ELSEVIER

NeuroRehabilitation 5 (1995) 3–25

NeuroRehabilitation
An Interdisciplinary Journal

New developments in functional assessment: Probabilistic models for gold standards

William P. Fisher*, Richard F. Harvey, Karl M. Kilgore

Center for Rehabilitation Outcome Analysis, Marianjoy Rehabilitation Hospital and Clinics, Wheaton, IL, USA

Accepted 10 September 1994

Abstract

Advances in fundamental measurement have led to many exciting new developments in functional assessment. This paper presents fundamental measurement theory and method in summary form, and briefly describes its various applications to functional assessment, program evaluation, and outcomes analysis in physical medicine and rehabilitation. The implications of computerized medical records and longitudinal patient tracking in national or global computer networks for functional status and health status gold standards are briefly addressed.

Keywords: Functional assessment; Health status; Rasch; Conjoint measurement; Computerization; Measurement standards

1. Measurement theory

1.1. Fundamental and derived measurement

Measurement is fundamental, as opposed to derived, when its quantities express amounts of the variable of interest, and not amounts derived from quantities read off instruments that measure other variables. Distance (length) and time are commonly expressed as fundamental measures, such as kilometers and hours. Speed, however, is usually measured in a derived fashion as a ratio of distance to time, as in kilometers per hour.

In so far as all measurement is a cultural phenomenon, all measurement is derived in the sense that no variable is directly and immediately observed as a quantity; anything that is counted is counted as something. The units in which things are counted are invariably cultural constructs that emerge in particular linguistic and historical contexts. Such units are not things that exist in themselves in nature, self evident to anyone who can see and think [1–13]. Rather, good measurement, like all scientific modelling, is a matter of making good metaphors, and the metaphors chosen are limited to those that are meaningful and relevant to the culture and language of the time [1–9]. The fact that metaphor plays a role in science does

*Corresponding author, 2539 Octavia Street, New Orleans, LA 70115-6535, USA.

SSDI 1053-8135(94)00101-4

not mean that scientists can make nature accord with any theory at all. In their struggle to obtain even the faintest response to the questions they put to nature, scientists are constrained by their ability to calibrate measuring instruments [1–11] and to make measures that confirm, or at least do not falsify [12], their hypotheses.

The preparation of models, metaphors, and measures involves establishing common units of measurement and standards of comparison. Laboratories are able to link up with each other and to extend their results into the world at large only to the extent that the models, metaphors, and measurements associated with the production of a phenomenon enable that phenomenon's reliable reproduction [13]. Instruments and other devices constrain the administration of treatments and the observation of their effects such that the laboratory conditions necessary for creation of a phenomenon can be extended into the real world. Poorly or sloppily calibrated instruments recreate the conditions necessary for observation of the relevant phenomena unpredictably and unreliably, making it impossible or difficult to set up the networks of relationships necessary for managing the structure, process, or outcome of interest.

1.2. Quantification and objectivity

An instrumentarium of the highest quality is therefore fundamental to the effective and efficient communication of meaningful measures and the relationships they represent. High quality instruments, i.e. those that are rigorously quantitative, scientific, and objective, are recognizable as those for which a given amount of something (an inch or 2.54 cm of length, or 20 units of functional independence, for instance) is the same amount no matter what particular person or thing exhibits it and no matter what particular (configuration or brand name of) instrument is used to measure it.

For instance, when we measure a person's height, we pay no attention to the particular ruler in use; the person's height will remain constant, within a negligible amount of error, no matter which ruler is used to measure, no matter who wields the ruler, where the measure is made, or what brand name marks the ruler. Conversely, we pay no attention to what length, height, or width is measured by a particular ruler; the ruler's units stay the same size and in the same order no matter what or who is measured with it.

