FROM THE 2006 NIDRR SCI MEASURES MEETING
Towards Guidelines for Evaluation of Measures: An Introduction With Application to Spinal Cord Injury

Mark V. Johnston, PhD[1]; Daniel E. Graves, PhD[2]

[1]College of Health Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin; [2]Baylor College of Medicine, Spinal Cord Injury Research, The Institute for Rehabilitation and Research, Houston, Texas

Abstract

Background: Both clinical practice and research in spinal cord injury (SCI) continue to struggle with issues of the quality and utility of outcome measures employed. Despite widespread deference to dicta on "reliability and validity," systematic means of grading the level of evidence for measures are lacking.

Objectives: This paper explains the methods and principles for use in systematic reviews of measures in SCI. It explains how extant measurement standards and principles can be elaborated for extant labels on various types of reliability and validity to define a more judicious method of grading level of evidence. We aim to initiate a process of discussion that will lead to improved systematic review of the measurement quality as a basis for long-term improvements in outcomes measures and their application.

Methods: This paper is a conceptual review, based on established measurement standards and principles and the incorporation of recent advances in measurement methodology. The scheme of grading of measurement quality is illustrated by examples of measures of health, function, activity/participation, and quality of life after SCI.

Results and Conclusions: It is possible to grade the quality of outcome measure in terms of level of evidence, provided the nature of the construct being measured is defined as well as its main use. Definite means of grading the level of evidence for measurement will help to identify priorities for measure development and facilitate more appropriate uses of measures.

J Spinal Cord Med. 2008;31:13–26

Key Words: Spinal cord injuries; Outcome assessment; Measurement; Disability; Quality of life; Health status indicators; Systematic review; Level of evidence; Reliability; Validity

INTRODUCTION

Valid, reliable, and sensitive outcome measures are required in all fields of health care and for all types of outcomes. The validity of existing scales is particularly questioned for interventions that aim to improve function or quality of life (QoL), but quality measurement is also needed in studies of emerging curative interventions. Both extant and newly developing interventions in rehabilitation commonly aim to improve particular aspects of function, and existing measures may not incorporate aspects of function or QoL that the intervention is likely to affect. The scarcity of fully validated outcome measures can be particularly problematic in many low-frequency conditions, including the different levels and types of spinal cord injury (SCI). Expressed another way, disability and rehabilitation measures are concerned with the needs and problems of people with disabilities, problems that differ from those of the general population. Complaints of the irrelevance of the common outcome measures employed in rehabilitation are commonly heard and have not diminished over the years. The adequacy of existing instruments for measuring outcomes needs to be systematically examined so that we can know what new or modified instruments need to be developed or tested (1). This paper addresses the requirements for such a systematic examination.

In recent years, hundreds and even thousands of outcome measures have been developed, including scores of scales assessing activities of daily living, as well as additional scores for instrumental activities of daily living, handicap/community activities, physical and mental health, and many diagnostic groups (2).

Please address correspondence to Mark V. Johnston, PhD, University of Wisconsin-Milwaukee, Occupational Therapy, 2400 E. Hartford Avenue, Milwaukee, WI 53211; phone: 414.229.3616; fax: 414.229.5100 (e-mail: johnst@uwm.edu).

The views expressed in this article are those of the authors and do not necessarily represent the views of the US Department of Education or the United States. Given in brief form as a presentation at the ASIA Conference, June 24, 2006, Boston, MA.

Researchers, clinicians, administrators, and clinical outcomes managers face the challenge of choosing the best measure for the patient problems they treat, a choice that is not always easy technically. Payers may require progress reports, and these need to provide meaningful, objective (reliable and valid) information on major patient problems and progress. Any assessments they require also need validation, and appeals based on recognized criteria for reliability and validity of measures can provide a strong argument. The difficulty of such tasks is multiplied when there is a lack of accurate and standardized information concerning the technical quality of measures, as well as confusion regarding standards and methods for grading the quality of evidence for measures. Researchers and practitioners are left to their own devices to locate and evaluate a bewildering assortment of data. Usually the most popular or traditional measure is chosen, an understandable compromise but not necessarily one that ensures that the best possible measure is chosen.

All of these problems have occurred despite widespread awareness of traditional notions of "reliability and validity" of measures. Measures of all types are described as "reliable and valid," without specification of how reliable or how valid, or in what ways; the terms are frequently used almost as a mantra, synonymous with "good," rather than reflecting an evaluation of the quality of a measure for a particular construct or application. The thesis of this article is that traditional notions of "reliability and validity" need to be reformulated—refined, elaborated, simplified, and stated in terms of grades or levels—to enable us to evaluate the level and type of evidence for measures and thus to provide guidance for how measures need to be improved and used.

OBJECTIVES

The purposes of this paper are:

- To provide a critique of existing measurement standards and to suggest profitable directions for future improvement. More specifically, we will explain how principles of measurement can be elaborated from the level of indefinite principles or dichotomous judgment to a more meaningful and judicious grading of level of evidence.
- To introduce a method of systematically summarizing the strengths and limitations of extant outcome measures. The method of grading measures will be illustrated by studies of measures of motor impairment, functional activities, participation, and QoL among people with SCI.

The paper notes issues and plans for the future. It is designed to spark discussion that will lead to better methods of grading the strengths and limitations of outcomes measures, leading in turn to better measures and to more useful research results in the future.

BACKGROUND AND BASIS

Motivation: Improving Measurement Standards in Rehabilitation

A number of professional associations have published standards for measurement of health-related outcomes and performance over the years, including:

> The American Psychological Association, in conjunction with the American Educational Research Association (3),
> The American Physical Therapy Association (4), and
> The American Congress of Rehabilitation Medicine, which published *Measurement Standards for Interdisciplinary Medical Rehabilitation* in 1992 (5).

These standards have encapsulated consensus regarding the most desirable scientific characteristics for measures of experienced or observed health, function, activity, and QoL. Based largely on "psychometric" principles and methods, the underlying statistical methods and criteria transcend psychology and have long been applied to a wide variety of outcome constructs whose measurement is largely probabilistic and that cannot be measured in terms of physical or other natural science quantities (3,6–8). Because many years have elapsed since their publication, the question of whether these measurement standards are still valid and up to date arises, and there has now been sufficient time to accumulate experience regarding needed improvements to facilitate the application of measurement principles.

Measurement Principles

Measurement involves a systematic procedure for assigning a number to an observation: permissible and firmly based uses and inferences need to be distinguished from false or misleading ones. The basic principles enunciated in *Measurement Standards for Interdisciplinary Rehabilitation* (5) still apply, but developments have occurred.

