criteria for measurement used in the physical sciences.[81] Vitally, the Rasch paradigm differs from the traditional statistical modeling paradigm, in that the latter approach is used to describe a set of data, whereas the former aims to obtain data which fit the model.[82]

In the Rasch model, the probability of a specified response to a given item (eg, "yes"/"no") is modeled as a logistic function of the difference between the person and item parameter (ie, the higher a person's ability with respect to the difficulty of an item, the higher the probability of a correct response). When applying the Rasch model, item locations are scaled first in a process known as "item calibration". Once item locations are scaled, the person locations are measured on the same scale. Each item and person estimate has an associated standard error of measurement, which quantifies the associated degree of uncertainty.

Rasch measurement methods are able to transform ordinal summed scores into linear measurements by paired comparisons of any two persons, any two items, or any one person and one item, defined by the logarithm of the relative probabilities.[81,83,84] Essentially, observed scores are replaced by the expected probabilities of occurrence, and relative differences are computed as ratios of the relative probabilities (as these are consistent indicators of relative differences). This ratio of the relative probabilities is then expressed on a linear scale in an additive form by taking logarithms. In addition, the Rasch model is able to transform summed scores into linear measures of persons and items that are on the same scale with a common unit, and freed up from the distributional properties of each other. Thus, the Rasch model realizes, mathematically, the requirements for scientific measurement of invariant comparisons of people, and items, on the same linear scale.[81 83,84]

Rasch measurement methods use the Rasch model to evaluate the legitimacy of summing items to generate measurements, and their reliability and validity. The model articulates the set of requirements that must be met for rating scale data to generate internally valid, equal-interval measurements that are stable (invariant) across items and people.[85] The central tenet of the Rasch measurement methods is that they examine the extent to which observed data (patients' actual responses to scale items) accord with ("fit") predictions of those responses from a mathematical (Rasch) model. Thus, the difference between what should happen (expected) and what does happen (observed) indicates the extent to which rigorous measurement is achieved. Statistical and graphical tests are used to evaluate the correspondence of data with the model. Certain tests are global, while others focus on specific items or persons. There are seven key measurement properties that should be considered, ie, thresholds for item response options, item fit statistics, item locations, differential item functioning, correlations between standardized residuals, person separation index, and individual person change statistics. We describe these in more detail elsewhere.[14]

Comparison of classical test theory and Rasch measurement

Direct comparisons of classical test theory and Rasch measurement methods in the medical literature are sparse, and at best superficial.[86,87] In part, this may be due to the fact that the two approaches cannot be compared easily, because they use different methods, produce different information, and apply different criteria for success and failure.

There are four main limitations of classical test theory. First, the data generated are ordinal rather than interval, the invariance of which is unknown.[85] Second, scores for persons and samples are scale-dependent because they lack the provision for varying item parameters, resulting in item parameters that must be regarded as fixed.[88] Third, scale properties, such as reliability and validity, are sample-dependent. As such, the marginal probabilities of measures (ie, the probability distribution of scale scores) vary across population subgroups, because these subgroups may vary in the rate of the construct being measured.[11] Fourth, the data are only suitable for group studies, and are not suitable for individual patient measurement.[89]

Rasch measurement methods address each of the four limitations of classical test theory. First, the approach offers the

ability to construct linear measurements from ordinal-level rating scale data, thereby addressing a major concern of using rating scales as outcome measures.[90,91] Second, Rasch measurement methods provide item estimates that are free from the sample distribution and person estimates that are free from the scale distribution, thus allowing for greater flexibility in situations where different samples or test forms are used.[92] Therefore, the methods allow for the use of subsets of items from each scale rather than all items from the scale, yet are still able to compare scores using different sets of items. This is the foundation for item banking and computerized adaptive testing.[66] Third, Rasch measurement methods enable estimates to be obtained suitable for individual person analyses rather than only for group comparison studies.[84,93]

Criticisms of the Rasch model include it being overly restrictive because it does not permit each item to have a different discrimination and because there is no provision in the model for other parameters (eg, guessing). Some also suggest that this model is also limited by the need for unidimensional data and is too simple to match the complexity of human behavior. Further, it is complex, and classical test theory test scoring procedures are simpler to compute.[86,94–96]