But when we measure with a ruler, we require attention to certain routine procedures of data quality. These are so basic to our thinking about measurement that we hardly recognize them for what they are: methods of maintaining the constant and invariant unit necessary for establishing a basis of comparison. For instance, would anyone ever measure height by holding a ruler on the floor for some people, at ankle level for others, and at knee level for others, and expect to have quantities that correspond to differences in people's heights? No, of course not. But our lack of quality control procedures in the use of rating scales create many situations directly analogous to this [14].

1.3. Deterministic vs. probabilistic conjoint measurement

Because the measurement constructs of the human sciences, including attitudes, feelings, health status, and abilities, such as functional independence, involve a degree of unpredictability and randomness, we cannot expect the kind of perfect accord that occurs between the units on measuring instruments and people's measures on constructs such as height [15]. This accord is referred to as conjoint order, with respect to the fact that more height on any ruler of any brand or metric always corresponds to more height as exhibited by the thing or person measured. For constructs of physical measurement, such as height, weight, blood pressure, temperature, etc., the conjoint order of the measuring instruments' units and the characteristics of the persons measured is deterministic, meaning that random variation within individual measures is virtually nonexistent.

For measuring abilities and attitudes, though, we need to calibrate instruments that can tolerate random noise; better yet, we need to devise methods that indicate how much noise is associated with any individual item calibration or person measure. Probabilistic conjoint measurement

models, such as those devised by Rasch [16] and his students Wright [17–20], Andrich [21] and Fischer [22], not only tolerate random noise, they require it for the estimation of scale values and use it as a basis for estimating individual errors of measurement and for indicating the quality of the data (the extent to which it fits the specifications of the measurement model).

Rigorous additive (interval level) conjoint measurement models have been researched in the psychometric, biometric, and econometric literature since at least the 1920s [23]. Deterministic conjoint models can be found in or derived from the works of Thurstone [24,25], Guttman [26], Bradley and Terry [27], Luce and Tukey [28], Krantz et al. [29], and Anderson [30]. These deterministic models have a history of [31–33] application in the area of functional status measurement, as well as in health status and quality of life measurement [34–39].

1.4. Gold standards

A great deal of interest and controversy in the application of deterministic models in health status and functional status measurement involves the issue of gold standards, instruments that are to be widely accepted as providing the reference unit of measurement in research and applications. The distinctions among three purposes of health status measurement identified by Kirshner and Guyatt [35] depend largely on the availability of gold standards calibrated within the overly-limiting constraints imposed on their development by the deterministic models used to assess their validity and reliability. The massive proliferation of health status, quality of life, and functional status instruments stems largely from the misperception that every new application or audience requires a new instrument, and that new instruments automatically entail a new quantitative unit of measurement. The result is that 'The profuseness of the measurement instruments reported in the literature seriously inhibits comparison of the results and ascertainment of significance of the results across different studies assessing similar problems' [39].

The difficulty in calibrating gold standards using deterministic models is that such models are bound to the particular items and rating scale found in a single instrument. The scale-dependency of most rating scale or test based units of measurement makes gold standards elusive. After all, different applications require different degrees of specificity, precision, and accuracy, and they may well also require different ranges in the measurement continuum [35–40]. Utilization management, third party payors, accreditation agencies, and case managers may want to know only if a patient can dress herself, but clinicians planning treatment and researchers studying treatment effect will need considerably more detail and less error in their measures. How could a single set of items attached to one rating scale ever meet all of these needs?

Probabilistic measurement models, such as those devised by Rasch [16] and Wright [17–20], are not constrained by the same factors that make deterministic models impractical [15]. Probabilistic models are sometimes referred to as scale-free, sample-free, or instrument-free [17] because their item parameter estimates do not vary according to the particular persons measured, and the person parameter estimates do not vary according to the particular set of items used as a basis of comparison. Scale-free measurement models bring gold standards within reach because they offer the possibility of calibrating different scales that measure the same thing by different means onto a single quantitative continuum, as is shown in some detail below in this paper.