- The need to understand measurement reliability (freedom from random error) has remained, although more sophisticated methods of computing reliability than those found in classical test theory (CTT) are increasingly employed. Improved methods of computing reliability include more meaningful ways of conveying reliability information (eg, the confidence interval for error of measurement rather than difficult-to-interpret statistics, such as the intraclass correlation coefficient); approaches that attempt to identify multiple sources of measurement error rather than reducing it to a single parameter; approaches that attempt to identify differences in reliability across individuals measured; approaches involving more elaborated measurement models than CTT. An improved version of measurement standards should make reference to these improvements.
- The need to understand and report measurement bias has grown over time. Concerns about measurement bias (viz, lack of double blinding) was at least as serious a problem as scarcity of randomized clinical trials in

systematic reviews of evidence in rehabilitation (9).

- Content and face validity remain a critical first step in establishing the validity of a test or scale in both traditional and contemporary approaches. A scale that does not contain items on the main, logically expected effects of an intervention, is not likely to be a valid or sensitive measure of its effects. The passage of time has taught measurement researchers the need to emphasize content validity and the need to explicate at least a simple theory of the needed content for a measure, that is, a framework from which construct validity can begin to be evaluated.
- Predictive or criterion-oriented validity remains essential because a useful scale predicts something outside of itself. Many scales may be used to predict a particular event, in which case knowledge of predictive validity for that criterion becomes critical.
- Construct validity remains the overarching criterion and still requires study of a network of relationships and incorporates both convergent validity (whether the measure is related to the things it should be related to, according to best theory and knowledge) and discriminant validity (the degree to which the measure is distinguished from confounding factors).
- The improved measurement models in Rasch and item-response theory (IRT) analysis can provide firmer evidence of internal measurement validity than older true score analyses of internal consistency (6,8,10–14). It is now possible to transform raw (observed) scores, whether they are observations of highly complex, varying biological systems, or reports of QoL, into objective linear measures with properties similar to physical measures (15), with the exception that they are more probabilistic and have greater random measurement error than typical measures of physical constructs, such as height, weight, and acceleration.

Validity remains the overarching criterion for evaluation of measurement procedures, but understanding of validity has grown over the years. It has become clear that the position enunciated in old measurement publications—that validity is dichotomous—is simplistic and needs to be revised. Validity is not a simple dichotomy (although dichotomous decisions may result from application of a measure, and a particular threshold may need to be set in a particular application or situation). Rather, validity is a complex concept that includes multiple characteristics of the construct to be measured, relationships to other constructs, limitations, and primary uses of the measure. Moreover, the validity of a measure is not static. Evidence of validity evolves over time: as one uses a measure in an increasing number of studies, one obtains more and more empirical information about allowable inferences from the measure (and limitations of it). One's understanding or theory of the construct being measured should change as new discoveries are made. At some point, development of a measure moves from basic validation to enhancement of understanding of the construct, which in turn can alter aspects of the measure and its use. Validity also is not entirely a property of the measure itself but involves the intended purposes, uses, or inferences as well. In sum, the validity of a measure has multiple attributes; it can and should be graded; and it will evolve over time.

Developments: New Scales

The fields of psychometrics and health outcomes measurement have continued to grow over the last 14 years. These years have seen the publication of large number of new measures of health and function, with research articles describing uses as well as refinements to older scales (2,16–18). This growth shows the need for measures that can be tailored to address particular measurement problems but also raises questions about proliferation of redundant measures and whether the growth in measures will continue without end or apparent order. Our own files have well more than 50 published scales of activities of daily living ADL alone. Does the world need new scales of ADL? Which scale should one choose? Are there still gaps in measurement? Such questions occur in many areas of outcomes measurement, and they deserve an answer. The quality of outcome measures remains important, as does the need to understand limitations, strengths, interrelationships, and the most appropriate applications of various measures.

Technical Developments

At the same time, uncertainties about technical psychometric issues continue to confound attempts at agreement on exact statistical criteria for measurement quality. The growth of new-and-improved but technically different psychometric methods (including Rasch analysis and IRT models) has complicated matters. The methods are based on different assumptions regarding probabilistic measurement modeling and give rise to different questions and technical criteria on measurement quality (11,19). These models were developed to overcome limitations and problems in classical test theory (CTT). In fact, several IRT models have been developed to address various measurement issues, including assumptions concerning the possible responses and the construct under consideration. There is an emerging consensus among psychometricians that these new models are superior to CTT in most circumstances (11,20), because CTT is a special and sometimes rather limited case. Both IRT and Rasch analysis alleviate problems with undetected ceiling and floor measurement, gaps in items, redundancy, and many other potential problems encountered with older measurement models (6,8,11,13,21).

Use of CTT, however, persists, and in most cases CTT methods do provide useful information. Additional technical comparisons and much communication will be needed to overcome the differences between different schools of metric analysis.

Both Rasch analysis and true IRT entail a reformulation of classical notions of reliability and validity. Both are methods of analysis of internal validity or structure (how items in a scale relate to each other and to the construct they are purported to measure), so that indicators of desirable internal structure (eg, item separation and separation reliability, information indicators) combine aspects of both reliability and validity (that is, a ratio of desirable to undesired or random variability is computed). Essentially, the newer methods provide stronger and more stringent statistical indicators of metric interrelationships among items in a possible scale. External validity (how a scale relates to important variables outside of itself) becomes quite separate and means essentially the same as the traditional term, predictive validity.

Major Limitation: Practical Implementation

Perhaps the most important limitation to *Measurement Standards* (5) has been the absence of an appropriate method of implementing them. The *Standards* were written at the level of scientific principle, not at a level of specificity that permits and encourages implementation. Extremely specific standards could also be criticized and resisted as being inappropriate and counterproductive in many circumstances: hence our caution in enunciating points on which there is widest agreement and our endeavor to be educational rather than prescriptive. Nonetheless, without greater specification, standards are of little specific use.

Developing and validating a new scale is a great deal of effort and requires a series of studies. Measurement standards should recognize the limitations of new measures but not derogate them because they have not yet achieved the sustained funding necessary to fully validate them: improved measurement standards should help us to judge whether there is a gap in extant measures that needs to be filled.

Moreover, measurement standards should help guide and encourage the development of better measures and assist with their interpretation: researchers should be able to use them to plan and justify their application of measures and measure development research, and peer review boards and editors should be able to use standards to evaluate measures in proposals and manuscripts. Finally, users of measures need guidance on what measure they should choose for what problem or domain. Improved measurement standards should help users to choose measures in a practical way using sound metric information presented in a consistent framework.