Item response theory and Rasch measurement

Item response theory is another body of psychometric analysis that provides a foundation for statistical estimation of parameters that represent the locations of persons and items on a latent continuum.[97] In particular, item response theory analyses are used to ascertain the degree to which a given model and parameter estimates can account for the structure of and statistical patterns within a response dataset.[82,97] Rasch measurement methods and item response theory are mathematically similar and, therefore, are often considered as members of the same family of statistical techniques.[82,98] This is inaccurate because practitioners of Rasch measurement methods and item response theory have different research agendas.[23,82,98]

The distinction between Rasch measurement methods and item response theory is subtle but important. Item response theory models are statistical models used to explain data, and as such, the aim of an item response theory analysis is to find the statistical model that best explains the observed data.[82,98] When the observed data do not fit the chosen item response theory model, we seek another model to explain the data better. In contrast, Rasch measurement methods provide a mathematical model for guiding the construction of stable linear measures from rating scale data.[81] Therefore, the aim of Rasch measurement methods is to determine the extent to which observed rating scale data satisfy the measurement model. When the data do not fit the model, we examine the data carefully to try and explain the misfit, but ultimately we choose data that satisfies the model's requirements. This is the central tenet of the Rasch model that distinguishes it from item response theory models. Specifically, its defining property is its mathematical embodiment of the principle of invariant comparison.

The above discussion invokes two questions, ie, which approach is better and does it matter which approach is used? The answers to both questions depend on which central philosophy is followed, because this divides proponents of item response theory and Rasch measurement. Because item response theory 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.[99,100] Because 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 item response theory and Rasch measurement have irreconcilable differences,[101] and the two groups have come into conflict regarding which approach is preferable.[82,102–104]

Problem: our understanding of exactly what rating scales are measuring is limited

We hope that, in the previous sections, we have made the case for the strong scientific basis that underpins the area and the progress that has been made, especially over the last 50 years. We also hope that we have illustrated some of the potential pitfalls, especially in the selection of appropriate scales and use of appropriate psychometric methods. In fact, it is our experience that the most common disagreements in health measurement surround the issues of methods and methodology. We also expect that the debate surrounding the relative merits of competing psychometric approaches will continue. This is an issue for health measurement but, over time, and with enough discussion and clarification, we hope that this situation will improve. However, in our opinion, there is a more pressing and fundamental problem that needs to be addressed in health measurement.

The rise in profile of health measurement requires rating scales that measure the health constructs they purport to measure (ie, are valid), and health constructs that are

clinically meaningful and interpretable. Unfortunately, the current methods of establishing rating scale validity rarely enable these goals to be confirmed, because they lack formal methods for defining and testing construct theories.[105] This situation 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. Furthermore, both methods of grouping items do not adequately address the issues of defining, conceptualizing, and operationalizing constructs, which are central to valid measurement.[106–109] Even if the circumstances were different, and scales were 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 nonstatistical and statistical tests, provide circumstantial evidence at best that a set of items is measuring a specific construct.

Nonstatistical tests of validity typically consist of assessments of content and face validity. Content validation assesses whether scale development has sampled all the relevant or important content or domains,[110] uses "sensible methods of scale construction", and a "representative collection of items".[111] Face validation assesses whether the final scale looks, on the face of it,[110] like it measures what is intended.[111] Over 50 years ago, Guilford named these evaluations "validity by assumption" and "faith validity",[71] yet they remain essentially unchallenged, except, perhaps for Feinstein's contribution of clinimetrics.[24]

Statistical tests of scale validity are more formal than their nonstatistical counterparts, but remain weak evaluations of the extent to which a set of items measures a construct. For example, statistical examinations of internal construct validity (eg, factorial validity[112] and internal consistency[113]) 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.