Rasch's [16] and Wright's [17–20] simple and easy to use probabilistic formulations of conjoint measurement principles have led to common references to 'Rasch measurement,' but 'probabilistic conjoint measurement' is more descriptive and avoids the implication that Rasch has invented something entirely new, when in fact his work derives from an ancient line of mathematical thought that includes Galileo and Plato [41,42]. Rasch's contributions to measurement theory and practice will revolutionize functional assessment in rehabilitation if only because of his model's capacities for tolerating missing data and for testing the hypothesis that a set of rating criteria (assessment items) produce empirically consistent

data and so can be calibrated onto an equal-interval, and hence additive, measurement continuum.

1.5. Mathematical vs. concrete objectivity

Probabilistic conjoint measurement [16–20] requires that data be fit to a model specifying the quality of relations in the data necessary for measurement to be achieved. This approach helps ensure that all respondents, raters, and items are talking about the same thing, and are engaged with the same object of inquiry. The fitting of rating scale data to probabilistic models requires that a mathematical sense of objectivity frame thinking about what counts as data good enough to measure with.

Rasch's mathematical sense of objectivity contrasts with the more commonly held sense of objectivity as a matter of concrete materiality. Functional assessment scales, for instance, are usually designed in the context of the belief that an observation is objective when the rating assigned is identical across independent observers. To take a common example, if I raise my hand over my head, virtually anyone looking at me would independently agree that I raised my hand over my head. Extending this notion to the design of rating scales, rating scales have been called equal interval (which is a matter of having a constant, and therefore additive, unit of measurement) simply because the distances between several different levels of hand-raising represented in the scale were all the same. According to this approach, if I want to put my ability to raise my hand on a five-point rating scale, I rate a 1 when I cannot raise my hand from a hanging position at all, a 2 if I can raise it 45°, a 3 for 90°, a 4 for 135°, and a 5 for 180° from hanging. Each difference of one point corresponds to 45° in my range of motion, so this is an equal interval scale, right? Wrong!

Degrees of arc are not the same thing as the ability to move one's arm up and down. Mathematically, the probabilities of achieving any of the ratings on this hand-raising rating scale will vary depending upon the abilities of the persons rated, the particular difficulties that person experiences in hand-raising, and the amount and timing of treatment applied that is targeted at increasing people's abilities to raise their hands. Given these factors, it may be very easy, as indicated by the frequency of the ratings, to advance from level 1 hand-raising to level 2, but considerably more difficult to advance from level 2 to 3; then, because of the type of clients involved or the advances facilitated by treatment, hardly anyone may be rated a 4 because they pass straight on to being rated a 5. Using distance between categories to represent differences in difficulty, this scenario could result in a rating scale structure that looks like this:

1 2 34 5

even though the concrete observations on which the ratings are based are equal interval. Practical and useful probabilistic conjoint measurement models establish the odds of observations being made in each particular category and structure the instrument accordingly.

Rating scale measurement is not objective when it is simply tied to equidistant differences in what is observed. The importance of focusing on such differences stems from their utility in documenting consistent variation in the persons rated across the rating criteria, and vice versa. Mathematical objectivity is a matter of establishing (relatively and probabilistically) invariant conjoint orders that facilitate the measurement of amounts of the thing of interest, amounts that remain the same, within an error, no matter what instrument is used and no matter who is using it.

1.6. Agreement vs. consistency

Accordingly, as strict agreement in ratings is abandoned in favor of a consistent ordering of the items on the persons and vice versa, another measuring problem emerges: the propensity of the person rating to assign more or less extreme ratings [43–47]. Where the usual sense of objectivity in rating demands perfect agreement in ratings for the same behavior, a mathematical sense of objectivity has the more reasonable and attainable requirement of consistency. Research shows that people using rating scales tend to develop their own system for assigning ratings [47]. Assuming a calibrated instrument and trained

raters, when the raters are doing their best to use the instrument as they were trained to, assessments will vary consistently. Persons exhibiting less functional independence will receive the lowest ratings, those exhibiting more functional independence the highest, and the most difficult areas of independent functioning will receive the lowest ratings, and the easiest, the highest. Data that are consistent are not necessarily in perfect agreement across raters, but such data do make it possible to adjust patients' measures of functional independence by the severity or leniency of the raters' ratings. When the training of raters strives for agreement, raters often wind up second-guessing themselves, never achieving perfect agreement and often destroying the data's consistency as well.