Models for Improved Standards

Since publication of *Measurement Standards*, many important standards relevant to health science have been published. These include:

- The Consolidated Standards of Reporting Trials (CONSORT) statement (22,23), which has been adopted by an extraordinary array of medical journals. More detailed criteria have been developed for specific types of trials, but the strategy of starting with criteria for randomized controlled trials in general is sensible. The CONSORT statement has led to a demonstrable improvement in reporting of clinical trials (24).
- Standards for Reporting of Diagnostic Accuracy (STARD) have similarly been published and are being adopted more and more widely (25).
- Standards and methods for grading the strength of evidence for therapeutic interventions as well as for diagnostic, screening, and prognosis studies have been published and widely applied by the Cochrane Collaboration (26), the AAN (27), and an increasing number of other organizations (28).

These standards and methods have been enormously and increasingly influential. They have had a demonstrable effect on scientific publications and clinical policy and decision-making. This effect is due to the extensive consensus-development work, resulting not only in a statement of principles but also detailed operational manuals with checklists or grading methods that permit editors to evaluate the adequacy of manuscripts and reviewers to grade the quality of evidence to be synthesized. To develop improved measurement standards, we can learn from these examples.

Specific checklists are used in CONSORT and in STARD, because the aim is to specify what scientific editors and reviewers should look for (and authors should include) in publications reporting clinical trials and diagnostic methods, respectively. The strength of evidence on a particular topic is based on a review of multiple studies, and the accepted standard is to grade strength in terms of multiple levels (26,27), typically: Level 1 (very strong evidence, a finding that can be considered to be established); Level II (good/probable evidence); and Level III (some evidence, worth considering). Recommendations for further research are reported as well, especially when evidence is insufficient to justify a clinical practice recommendation. It should be possible to employ such graded methods also to the evaluation of the quality of measures for a well-defined construct or application. Such manuals, checklists, and grading methods are needed to advance measurement to the next level.

Grading the Quality of Measures

Measurement Standards need to move from the level of textbook statements of principles to more specific methods of grading a scale's quality and evaluating the appropriateness of a particular application of that scale. If measurement standards are to be applied fairly and systematically, reviews of measures will need to employ an "evidence table," a grid, checklist, or rating scheme based on criteria that can in principle be applied uniformly across potentially relevant measures. At present there are no accepted or published standards or methods for such grading. The absence of an established method of grading evidence for outcome measures is a major problem in rehabilitation science, because the goals of rehabilitation interventions may not map well into extant measures. Objective criteria for evaluating the severity of this problem, however, are lacking. Prior published standards and texts on development of health and rehabilitation measures provide a basis in principle for grading the quality of measures. These principles can be elaborated and clarified to create a grading system. Substantial work is needed to devise clear grading criteria and instructions for grading the quality of measures, to reach consensus, and to obtain experience and data on their utility.

MEASURES OF HEALTH, FUNCTION, AND QoL

Working with a group of experienced researchers, we have devised a method of grading the quality of measures of health, function, and QoL and have begun to apply it to SCI outcome measures. In the section below, we introduce this method.

The method is primarily a grid that directs attention to primary indicators of measurement quality. The grid is a way of collecting key information and asking uniform questions across scales. We also have attempted a summary rating of the overall quality (validity) of each measure reviewed. In the text below, we will explain the principles involved and provide a few examples.

Operationally, the grids are provided in a blank document so that space can expand when needed. Extensive room is provided for comments, because measures are complex and important points may not fall into neat boxes. We recommend reviews of measures be formatted to provide both a systematic evidence table and a substantial textual introduction with a good deal of qualification and explanation.

The First Step: Explain the Construct and Use of the Measure

Evaluating the quality of a measure begins with (and after much consideration, returns to) the construct being measured. One cannot evaluate any measure or scale without specifying what it is that one is attempting to quantify, evaluate, or categorize. The authors of different measures interpret a construct or rubric differently, and one needs to know these differences to choose which measure to use in a clinical trial or other research application: a degree of elaboration (perhaps at least a paragraph) is needed to explain the construct. Simple rubrics (eg, "function," "quality of life") are not enough of a descriptor, as different authors interpret these terms in very different ways. The most important criterion for the quality of a measure can vary depending on the specific construct, so that it is difficult or impossible to uniformly and reliably grade the quality of measures of a vague construct or broad domain.

The construct not only needs to be defined as well as possible but also needs to be delimited, qualified, or distinguished from other constructs with which it might be confused. The aim is clarity, but not necessarily the perfect clarity of a univocal attribute, because relevant outcome constructs, such as health, function, and QoL, have degree of richness, and there are almost always uncertainties about how far a construct goes and when it needs to be distinguished from similar or confounding variables.

The use or main purpose of a scale will affect one's evaluation of its quality. At least the primary application of the scale should be specified, as well as the main criterion that one would expect a valid scale would predict, because one expects that a valid measure will predict something important outside of itself.

Construct validity is generally considered to be the ultimate form of validity, and it can be simply explained as whether a measurement procedure yields numbers that "behave" as they should according to theory. Construct validity is evaluated in terms of an interrelated set of ideas regarding a measure's content, internal structure, and what it should do, and what it should not do. One cannot judge construct validity if one does not know the construct and its closely associated theory or framework. In sum, at least a simple "theory" of the construct should be explicated to enable the validity of the measure to be judged.

Scale Content

Systematic description of a measure begins by describing the name of the scale and its content (Table 1). This can usually be done using language provided by the scale's author; examples of scale items may be provided to clarify scale content. These descriptions help a potential user to judge "face validity" and whether the scale assesses the domain of interest. When reviewing a group of scales that assess a similar domain, description of the usual content might be abbreviated, emphasizing nuances or variations of content provided by the particular scale.

A basic issue in evaluating a scale is whether it measures one construct or dimension, or two, or several. This is a general consideration for an evidence review on measurement, and it is considered in the second row of the grid, because other ratings will not make sense unless it is considered up front.