Statistical tests of external construct validity consist of a range of examinations (including correlations with other measures,[114,115] testing known group differences,[116] and hypothesis testing[113,114]) which assess the extent to which scale scores "behave" as predicted, and seek to determine if a scale "does what it is intended to do".[74] The examination considered to provide the strongest statistical evidence of scale validity is called convergent and discriminant construct validity.[115] Here, a range of scales measuring similar and dissimilar constructs are administered to a sample. Their scores are correlated, and the pattern and magnitude of correlations are examined to determine if the scale being validated correlates better with scales measuring similar constructs than dissimilar constructs. The limitation of this approach is that showing a scale does not correlate highly with measures of a dissimilar construct tells us nothing about what the scale actually measures. Similarly, showing that a scale correlates highly with measures of similar constructs only tells us that the two are related.

A key problem with all statistical tests of validity is that they focus on person scores and between-person variation in these scores. They are weak because there is no independent means of assessing the extent to which the intention of the scale is attained.[117] Consequently these validation techniques entail circular reasoning,[117] generate only circumstantial evidence of validity,[98] enable limited development of construct theories, and result in "primitive" understandings of exactly what is being measured.[105] Like their nonstatistical counterparts, they have remained essentially unchallenged for decades.

Can we solve the problem?

Encouragingly, PRO guidelines, such as the current scientific requirements of the US Food and Drug Administration (FDA) for patient-reported rating scales in clinical trials,[2,118] highlight the importance of establishing validity. In particular, the FDA emphasizes appropriate conceptual frameworks and definitions as being fundamental. However, the FDA document provides little detailed guidance on how these can be achieved, largely because the field is poorly developed. We would argue that greater use of qualitative assessments is vital, and should include evaluating the extent to which the items of a scale map out the construct to be measured, establishing the most appropriate item phrasing, structuring and context, and cognitive debriefing to ensure consistency in meaning. In particular, we advocate the use of inductive and deductive approaches to develop explicit theories of the constructs being measured, and explicit methods of testing those theories.[105,117,119]

Rating scale development would benefit from being "bottom-up" (from a construct definition), rather than "top-down" (from a method of grouping items) to ensure that a substantive construct theory determines scale content, and

validation tests construct theories. This would require the development of robust guidelines for defining constructs and explicit definitions for content and face validity. Rating scale evaluation should fully acknowledge the equally important and complementary roles of qualitative and quantitative evaluations. In fact, scale evaluation could be considered under these two headings. The aim of qualitative evaluation could be defined as determining the extent to which the items of a scale map out 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 determining the extent to which the numbers generated by scales are measurements rather than numerals.

This analysis of scale validity implies that two things are needed, ie, explicit theories of the constructs being measured and explicit methods of testing those theories. Over the last 25 years, one group outside of health measurement has developed these ideas to an advanced level.[105,117,119] This group, led by Stenner, has argued for a change in focus of assessing validity from studying the people to the items,[105] and in particular the relationships between item characteristics and item scores. This forms the building blocks of the theory of the construct, and the validity of the construct theory becomes established when it predicts variation in item scales values. Stenner asks three key questions: Why are items ordered in a particular way? How can we explain variation in item scores, (ie, item difficulty)? What is the "something" that causes this variation?

The approach of Stenner et al is illustrated by their Lexile framework for measuring people's reading ability.[119] The reading ability continuum is mapped out by a set of items, each of which is a passage of reading text with different levels of readability (reading difficulty). People's responses to the items are scored to give a measure of their reading ability. The Lexile framework was constructed using Rasch measurement methods, thus people are measured in linear units (called Lexiles), and legitimate individual person measurement is possible. Theory suggests that the reading difficulty of a passage of text (item difficulty) is determined by two characteristics, ie, the frequency of the words as they are used in everyday written and oral communications, and the length of the sentences. These two variables combine in the form of a construct specification equation that consistently explains more than 80% of the variation in text difficulty.[119] Thus, empirical evidence strongly supports the construct theory. Stenner calls this approach "theory-referenced measurement".[119] We provide more detail about his work elsewhere.[23]

There are currently no examples of scales developed using theory-referenced measurement in health measurement, but it would not be hard to imagine instances where we could apply this approach. One example could be measuring the impact of disability. We would argue that it should be possible to take any aspect of impact (eg, upper limb functioning), and ask the same questions as Stenner's group. Thus, why are upper limb physical functioning items ordered and separated as they are? What specific item characteristics (eg, task variables) determine item difficulties (eg, task abilities)? We could identify the motor components of tasks that may characterize a theory of upper limb functioning, and examine items to identify their characteristics (variables) that account for these task difficulties. In doing so, we would begin to assemble the building blocks of a new construct theory and then move towards an appropriate construct specification equation.