1.7. Logarithms and logits

An aspect of the mathematical modelling of rating scales that seems to be almost universally mystifying to non-mathematicians involves the use of the logarithm in the estimation of the logit (logarithmic odds unit or log odds unit), which is what the quantities produced in the application of probabilistic conjoint measurement models are called. The odds part of the logit helps establish the order, and some of the spacing, of the rating scale categories. Because odds and the probabilities used in their calculation include reference to the number of items actually administered, they are able to maintain their ordinal properties when the number of items varies from person to person [19–21].

The logarithm, usually a natural logarithm, is used to linearize the odds. Linearization is necessary because odds necessarily range from zero to one and so suffer from floor and ceiling effects — absolute limits that misrepresent and distort the characteristics of variables at the extreme ends of the range of measurement currently investigated. What the logarithm does, in effect, is to give credit where it is due, expanding the size of the measuring unit as it is repeated away from the center of the distribution toward the extremes.

A simple example will illustrate the value of logarithms in estimating measures. Imagine that a new therapeutic treatment was found to reduce disability from a 50% to a 49.5% frequency of occurrence, and that another did the same thing, but the reduction was from a 1% to a 0.5% frequency. The former reduction is a negligible 1% change, but the latter amounts to a 50% overall reduction in the incidence of this type of disability. Should not the reduction that cuts a problem in half be weighted as more valuable than one that drops it by a mere 1% (the 0.5% reduction from 50%)? The logit provides that credit by incrementally decreasing the size of the raw score measuring unit as it is repeated from one end of the scale toward the middle of the scale, and then increasing it again as the other end of the scale is approached. Such logarithmic linearization has a long history of use in measurement, with common applications in barometric pressure and earthquake magnitude (Richter) scales.

1.8. Transparency

A basic goal of measurement system design is instrument transparency [48–50]. We want to be able to look through the instrument, as if it was not there, at the thing of interest. We can think clearly about a variable only in so far as the instrument acts as a window or lens on the variable. All instruments measure with some degree of error, so they are never perfectly transparent. However, to the extent that a variable's properties can be separated from the characteristics of the instrument, error is lowered and transparency is increased.

Transparent instruments make it possible to reproduce the most probable observations (ratings) that produced the measure, from the measure alone. Feinstein [48] points out that an instrument such as the Apgar scale, which is composed of five items on a three-point (0, 1, 2) rating scale, is transparent at its extremes (see Table 1). A score of 0 or 10 tells us what every rating on every item was, and a score of 1 or 9 gives a strong sense of what every rating on every item was. But according to Feinstein, scores of 4, 5, or 6 do not, and by implication cannot, provide a transparent sense of what a baby's condition is,

Table 1
Fictional Apgar scale data

Persons	Items					Person Scores
	3	2	5	1	4	
Nathan	1	2	1	2	1	7
John	0	0	0	1	0	1
Laura	2	2	1	2	2	9
Martha	1	1	0	2	0	4
Diane	1	2	0	2	1	6
Luc	0	0	0	0	0	0
Jon	2	2	1	2	1	8
Louise	0	1	0	2	0	3
Alissa	2	2	2	2	2	10
Jimi	1	2	0	2	0	5
Item score	10	14	5	17	7	

since a mid range score indicates that something is wrong, but gives no indication as to what in particular needs attention.