As an example of a construct definition, QoL is universally accepted as an important outcome domain but is defined in various ways. The term QoL is used to denote health-related QoL and global well-being and as a superordinate construct embracing virtually all of experienced outcomes (29). QoL and life satisfaction are separate constructs from functional independence or such mental states as anxiety (30). One must inspect the

Table 1. Scale Content

Scale Name	Name $^{\text{(with reference number)}}$
Description	Explain the construct being measured, emphasizing content in plain English. Provide the author's labels and describe the nature of items if, as is often the case, the label does not fully explain the content. The construct and domain being measured should be further explained in text.
Subscales/internal structure	Note each subscale (or dimension) using labels of author. Brief introductory description only.
Statistical support	Report statistics used by the author to support claims of unidimensionality/multidimensionality, usually a form of factor analysis. For single item scales, say "1 item/NA." For unidimensional scales, say "see reliability." If the support is conceptual rather than empirical, say so. (In a Rasch framework, look for factor structure of residuals; cross-reference separation reliability.)
Comment	Comment on content or face validity or method of scale development. For example, was the scale developed by experts or persons with a disability? Are there clear indicators of sensitivity/insensitivity to the concerns of people with spinal cord injury/disability? Similarly, comment on internal reliability and structure, such as whether you feel evidence of dimensionality is strong, medium, weak, or misleading, whether alternate solutions are possible but not stated above.

particular QoL measure used in a particular study to identify the QoL domain that is measured (29).

An example of internal structural analysis is provided by Graves and colleagues' IRT analysis of ASIA Motor Scale scores (31). The analysis demonstrated that measurement error is reduced by use of separate upper and lower extremity subscales, a result that neatly parallels Marino's finding that predictive validity for functional independence is improved by use of such subscales (32).

Administration

To choose a scale, users need to know basic characteristics of scale administration, such as:

- The basic type of scale or mode of administration.
- Whether it is adapted for use by key disability groups (eg, self-controlled computer administration for use by persons with tetraplegia, Braille, spoken versions for persons with vision impairment).
- Burden indicators, such as number of items, time to completion, and cost. Space is provided to summarize each of these scale characteristics (Table 2).

Reliability

The reliability of a measurement procedure, which we broadly define as freedom from random error, needs to be understood to interpret its results and to answer questions, such as whether 2 numeric results really (probably) differ and whether one should have high, moderate, or low confidence in inferences from the measure: unreliability constrains validity. For a measure to be accepted as reasonably "reliable and valid," information on scale reliability or reproducibility should be reported (Table 3).

A difficulty is that different measurement frameworks (CTT, IRT, and Rasch) provide different ways of computing reliability. Although the authors recommend Rasch or IRT models, the current condition of the literature is that many measures have not yet received this level of analysis. Meanwhile, CTT coefficients, such as Cronbach's coefficient alpha, are still widely employed and provide some information. To be inclusive, space is provided in the grid for summary statistics and evidence from both CTT and IRT/Rasch analytical methods.

Table 2. Administration

Type/mode	Note the basic type/nature of the instrument in terms of data collection, (eg, questionnaire on subjective or objective items, rating of performance, instrumented assessment, computerized).
Disability adaptation	Note issues regarding necessary adaptations for tetraplegia/other disabilities.
Burden indicators	Number of items, expense, time to administer, special equipment, training required or available. Also record indications of risk or discomfort.
Comment	Optional comment on administration, burden/practicality to person and others.

Table 3. Reliability/Reproducibility

Internal consistency classical framework	Relevant to all multi-item summative scales. Report primary available statistics. Reliability statistics, such as SEM and others, may also be reported if they are the only information available. Cronbach alpha or standardized item alpha are most commonly reported. Not applicable for single-item scales.
Reliability: item-response theory (IRT) or Rasch framework	In 2p or 3p IRT framework, report marginal reliability (or test information function). In Rasch framework, report item separation reliability in decimal terms (eg, 9); measure-to-residual variance ratio could also be reported.
Interrater, test-retest, and other reliability/reproducibility characteristics	Key statistics and information are reported here. In grading the adequacy of reliability information, the needed form and degree of reliability depends on the nature and application of the test. For multipoint scales, relevant statistics on reliability/reproducibility include, among other possible statistics, the following: • Bland-Altman Limits of Agreement are commonly used to describe reproducibility (33). • Intraclass correlation coefficients are commonly reported. • Measurement error is an interpretable way of conveying reliability information, because uncertainty can be conveyed in terms of a band of uncertainty for both research and clinical decisions. For ratio-level scales (eg, measures of physical quantities), Lin's Concordance Correlation Coefficient is an excellent statistic to summarize overall accuracy or reproducibility (34). For observations of performance or other judgments of complex observable phenomena, inter-rater reliability needs to be reported. Not directly applicable to subjective quality-of-life assessments. Because most measures are used to evaluate degree of change, test-retest reliability or stability should be tested. This test characteristic affects sensitivity to change. Report indicators of bias, such as higher scores are obtained in morning, higher ratings of less depressed patients. Information on sources of unreliability enhances understanding and use of measures.
Bias	Bias is when a measurement procedure gives results that are systematically too high or low with different raters, situations, or personal incentives. Although it is often not reported, bias in measurement is potentially very serious and needs to be reported whenever possible.
Comment	Comment on reliability or reproducibility, such as whether it is so low that it can be used only with repeated testing or in group studies.

Which form of reliability is most important also depends on the nature of the scale and its application. Inter-rater reliability should be reported for a judgmental rating of patient performance: internal consistency scores are not enough. Test-retest reliability or stability is important when assessing a construct (eg, behavior problems, pain, autonomic function) that varies over time. Instructions in Table 3 provide some technical information but are not meant to be comprehensive.

It is also valuable to report reliability in terms that a clinical professional can understand (eg, in terms of standard error of measurement or confidence intervals rather than in terms of ICCs).

Finally, most measures are subject to biases. Self-reports of subjective states, for instance, are subject to social desirability reporting biases. Bias is also a validity issue, but validity can remain if one accounts for magnitude and direction of likely bias. At the same time, results of a study or clinical assessment procedure may be discredited if biases go unaddressed.

Indicators of internal validity and reliability should be distinguished from indicators of external validity. Internal validity includes internal consistency, item separation reliability, factor structure, and all other indicators of internal structure in static analysis (stability as a measure of change being a separable issue). External validity includes predictive validity, discriminant validity, and other data on generalizability and relationships to variables outside of the scale's item set itself. Although this remains a logical structure, we have found that many raters, having been trained in CTT, have difficulty with this distinction, and so the grid here retains traditional labels.