Conclusion

In a 1997 editorial, Sonja Hunt, codeveloper of one of the first generic measures, ie, the Nottingham Health Profile,[49] warns us about the dangers of using quality of life instruments for decision-making: "From the perspective of scientific method it seems that there is a considerable way to go before any of the existing models or 'theories' can be considered definitive enough to justify application in the lives of patients ... where the results may be used to guide decision-making in the real world is not only unscientific, it is unethical". [47]

Fourteen years later, we find ourselves in a position where the field now stretches far beyond quality of life, into all aspects of health, and clinician-report and patient-report rating scales are being used as part of the patient decision-making process. However, in terms of the application of scientific methods to ensure that we have a clear understanding of what we are measuring, much less progress has been made. Thus, whereas we feel the intention behind the use of rating scales as health measurement tools in high stakes decision-making is well meant, we believe that there is a way to go before we can be confident that these tools are providing accurate information about their target constructs. The potential consequences in terms of rating scales misguiding patient care and misleading research, we believe, are underappreciated by clinicians and researchers.

Although construct specification equations are some way off, a move towards developing consensus guidelines to strengthen the theoretical underpinnings of new scales and the evaluation of existing scales would benefit health measurement. In particular, we would like to see greater use of qualitative assessments including: the adoption of

inductive and deductive approaches to construct theory building and development; evaluations of the extent to which the items of a scale mark out the construct to be measured; establishing the most appropriate item phrasing, structuring, and context; and cognitive debriefing to ensure consistency in meaning.