What if, however, Apgar scale data lend themselves to something approximating a deterministic conjoint order, such as that shown in Table 2? In this case, scores of 4, 5, or 6 are virtually as transparent as scores of 0, 1, 9, or 10. If this sort of consistency in the item order across the persons were obtainable with Apgar scale data, item 5 would represent the most likely area in which problems arise for babies, item 4 the next most likely area, etc.

However, for transparency to be more than a hollow technicality, the order to the items has to mean something; reliability is nothing without validity. Taking the data in Table 2 as an heuristic example, one would ask why babies usually receive the highest ratings on item 1 and the lowest on item 5? What is it about the observations provoked by these items that reasonably leads to their being positioned on the ends of the scale? And what about the ones in the middle; why are they ordered as they are? What is the one thing touched on by all of the items, more or less? Each item on the scale must be an indicator of how much health a baby exhibits. If the rating probabilities can be conjointly ordered, the log of the rating odds will produce an interval-level, additive scale.

Table 2
Fictional Apgar scale data tested for conjoint order

Persons	Items					Person scores
	1	2	3	4	5	
Luc	0	0	0	0	0	0
John	1	0	0	0	0	1
Louise	2	1	0	0	0	3
Martha	2	1	1	0	0	4
Jimi	2	2	1	0	0	5
Diane	2	2	1	1	0	6
Nathan	2	2	1	1	1	7
Jon	2	2	2	1	1	8
Laura	2	2	2	2	1	9
Alissa	2	2	2	2	2	10
Item score	17	14	10	7	5	

If the ratings do not order persons and items along a common continuum, pointing in the same direction, more work needs to be done to test and substantiate the hypothesis that the variable is quantitative [23]. Failure to confirm this hypothesis means only that the current approach to the variable, as realized by the recorded observations, does not lend itself to the variable's quantification. A different observational framework may show that the variable is quantitative. However, even if the quantitative hypothesis is repeatedly falsified for a given variable, that does not doom research on the variable to a non-scientific status. It merely means that the matter is qualitative and must be treated as such.

2. Advantages of probabilistic conjoint measurement for functional assessment

Probabilistic models have been used to study several kinds of instruments in general use in physical medicine and rehabilitation [49–77]. In addition to functional independence in motor skills (combined mobility and ADL functions) [51–62,74–77], these instruments include measures of impairment severity [51–54], applied self care [51–54], community reintegration [54], cognition and neuropsychiatric disorders [51–62,66,72,74–77], pain [63–65], pediatric developmental sequences [67–69], spiritual well-being [70,71], and patient satisfaction [73], with at least one instrument explicitly designed to produce data

fitting multifaceted probabilistic conjoint measurement models [74–77]. Each of these instruments' applications stand to benefit from the following advantages.

2.1. General points

Missing data are no problem. For instruments that add up raw ratings into total scores, such as the Apgar scale or the Functional Independence Measure (FIM), missing data is anathema because the meaning of the score depends on the number of items rated. Users of such scales are often instructed to assign a rating of 1 rather than not make any rating at all on an item. In contrast, the procedures for calculating logits deal only with the number of items for which ratings were convenient, possible, or meaningful [19,21]. The only consequence is that fewer items mean a higher error of measurement, something made explicit in the output produced by most computer programs that fit data to probabilistic conjoint measurement models. With fewer than a minimum number of items, probably 10–15, error becomes unmanageably large because reliability and the distinctions one is ostensibly trying to make among those measured are lowered.

Item or task calibrations and the measures of persons. Item or task calibrations and the measures of persons are logarithmically transformed, which is crucial for interpreting the value of differences involving comparisons made at various points along the measurement continuum.

Construct validity is placed at center stage. The grande dame of psychometrics, Jane Loevinger, claims that, 'Since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view' [78]. Samuel Messick, a vice president at Educational Testing Service and authority on the principles of construct validity, concurs, saying 'all measurement should be construct-referenced' [79]. Loevinger's [80] strong appreciation for Rasch's mathematics goes hand-in-hand with her position on construct validity, since, as Messick points out, 'any concept of validity of measurement must include reference to empirical consistency' [79].