Table 4a. Validity Indicators

Sensitivity to change	Report evidence regarding scale sensitivity or changes in scale properties over time, emphasizing change relevant to interventions in spinal cord injury (SCI). Cross-reference above information on test-retest reproducibility/reliability, which constrains sensitivity. In Rasch or item-response theory (IRT), report whether dynamic stability has been tested. Note other reports of insensitivity to change that actually occurs or of sensitivity to meaningless changes.
Ceiling/floor	Measurement ceiling and floor issues can determine whether change is detected or obscured by the measure. Report indicators of ceiling and floor effects, such as percentage at highest or lowest scale levels (eg, the great majority of SCI patients have maximum scores on Functional Independence Measure (FIM) cognitive items). If possible, also report highest and lowest level items, such as stairs is the most difficult item in the motor FIM.
Comment	Is the scale insensitive to clinically significant changes? Or does it detect changes that are meaningless to most people with SCI/the problem at issue?
IRT/Rasch validity coefficients	Report main statistics on model fit. In 2p and 3p IRT, report information function. In a Rasch framework, report item and person separation (and number of strata measured, if provided). Not applicable for single-item ratings and physical scales.
Other key indicators	IRT: best performing item, number of items at which information peaks. IRT or Rasch: other indicators the author considers important.
Comment	Summarize/comment.

Validity

Validity is a complex and multifaceted construct, and most of the rating grid is therefore devoted to different aspects of validity (Tables 4a and 4b). To provide a proper grading of a measure, it is important to prespecify the most important validity characteristics of the construct and the main use of the measure in question. If, for instance, a functional measure is used primarily to forecast future continuing support requirement, then it should be validated against measures of support needs, types, and hours.

Sensitivity to change remains a critical criterion for clinical measures (Table 4a). Acute rehabilitation populations are, as a rule, changing in function, and a measure that does not show change in the clinic is likely to be insensitive and invalid. Although greater sensitivity is usually good, outcome measures should not be chosen purely to be sensitive to the effect of the intervention or to show change: the main outcome should reflect activities, capabilities, or effects that are valued by people served ("clinically significant"). Over time, awareness has increased that sensitivity to change can and should be quantified (eg, in terms of standard error of measurement or other coefficients of reproducibility that provide a confidence interval). Patients whose improvement scores are less than the SEM may not really have improved at all. Newer statistical methods can also test whether the equal-interval structure of a measure is maintained upon retest (35).

As examples of sensitivity issues, it is well known that SCI patients' scores on the Functional Independence Measure (FIM) improve during rehabilitation. The cognitive items on the FIM, however, are not designed to be sensitive to possible improvements in attention or recall memory among patients who have both traumatic brain injury and SCI. The Spinal Cord Independence Measure is somewhat more sensitive to improvement than the FIM among patients with SCI, apparently because it includes additional domains and items (36).

Understanding of "ceiling" and "floor" issues is also essential to selection and interpretation of a scale. One would not use an oven thermometer to measure the temperature of a person, because the lowest grade on the oven thermometer may be 150 degrees: the measure has a floor problem. Similarly, the human thermometer has a "ceiling" problem in measuring oven temperature. A scale like the PF-10 (the 10 physical function items in the Short-Form 36 [SF-36]) cannot be expected to show improved function among people with complete tetraplegia, even if they in fact improve, because their physical functioning is below the floor of that scale. Conversely, a person may be discharged from inpatient rehabilitation at the top of the FIM scale but may need additional speed, strength, and endurance to successfully manage a household and return to work: the FIM has a ceiling problem for studies of community participation (37,38).

Developments in psychometrics promise a better and clearer way of evaluating the technical quality of an additive, conjoint measure. In Rasch analysis, the core indicator of the quality of a measure is item separation (also expressible as item separation reliability): a set of items with high item separation reliability can in all probability be employed as a scale. (Other questions can and should also be asked, eg, dimensionality, interpreta-

Table 4b. Criterion-Oriented Validity

Concurrent validity	Whether the scale predicts other measures of the same construct (or same labeled construct) measured at the same time. For diagnostic studies, you can report concordance with other useful diagnostic procedures here.
Criterion-oriented validity: predictive (and discriminant)	Report the predictive coefficient for the most important predictive use. If there is a clear categorical "gold standard" (eg, diagnostic or prognostic study), report accuracy, sensitivity, specificity, and positive predictive value. Discriminant validity could also be reported. Does the scale distinguish between 2 outcomes or 2 groups that need to be distinguished?
Clinical utility	Also called prescriptive validity and consequential validity. Do decisions in clinical practice alter depending on the measure? Note or rate extent of use in clinical practice per expert knowledge: not used, rarely, occasionally, frequently, very frequently/routinely.
Other validity	Report any other validity data here.
Comment	Space for optional comment on details of above validity information.
Population Applicability	
Applicability in SCI (vs other groups)	Describe use in SCI. (Describe other major groups and settings in which the scale has shown utility if use in SCI is limited.) Note any important indicators of problems/misfit in application to some subgroups of persons with SCI. (If Rasch analysis or other advanced metric analysis has been performed, person fit statistics should/could be reported. Note type of people misfitting, if reported, in comment section.)
Language(s)/multicultural issues	Note if available in alternate languages. If desired, also note if special procedures have been employed to ensure accuracy of translation (eg, back translation). If desired, note important research on multicultural issues.
Norms	Note whether reliable information on norms is available, such as population norms for SCI graded by level and completeness of impairment, age, or gender.
Extent of use in SCI	If an exact count is not feasible, use a gross indicator to quantify extent of use in SCI: none/virtually none, a few (eg, 2–4), many (eg, 5–10), extensive use (eg, 10 or more).
Comment	Comment on comparative extent and validity of use in SCI and other populations. If the scale was developed primarily in another group, comment on whether it is promising in SCI.

tion, stability over time, fit to a population.) The information function (a function that is inversely related to the error of measurement over the different levels of theta) provides a method of evaluating the efficiency and reliability of a measure for the population (39).

Criterion-Oriented Validity

Concurrent validity. The term "gold standard" is often used to describe the current best-accepted measurement methods. Reviewers should be aware that the "gold standard" may be brass at best: the criterion measure is likely to have measurement limitations and flaws as well. Consequently, we initially de-emphasized concurrent validity in our rating grid (Table 4b): validating a new experimental scale against an existing-but-also-weakly-validated scale provides only a questionable scientific advance, possibly adding redundancy and confusion rather than an improved tool.

Nonetheless, concurrent validity remains important. Those considering adoption of a new assessment method will want to know the relationship of the new scale to the best accepted of current methods, even if it is a clinical judgment of unknown reliability and validity. Scales benefit from being validated against the current best-accepted measurement method because this adds to understanding of the scale and to the interpretations that are logically possible (validity). For instance, a new scale might be essentially identical to the established scale, but it may be simpler and may also have greater range, avoiding ceiling and floor problems for some patients. The validity correlations and inferences of the prior scale then hold for the new scale, which is more practical, and the new scale would be preferred for persons who face high-level, or very basic, challenges.