We have two key messages from our review. First, clinical researchers should be aware that there is a wealth of information regarding psychometrics out there. However, considered in isolation, psychometric statistics can be misleading. They cannot be expected to produce consistently meaningful results when considered apart from qualitative scale content evaluations. Second, establishing clinically meaningful content validity from the onset by defining, conceptualizing, and operationalizing the constructs intended to be measured is a vital step. Unfortunately, in health measurement, such strong conceptual underpinnings and therefore explicit construct theories are uncommon,[47] and clinicians, researchers, and policy makers should bear this in mind when engaging with health measurement at all levels. Stenner et al use the following analogy to describe a construct theory: "The story we tell about what it means to move up and down the scale for a variable of interest (eg, temperature, reading, ability, short-term memory). Why is it, for example, that items are ordered as they are on the item map? [This] story evolves as knowledge increases regarding the construct".[119] We would suggest that we need to be able to tell clearer and more detailed stories about what underpins our rating scales before we can start to use them confidently to make decisions about patient's lives.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Darzi A. High Quality Care for All: NHS Next Stage Review Final Report. London, UK: Department of Health; 2008.
2. Food and Drug Administration. Patient reported outcome measures: Use in medical product development to support labelling claims. Available from: www.fda.gov/cber/gdlns/prolbl.pdf. Accessed May 17, 2011.
3. Food and Drug Administration. Qualification process for drug development tools. Available from: http://www.fda.gov/cder/guidance/index.htm. Accessed May 17, 2011.
4. Department of Health. *Equity and Excellence*: *Liberating the NHS*. London, UK: Her Majesty's Stationery Office; 2010.
5. MAPI Trust. Available from: http://www.mapi-trust.org/about-the-trust. Accessed May 17, 2011.
6. Hobart J, Lamping D, Thompson A. Evaluating neurological outcome measures: The bare essentials. *J Neurol Neurosurg Psychiatry*. 1996;60:127–130.
7. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: The application of psychometric methods to clinical intuition. *Brain*. 2000;123:1027–1040.
8. Cano S, Klassen A, Pusic A. The science behind quality-of-life measurement: A primer for plastic surgeons. *Plast Reconstr Surg*. 2009;123:98e–106e.
9. Cano S, Klassen A, Scott A, Thoma A, Feeny D, Pusic A. Health outcome and economic measurement in breast cancer surgery: Challenges and opportunities. *Expert Rev Pharmacoecon Outcomes Res*. 2010;10:583–594.
10. Cano S, Hobart J, Hart P, Kolipara L, Schapira A, Cooper J. The International Co-operative Ataxia Rating Scale (ICARS): An appropriate rating scale for Friedreich's ataxia? *Mov Disord*. 2005;20:1585–1591.
11. Cano S, Posner H, Moline M, et al. The ADAS-cog in Alzheimer's disease clinical trials: Psychometric evaluation of the sum and its parts. *J Neurol Neurosurg Psychiatry*. 2010;81:1363–1368.
12. Hobart J, Lamping D, Fitzpatrick R, Riazi A, Thompson A. The Multiple Sclerosis Impact Scale (MSIS-29): A new patient-based outcome measure. *Brain*. 2001;124:962–973.
13. Cano S, Browne J, Lamping D, Roberts A, McGrouther D, Black N. The Patient Outcomes of Surgery-Head/Neck (POS-Head/Neck): A new patient-based outcome measure. *J Plast Reconstr Aesthet Surg*. 2006;59:65–73.
14. Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: The role of new psychometric methods. *Health Technol Assess*. 2009;13:1–200.
15. Streiner D, Norman G. *Health Measurement Scales*: *A Practical Guide to their Development and Use*. 4th ed. Oxford, UK: Oxford University Press; 2008.
16. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality of life instruments: Attributes and review criteria. *Qual Life Res*. 2002;11:193–205.
17. Mokkink L, Terwee C, Patrick D, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Qual Life Res*. 2010;19:539–549.
18. Cano S, Browne J, Lamping D. Patient-based measures of outcome in plastic surgery: Current approaches and future directions. *Br J Plast Surg*. 2004;57:1–11.
19. Cano S, Hobart J, Linacre J, et al. Patient-based outcomes of cervical dystonia: A review of rating scales. *Mov Disord*. 2004;19:1054–1059.
20. Pusic A, Liu J, Chen C, et al. A systematic review of patient-reported outcome measures in head and neck cancer surgery. *Otolaryngol Head Neck Surg*. 2007;136:525–535.
21. Kosowski T, McCarthy C, Reavey P, et al. A systematic review of patient-reported outcome measures after facial cosmetic surgery and/or nonsurgical facial rejuvenation. *Plast Reconstr Surg*. 2009;123:1819–1827.
22. Chen C, Cano S, Klassen A, et al. Measuring quality of life in oncologic breast surgery: A systematic review of patient-reported outcome measures. *Breast J*. 2010;16:587–597.
23. 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.
24. Feinstein A. *Clinimetrics*. New Haven, CT: Yale University Press; 1987.
25. Stewart A, Ware J, editors. *Measuring Functioning and Well-being*: *The Medical Outcomes Study Approach*. Durham, NC: Duke University Press; 1992.
26. Nunnally J. *Psychometric Theory*. 2nd ed. New York, NY: McGraw-Hill; 1978.
27. Thurstone L. A method for scaling psychological and educational tests. *J Educ Psychol*. 1925;16:433–451.
28. Guttman L. A basis for analysing test-retest reliability. *Psychometrika*. 1945;10:255–282.
29. Gulliksen H. *Theory of Mental Tests*. New York, NY: Wiley; 1950.
30. Torgerson W. *Theory and Methods of Scaling*. New York, NY: John Wiley and Sons; 1958.