As an example of knowledge provided by study of concurrent validity, Andresen has reported that similar constructs on the SF-36 and the Behavioral Risk Factors Surveillance System are correlated (40). The SF-36 vitality scale and the single Behavioral Risk Factors Surveillance System item asking about "days full of energy" correlate highly ($r = 0.8$). In contrast, Instrumental Activities of Daily Living and the Quality of Well Being correlate inversely ($r = -0.45$), because these measure different outcome constructs.

Predictive validity. A useful scale should predict something outside of itself. Indicators of desirable internal structure do not tell a user whether a measure will work well for any given purpose. Full validation of a scale entails gathering of data on predictive validity for its common uses. "Pre"-diction here should usually be of a future event, preferably a "gold standard" or other important criterion. For example, a measure of functional independence should predict minutes of assistance per day. Some scales predict multiple useful criteria, enhancing their utility and multipurpose validity, but predictive validity should at least be reported for the main use of a scale. Information on discriminant validity can be recorded in the box as well.

Predictive validity coefficients are a function of the degree of variability in the group sampled. Use of a "pure" sample is not necessarily a virtue in predictive studies: excluding people and circumstances that are commonly encountered in practice but that confound prediction artificially inflates predictive coefficients. The highest quality predictive studies will apply to the range of individuals commonly encountered in clinical practice: predictive or prognostic validity itself can be graded using Cochrane (26) and AAN (27) criteria.

Knowledge of predictive relationships is needed to enhance and confirm understanding of measures. For example, the Quality of Well-Being scale, Instrumental Activities of Daily Living, and physical health measures of the Behavioral Risk Factors Surveillance System and SF-36 all show greater impairment for quadriplegia than paraplegia; this provides a basic, although not surprising, confirmation of validity (40). Studies have found predictive relationships between neurologic level of injury and physical functioning measures; as one would expect, SCI affects health-related QoL, but it most strongly impairs QoL associated with physical functioning (41).

Clinical utility. Clinical utility, or prescriptive validity, is sufficiently important that we listed it separately on the grid. A minimum criterion for a claim that a measure is of clinical utility is that some practical clinical or policy decision (or decision by the person or family) changes as a consequence of the measure. Ideally, decisions are improved, and employing the measurement procedure leads to improved treatment and outcomes for the person. A surprising number of measures of function and QoL lack specific evidence of utility.

Population applicability. Issues of population applicability need to be addressed, even beyond application to the primary diagnostic group, such as SCI. Information on language versions and any information on multicultural validity or biases should be conveyed to assist with evaluation of a scale.

In addition, modern IRT and Rasch measurement models provide substantial information on misfitting individuals. The Rasch measurement model intrinsically and routinely reports "person separation," which is related to how well individuals are discriminated in the study sample and whether persons themselves form an ordered hierarchy. Rasch analysis also naturally provides indicators of person fit and misfit: some scales fit 99% of persons measured, whereas others fit small percentages, because some people have an unusual hierarchy of abilities. Data on whom the measure works and on who misfits should be reported.

Measures assess a defined content or domain, and any individuals may be concerned with a different set of items from that in any particular scale, giving rise to continuing controversy about the use and validity of standardized assessment itself, as opposed to qualitative research methods or even ad hoc and intuitive approaches. Measures that perfectly fit every individual may never be possible, but improved information on those to whom a measure logically applies (and does not apply) will ameliorate inappropriate applications of measures.

Finally, data on norms assist in interpretation and use of a measure, and the extent of use of the measure (number of studies employing it) is a potent if not precise indicator of scale quality.

Overall Summary Rating

In fields as diverse as restaurant rating and systematic reviews of the evidence on health interventions, quality ratings are summarized into a simple multipoint rating. Such ratings may be viewed as useful simplifications: potential users need to have an efficient way of locating relevant scales with at least some information on scale quality, and they need guidance in judging the amount of evidence in favor of the scale.

After summarizing data on scale reliability and validity, reviewers are asked to summarize the extent of favorable evidence for the scale, that is, to rate overall validity (Table 5). Reliability, validity, and other characteristics are to be considered for the construct and main use(s) specified initially. Reliability and validity are rather easily grouped together under validity, because reliability is a component of validity; unreliability generally constrains validity.

We have used a 5-point rating system here, but in practice we expect that 3 levels will be employed: one would not usually review a scale lacking published reliability/validity evidence (dot, no stars), and very few scales will have truly comprehensive validity information (4 stars). The SF-36 would be rated as having such comprehensive information as a generic scale of general population health, but there is much less evidence for its use in disability research, including SCI (1,42).

An innovation in the rating is the explicit consideration of evidence from outside of the diagnostic group (here, SCI). A scale of general function, participation, or QoL that is well validated in other groups should not be excluded from a review of activity, participation, or QoL

Table 5. Overall Summary Rating

Overall rating of quality in spinal cord injury (SCI)	○ No formal validity/reliability information published or content inappropriate. Do not review. ***Questionable or insufficient.** Little or no formal validity or reliability evidence, possibly questionable content for SCI. Development is required for application to SCI. ****Minimal validity.** Apparently applicable content with good validity/reliability in another group but little use in SCI. Or used in SCI but some limitations shown or with little reliability/validity information. Further development is desirable. Scale can be used if there are no alternatives, but firm conclusions are not possible given the preliminary and modest degree of validity evidence. *****Content and metric reliability and validity shown.** Adequately/reasonably valid for the main defined purpose. (Widely used outside of SCI, with formal studies/use in SCI.). OK to use in studies, although checking of assumptions or small improvements may be desirable to further improve the measure (eg, classical measures would benefit from item-response theory or Rasch analysis). ******Extensively validated and widely used.** Very well established as valid for the specific construct (eg, Short Form-36 for primary care).
Summary comment	*Comment on extent of use and validity in general.* Has it been used in clinical practice, research, or policy? Does the scale have construct validity, as indicated by a complex predicted pattern of theoretically expected relationships, free from confounding? What are biases/problems? These comments will ultimately go best in text.

outcome measures, because it may well be relevant to the particular diagnostic or disability group under consideration. At the same time, its validity for the disability group of concern would ordinarily be downgraded, because there are likely to be concerns regarding applicability of some of its items. Greater specification of these caveats will be valuable. A review of measures of disease symptoms or specific outcomes would not ordinarily consider studies of other populations, but the impact of disease is a different construct from general functional level, activity, participation, or QoL.

We realize that this overall rating is experimental and more evidence of its reliability and validity is required. At the same time, we feel that it is no longer satisfactory to leave the issue of measurement quality to the informal understanding that supports the current mix of high-quality measures and chaos. There is widespread agreement among scientists at least on basic principles of measurement, and if these principles can be operationalized in a grading system, they can be tested and improved and will ultimately have greater reality and utility.