31. Edwards A. *Techniques of Attitude Scale Construction*. New York, NY: Appleton-Century-Crofts; 1957.
32. Likert R. A technique for the measurement of attitudes. *Arch Psychol*. 1932;140:5–55.
33. Likert R, Roslow S, Murphy G. A simple and reliable method of scoring the Thurstone attitude scales. *J Soc Psychol*. 1934;5: 228–238.
34. Ware J, Snow K, Kosinski M, Gandek B. *SF-36 Health Survey Manual and Interpretation Guide*. Boston, MA: Nimrod Press; 1993.
35. Ware J, Kosinski M, Keller S. *SF-36 Physical and Mental Health Summary Scales*: *A User's Manual*. Boston, MA: The Health Institute, New England Medical Center; 1994.
36. Goldberg D. *Manual of the General Health Questionnaire*. Windsor, UK: NFER-Nelson; 1978.
37. Zigmond A, Snaith R. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand*. 1983;67:361–370.
38. Kaska S, Weinstein J. Historical perspective. Ernest Amory Codman,1869–1940. A pioneer of evidence-based medicine: The end result idea. *Spine*. 1998;23:629–633.
39. Neuhauser D. Ernest Amory Codman, M.D., and end results of medical care. *Int J Technol Assess Health Care*. 1990;6:307–325.
40. Visick A. A study of the failures after gastectomy. *Ann R Coll Surg Engl*. 1948;3:266–284.
41. Karnofsky D, Abelmann W, Craver L, Burchenal J. The use of nitrogen mustards in the treatment of carcinoma. *Cancer*. 1948;1:634–656.
42. Fraser S. Quality-of-life measurement in surgical practice. *Br J Surg*. 1993;80:163–169.
43. Katz S, Downs T, Cash H, Grotz R. Progress in development of the index of ADL. *Gerontologist*. 1976;10:20–30.
44. Herndon R. *Handbook of Neurologic Rating Scales*. New York, NY: Demos Medical Publishing; 2006.
45. World Health Organisation. *Constitution of the World Health Organisation*. Geneva, Switzerland: World Health Organisation; 1948.
46. Robinson R. The policy context. *Br Med J*. 1993;307:994–996.
47. Hunt SM. The problem of quality of life. *Qual Life Res*. 1997;6: 205–212.
48. Bergner M, Bobbitt R, Pollard W, Martin D, Gilson B. The Sickness Impact Profile: Validation of a health status measure. *Med Care*. 1976;14:57–67.
49. Hunt S, McEwen J, McKenna S. *Measuring Health Status*. London, UK: Croom Helm; 1985.
50. Ware J, Sherbourne D. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. *Med Care*. 1992;30:473–483.
51. Aaronson N, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*. 1993;85:365–376.
52. Dawson J, Fitzpatrick R, Murray D, Carr A. Comparison of measures to assess outcomes in total hip replacement surgery. *Qual Health Care*. 1996;5:81–88.
53. O'Boyle C, McGee H, Hickey A, Joyce C, Browne J, O'Malley K. *The Schedule for the Evaluation of Individual Quality of Life (SEIQoL)*: *Administration Manual*. Dublin, Ireland: Royal College of Surgeons in Ireland; 1993.
54. Revicki D, Cella D. Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Qual Life Res*. 1997;6:595–600.
55. Fries JF, Cella D, Rose M, Krishnan E, Bruce B. Progress in assessing physical function in arthritis: PROMIS short forms and computerized adaptive testing. *J Rheumatol*. 2009;36:2061–2066.
56. Garcia S, Cella D, Clauser SB, et al. Standardizing patient-reported outcomes assessment in cancer clinical trials: A patient-reported outcomes measurement information system initiative. *J Clin Oncol*. 2007;25:5106–5112.
57. Browne J, McGee H, O'Boyle C. Conceptual approaches to the assessment of quality of life. *Psychol Health*. 1997;12:737–751.