Lessons from Employing the Grids and Grading System

Gathering all of the information about measures of a construct can be a considerable amount of work, more than is usually possible without adequate resources and funding. Reviewers found that it took considerable effort to employ the grids. Reviewers also need training in psychometric terminology, and expert support is needed to answer questions. Measurement studies are often not indexed in a simple or direct way. As with other complex review topics, multiple searches are needed, experts need to be contacted, and many studies need to be evaluated for relevance.

Reviewers need to understand that it is acceptable for many or even most cells to be empty. A comprehensive set of cells and criteria was created because each can be relevant, but one expects that very few measures will have been studied in all aspects. Nonetheless, the application of a uniform grid enables fairer comparison of the extent of study of different measures and illustrates the research that needs to be done to fully validate outcome measures.

The overall rating of quality at the end proved feasible, but there was little variation, with most scales of function and QoL in SCI being rated as having minimal validity (2 stars). This was expected: scales with almost no validity information were excluded from review, and few outcome measures have had such extensive study that they can be rated as fully or satisfactorily validated in SCI. The reliability of the rating scheme will require further study.

Our experience has not suggested that dichotomous yes-no judgments would be easier or more accurate than the multilevel grading. Perhaps a yes-no judgment would be possible if a construct and application of the measure were specified with precision, but the practice is always to describe a generalizable construct and at best a main use. Moreover, constructs are usually complex and even vague, injecting uncertainty into attempts to evaluate the quality of measures. Different uses will also require different levels of strength of evidence about a measure, and a rating of measurement quality will be more useful if it applies to the construct at issue and is robust at least across a group of interrelated practical uses. Moreover, evidence regarding the multiple characteristics of a measure evolves and increases over time, and only a multilevel or graded vocabulary can convey levels of evidentiary development.

Questions have arisen as to whether more finely graded ratings are possible. One suggestion is to devise a grading method for components of measurement quality (eg, for reliability, internal structural characteristics, predictive validity for the primary criterion). The possibility of multilevel grading of components of measurement quality deserves investigation. Systematic reviews in other fields of health care have found that strength of evidence can only be clearly graded with regard to a particular clinical assertion (5,26,27). Reliable, meaningful, multilevel ratings of the quality of measures are likely to require precise specification of the construct and/or uses of the measure.

Comment on Measures of Physical and Biological Constructs

Principles of reliability (freedom from random error) and validity in psychometrics were originally derived from scientific and statistical principles employed in the natural sciences for accuracy and reliability of measures. Principles of reliability and validity are applicable to measures of physical/biological constructs, as well as to nonphysical or psychosocial constructs, at least in an abstract, generic way:

- Reliability and error rates need to be evaluated and tested, rather than assumed, for both. Unreliability, that is, the degree of random error, needs to be quantified for measures of physical and biological constructs (43).
- Statistical methods for characterizing error rates, reliability, accuracy (if there is a definite quantitative criterion), and predictive validity are largely the same, or overlap.
- Statistically, distinctions between levels of measurement (categorical vs ordinal vs equal-interval) are more important than whether the construct being measured is physical or psychosocial.
- Evidence of the predictive validity and utility of measures of physical and biological constructs is also valuable. Important limitations, likely biases, and caveats should be specified.

At the same time, there are important differences. Physical and biological constructs may not be measured using additive, conjoint (averaging) methods. Greater precision is often (but not always) possible when measuring physical quantities compared with psychosocial ones. A rating form for physical and biological measures is available upon request.

SUMMARY

The application of research findings to practice commonly requires years of effort and experience. Similarly, the application of measurement principles to research and practice takes effort and recurrently improved formulations. We know now that the terms “validity” and “reliability” need to be qualified and graded. Beware of unqualified statements that a measure is “reliable and valid”: ask more questions.

- What is the degree of reliability? No scale is perfectly “reliable” without qualification. Is the measure reliable enough to use for the applicable decision about individuals? Or is reliability only sufficient for inferences about a group average? What factors bias measurement results?
- What is the exact construct that the scale attempts to measure? A single word is not enough.
- For what inference or application is the scale valid? What is the nature and strength of evidence for this use? Has the inference or use been definitely demonstrated? Or is the inference merely probable? Based on authority and popularity? Logical and desired but unproven?
- All scales and all constructs have limitations. What are sensible caveats or limitations?
- For whom is the scale most useful or valid? Scales often do not apply to all individuals, even all individuals in the group for which they were designed.

If we can initiate a process whereby the terms “reliable and valid” are routinely elaborated with more graded but still simple language—“minimally/moderately/established as valid to measure x for population y” —we will have achieved an improvement over current practice.

Improving Measurement Standards

We are now in a position to know how to improve extant measurement standards, making them more operational and useful in practice. The 1992 *Measurement Standards* need (a) technical revisions and elaborations at certain points, (b) a checklist to make it operational for journal editors and reviewers, and (c) a method of synthesizing and grading evidence from multiple measurement studies so that the quality of alternative measures of a construct can be summarized. Reliable grading of the quality of a measure is only possible with clear specification of the construct being quantified and a purpose or application. When this is done, the quality or validity of the measure can be judged, and strongly based inferences can be distinguished from probable or merely possible ones.

Improved measurement can move the whole field of rehabilitation and health outcomes research forward, including SCI research and practice. Further work will be needed to:

- Test and demonstrate the acceptability, reliability, and utility of the rating scheme.
- Refine the rating scheme and to clarify and elaborate instructions.
- Network with other organizations and experts in the US

and international community concerned with measurement quality.

Methods can be devised to grade the quality of measures. Development and testing of such methods should help us to know when an outcome domain has strong, reasonably adequate, or weak measures; to identify gaps in measurement; to identify the best measures; and to identify the ways in which extant measures require further validation or improvement. The authors of this paper invite critique and comment on these important issues.

ACKNOWLEDGMENTS

We thank Marcel Dijkers, PhD, David Tulsky, PhD, Allen Heinemann, PhD, and anonymous reviewers for their contributions to this manuscript. Special thanks are also given to NIDRR staff members, particularly Ruth Brannon and Arthur Sherwood, for leadership and support of processes designed to lead to long-term improvements in outcome measures for SCI and other disabilities.

REFERENCES

1. Jette AM, Keysor JJ. Uses of evidence in disability outcomes and effectiveness research. *Milbank Q.* 2002;80(2):325–345.
2. Bowling A. *Measuring Health: A Review of Quality of Life Measurement Scales.* Maidenhead, Berkshire, England: Open University Press; 2005.
3. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
4. Task Force on Standards for Measurement in Physical Therapy. Standards for tests and measurements in physical therapy practice. *Phys Ther.* 1991;71:589–622.
5. Johnston MV, Keith RA, Hinderer SR. Measurement standards for interdisciplinary medical rehabilitation. *Arch Phys Med Rehabil.* 1992;73:S3–S23.
6. Andrich D. *Rasch Models for Measurement.* Newbury Park, CA: Sage Publications; 1988.
7. Andrich D. Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. *J Appl Meas.* 2002;3:325–359.
8. Wright BD, Stone MH. *Making Measures.* Chicago, IL: Phaneron Press Inc; 2004.
9. Johnston MV, Sherer M, Whyte J. Applying evidence standards to rehabilitation research. *Am J Phys Med Rehabil.* 2006;85:292–309.
10. Tennant A, Penta M, Tesio L, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Med Care.* 2004;42(I Suppl):I37–I48.
11. Conrad KJ, Smith EV Jr. International conference on objective measurement: applications of Rasch analysis in health care. *Med Care.* 2004;42(I Suppl):I1–I6.
12. Smith EV Jr, Conrad KM, Chang K, Piazza J. An introduction to Rasch measurement for scale development and person assessment. *J Nurs Meas.* 2002;10:189–206.
13. Bond TG, Fox CM. *Applying the Rasch Model Fundamental Measurement in the Human Sciences.* Mahwah, NJ: L. Erlbaum; 2001.
14. Fisher WP. *Truth, Method, and Measurement: The Hermeneutic of Instrumentation and the Rasch Model* [PhD dissertation]. Chicago, IL: University of Chicago; 1988.
15. Tesio L. Measurement in clinical vs. biological medicine: the Rasch model as a bridge on a widening gap. *J Appl Meas.* 2004;5:362–366.
16. Bowling A. *Measuring Disease: A Review of Disease-Specific Quality of Life Measurement Scales.* Buckingham, UK: Open University Press; 2001.
17. Finch E. Canadian Physiotherapy Association. *Physical Rehabilitation Outcome Measures: A Guide to Enhanced Clinical Decision Making.* Hamilton, ON: BC Decker; 2002.
18. Dittmar SS, Gresham GE. *Functional Assessment and Outcome Measures for the Rehabilitation Health Professional.* Austin, TX: Pro-Ed; 2005.
19. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med Care.* 2004;42(I Suppl):I7–I16.
20. Andresen EM. Criteria for assessing the tools of disability outcomes research. *Arch Phys Med Rehabil.* 2000;81:S15–S20.
21. Embretson SE, Hershberger SL. *The New Rules of Measurement: What Every Psychologist and Educator Should Know.* Mahwah, NJ: L. Erlbaum Associates; 1999.
22. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA.* 2001; 285:1987–1991.
23. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA.* 1996;276:637–639.
24. Plint AC, Moher D, Morrison A, et al. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust.* 2006; 185:263–267.
25. Bossuyt PM, Reitsma JB. The STARD initiative. *Lancet.* 2003; 361:71.
26. Higgins JPT, Green S. *Cochrane Handbook for Systematic Reviews of Interventions 4.2.6.* Chichester, UK: John Wiley & Sons Ltd; 2006.
27. Edlund W, Gronseth G, So Y, Franklin G. *American Academy of Neurology Clinical Practice Guideline Process Manual.* American Academy of Neurology; 2004.
28. West S, King V, Carey TS, Lohr KN, Sutton SF, Lux L. *Systems to Rate the Strength of Scientific Evidence: Summary. AHRQ Evidence Report/Technology Assessment: Number 47.* Agency for Healthcare Research and Quality; 2002. AHRQ Publication No. 02-E015.
29. Post M, Noreau L. Quality of life after spinal cord injury. *J Neurol Phys Ther.* 2005;29:139–146.
30. Wood-Dauphinee S, Exner G, Bostanci B, et al. Quality of life in patients with spinal cord injury: basic issues, assessment, and recommendations. *Restor Neurol Neurosci.* 2002;20:135–149.
31. Graves DE, Frankiewicz RG, Donovan WH. Construct validity and dimensional structure of the ASIA motor scale. *J Spinal Cord Med.* 2006;29:39–45.
32. Marino RJ, Graves DE. Metric properties of the ASIA motor score: subscales improve correlation with functional activities. *Arch Phys Med Rehabil.* 2004;85:1804–1810.

33. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res.* 1999;8:135–160.
34. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45:255–268.
35. Wolfe EW, Chiu CW. Measuring pretest-posttest change with a Rasch rating scale model. *J Outcome Meas.* 1999;3: 134–161.
36. Catz A, Itzkovich M, Agranov E, Ring H, Tamir A. The spinal cord independence measure (SCIM): sensitivity to functional changes in subgroups of spinal cord lesion patients. *Spinal Cord.* 2001;39:97–100.
37. Turner-Stokes L. Standardized outcome assessment in brain injury rehabilitation for younger adults. *Disabil Rehabil.* 2002;24:383–389.
38. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the functional independence measure in traumatic spinal cord injury. *Arch Phys Med Rehabil.* 1999;80:1471–1476.
39. Lord FM. Information functions and optimal scoring weights. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1980:65–82.
40. Andresen EM, Fouts BS, Romeis JC, Brownson CA. Performance of health-related quality-of-life instruments in a spinal cord injured population. *Arch Phys Med Rehabil.* 1999;80:877–884.
41. Celik B, Gultekin O, Beydogan A, Caglar N. Domain-specific quality of life assessment in spinal cord injured patients. *Int J Rehabil Res.* 2007;30:97–101.
42. Ku JH. Health-related quality of life in patients with spinal cord injury: review of the short form 36 health questionnaire survey. *Yonsei Med J.* 2007;48:360–370.
43. Dieck RH. *Measurement Uncertainty: Methods and Applications*. Research Triangle Park, NC: Instrument Society of America; 1992.

ERRATUM

Re: Akhavan A, et al. Pilot evaluation of functional questionnaire for predicting ability of patients with tetraplegia to self-catheterize after continent diversion. *J Spinal Cord Med.* 2007;30:491–496.

Appendix 1 (Functional questionnaire) should have been titled, "The Capabilities of Upper Extremity Instrument" as originally published in Marino RJ, Shea JA, Stineman MG. *Arch Phys Med Rehabil.* 1998;79:1512–1521. The appendix should have clearly indicated the original source and that the authors modified the questionnaire by adding questions 18–22.