58. Bowling A. *Measuring Health*: *A Review of Quality of Life Measurement Scales*. 3rd ed. Milton Keynes, UK: Open University Press; 2005.
59. Bergner M. Health status measures: An overview and guide for selection. *Annu Rev Public Health*. 1987;8:191–210.
60. Patrick D, Deyo R. Generic and disease-specific measures in assessing health status and quality of life. *Med Care*. 1989; 27(3 Suppl):S217–S232.
61. Fletcher A, Gore S, Jones D, Fitzpatrick R, Spiegelhalter D, Cox D. Quality of life measures in health care. II: Design, analysis, and interpretation. *Br Med J*. 1992;305:1145–1148.
62. Goldberg D, Hillier V. A scaled version of the General Health Questionnaire. *Psychol Med*. 1979;9:139–145.
63. Ruta D, Garratt A, Leng M, Russell I, MacDonald L. A new approach to measurement of quality of life: The patient-generated index. *Med Care*. 1994;32:1109–1126.
64. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess*. 1998;2:1–74.
65. Choppin B. An item bank using sample free calibration. *Nature*. 1968;219:870–872.
66. Linacre J. Computer-adaptive testing: A methodology whose time has come. In: Chae S, Kang U, Jeon E, Linacre J, editors. *Development of Computerised Middle School Achievement Tests*. Seoul, Korea: Komesa Press; 2000.
67. Ware J, Bjorner J, Kosinski M. Practical implications of item response theory and computer adaptive testing. A brief summary of ongoing studies of widely used headache impact scales. *Med Care*. 2000;38: 73–82.
68. Ware J, Brook R, Davies-Avery A, et al. *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Volume I, Model of Health and Methodology*. Santa Monica, CA: The Rand Corporation; 1980.
69. McDowell I, Newell C. *Measuring Health*: *A Guide to Rating Scales and Questionnaires*. 1st ed. Oxford, UK: Oxford University Press; 1987.
70. Streiner D, Norman G. *Health Measurement Scales*: *A Practical Guide to Their Development and Use*. 1st ed. Oxford, UK: Oxford University Press; 1989.
71. Guilford J. *Psychometric Methods*. 2nd ed. New York, NY: McGraw-Hill; 1954.
72. Nunnally J. *Tests and Measurements*: *Assessment and Prediction*. New York, NY: McGraw-Hill; 1959.
73. Thurstone L. Fechner's law and the method of equal-apprearing intervals. *J Exp Psychol*. 1929;12:214–214.
74. Nunnally J. *Psychometric Theory*. 1st ed. New York, NY: McGraw-Hill; 1967.
75. Brook R, Ware J, Davies-Avery A, et al. Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Volume VIII, Overview. Santa Monica, CA: The Rand Corporation; 1979.
76. Stewart A, Greenfield S, Hays R, et al. Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. *J Am Med Assoc*. 1989;262:907–913.
77. Spearman C. The proof and measurement of association between two things. *Am J Psychol*. 1904;15:72–101.
78. Novick M. The axioms and principal results of classical test theory. *J Math Psychol*. 1966;3:1–18.
79. Stevens S. On the theory of scales of measurement. *Science*. 1946;103:677–680.
80. Michell J. Measurement scales and statistics: A clash of paradigms. *Psychol Bull*. 1986;100:398–407.
81. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Education Research; 1960.
82. Andrich D. Controversy and the Rasch model: A characteristic of incompatible paradigms? *Med Care*. 2004;42:I7–I16.
83. Wright B, Stone M. *Best Test Design*: *Rasch Measurement*. Chicago, IL: MESA College Press; 1979.

84. Andrich D. *Rasch Models for Measurement*. Beverley Hills, CA: Sage Publications; 1988.
85. Wright B, Linacre J. Observations are always ordinal: Measurements, however must be interval. *Arch Phys Med Rehabil*. 1989;70:857–860.
86. McHorney C, Haley S, Ware J. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *J Clin Epidemiol*. 1997;50:451–461.
87. Prieto L, Alonso J, Lamarca R. Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health Qual Life Outcomes*. 2003;1:27.
88. Embretson S, Hershberger S, editors. *The New Rules of Measurement*. Mahwah, NJ: Lawrence Erlbaum Associates; 1999.
89. McHorney C, Tarlov A. Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Qual Life Res*. 1995;4:293–307.
90. Whitaker J, McFarland H, Rudge P, Reingold S. Outcomes assessment in multiple sclerosis trials: A critical analysis. *Mult Scler*. 1995;1:37–47.
91. Platz T, Eickhof C, Nuyens G, Vuadens P. Clinical scales for the assessment of spasticity, associated phenomena, and function: A systematic review of the literature. *Disabil Rehabil*. 2005;27:7–18.
92. Wright B, Masters G. *Rating Scale Analysis*: *Rasch Measurement*. Chicago, IL: MESA College Press; 1982.
93. Wright B. Solving measurement problems with the Rasch model. *J Educ Meas*. 1977;14:97–116.
94. Lord F. *Applications of Item Response Theory to Practical Testing*. Mahwah, NJ: Lawrence Erlbaum Associates; 1908.
95. Hambleton R. *Fundamentals of Item Response Theory*. London, UK: Sage Publications; 1991.
96. Norquist J, Fitzpatrick R, Dawson J, Jenkinson C. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Med Care*. 2004;42:I25–I36.
97. Lord F, Novick M. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley; 1968.
98. Massof R. The measurement of vision disability. *Optom Vis Sci*. 2002;79:516–552.
99. Cook K, Monahan P, McHorney C. Delicate balance between theory and practice. *Med Care*. 2003;41:571–574.
100. Fisher W. The Rasch debate: Validity and revolution in education measurement. In: Wilson M, editor. *Objective Measurement*: *Theory into Practice*. Norwood, NJ: Ablex; 1992.
101. Goldstein H. Consequences of using the Rasch model for educational assessment. *Br Educ Res J*. 1979;5:211–220.
102. Wright B. Misunderstanding the Rasch model. *J Educ Meas*. 1977;14:219–225.
103. Divgi D. Does the Rasch model really work for multiple choice items? Not if you look closely. *J Educ Meas*. 1986;23:283–298.
104. Goldstein H, Wood R. Five decades of item response modelling. *Br J Math Stat Psychol*. 1989;42:139–167.
105. Stenner A, Smith M. Testing construct theories. *Percept Mot Skills*. 1982;55:415–426.
106. Nicholl L, Hobart J, Cramp A, Lowe-Strong A. Measuring quality of life in multiple sclerosis: Not as simple as it sounds. *Mult Scler*. 2005;11:708–712.
107. Andrich D. A framework relating outcomes based education and the taxonomy of educational objectives. *Stud Educ Eval*. 2002;28:35–59.
108. Andrich D. Implication and applications of modern test theory in the context of outcomes based education. *Stud Educ Eval*. 2002;28:103–121.
109. Hobart J, Riazi A, Thompson A, et al. Getting the measure of spasticity in multiple sclerosis: The Multiple Sclerosis Spasticity Scale (MSSS-88). *Brain*. 2006;129:224–234.
110. Streiner D, Norman G. *Health Measurement Scales*: *A Practical Guide to Their Development and Use*. 2nd ed. Oxford, UK: Oxford University Press; 1995.
111. Nunnally J. *Introduction to Psychological Measurement*. New York, NY: McGraw-Hill; 1970.
112. Maurischat C, Ehlebracht-Konig I, Kuhn A, Bullinger M. Factorial validity and norm data comparison of the Short Form 12 in patients with inflammatory-rheumatic disease. *Rheumatol Int*. 2006;26:614–621.
113. Bohrnstedt G. Measurement. In: Rossi P, Wright J, Anderson A, editors. *Handbook of Survey Research*. New York, NY: Academic Press; 1983.
114. Cronbach L, Meehl P. Construct validity in psychological tests. *Psychol Bull*. 1955;52:281–302.
115. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*. 1959;56:81–105.
116. Kerlinger FN. *Foundations of Behavioural Research*. 2nd ed. New York, NY: Holt, Rinehart and Winston; 1973.
117. Stenner A, Smith M, Burdick D. Towards a theory of construct definition. *J Educ Meas*. 1983;20:305–316.
118. Revicki D. FDA draft guidance and health-outcomes research. *Lancet*. 2007;369:540–542.
119. Stenner A, Burdick H, Sandford E, Burdick D. How accurate are Lexile text measures? *J Appl Meas*. 2006;7:307–322.