As has already been shown, empirical consistency is the cornerstone of probabilistic conjoint measurement. Thus, the question, 'How do I know this instrument measures what it is supposed to?' becomes a matter of immediate importance. Computer programs such as BIGSTEPS [81] and FACETS [44] provide figures and tables with items and persons in calibration order as a matter of course, facilitating determination of how much of the variable each item measures and each person exhibits. These displays aid in the task of inquiring whether the persons, items, and rating scale categories expected to have the highest or lowest scale values in fact do. The visual and graphical grasp of variables as continua of more and less, like a thermometer or ruler, with the items and persons positioned on it, is crucial to understanding the measured quantities as an answer to the question 'How much?'

Analyses of the Functional Independence Measure [56–59] and the Levels of Rehabilitation Scale-III (LORS-III) [60,61], for instance, show that the common practice of aggregating ratings of functional independence from both the motor and cognitive domains is misguided. The two groups of items included on both of these instruments vary in the consistency of the data they produce. This means that scores in either instrument's middle range could be produced by persons with either high cognitive ratings and low motor ratings, or by the opposite, resulting in two very different sorts of patients having the same score. The two groups of items do not share a consistent order across all of the persons rated, making it impossible to determine an answer to the question 'How much?' because the scores are varying in two areas of functional independence at the same time.

Something similar was discovered in early studies of the Patient Evaluation Conference System (PECS©) [52,53], although its designers never intended its ratings to be summed across items into a total score [82]. Rather, cognitive, speech communication, and affect ratings were grouped together to test the tenability of a bread psychological construct. The data on psychological affect did not share the cognition-communication items' consistency, but eventually formed the basis of a community reintegration scale [54].

Computerized administration of instruments is greatly enhanced. The large numbers of items typically included in item banks preclude complete data for any one [83,84] person, so the probabilistic models' capacity to tolerate missing data makes them well positioned for the computerized administration of these banks [85]. Precalibrated item banks make it possible to target items at a person's ability; the computer draws items from the bank such that the next item is targeted at the person's current ability estimate. As someone continues to receive high ratings on mobility items, for instance, the computer administers progressively more difficult items until the ratings drop. The greater speed, reduced paper flow, and enhanced measurement precision of computerized instrument administration will be important as we face the demands of increasing pressure to cut costs, maintain outcomes, and document quality.

Measures are easily adjusted for rater severity / leniency. Clinicians who rate clients' performance on various functional tasks add an important dimension to the measurement of impairment, disability, and handicap. Multifaceted measurement designs [44–47,74–77] — those involving more than two facets — apply probabilistic conjoint measurement models in order to account for, and to remove from the measures, consistent variation in ratings caused by differences in rater severity or lenience, or by differences caused by any other uncontrolled but consistent influence on ratings.

The probabilistic models applied to functional assessment data thus far have most usually been two-faceted; the consistency of the data is hypothesized to result only from the interaction of the abilities of the persons measured and the difficulties of the items measuring. A three-faceted design includes the propensities of raters to assign consistently higher or lower ratings (their 'severity' or 'leniency') than the average as an influence on the consistency of the data. Calibrating the raters in this way makes it possible to adjust person ability measures and item difficulty calibrations according to the severity or leniency of the rater, so removing this uncontrollable factor from the resulting scale values. Multifaceted measurement analyses could also prove useful in accounting for consistent variations in functional independence that might be found to occur across facilities that encounter different degrees of impairment severity. If so, some scientific basis for the comparison of outcomes across facilities might be obtained.

Self-scoring report forms can provide quantitative information with no need for costly and time-consuming computer analyses. The instruments of the physical sciences typically embody something learned in a classic experiment. Thermometers, for example, look the way they do because of what was learned of the consistent relationships between heat and the expansion and contraction of fluids and gases in enclosed spaces. Only rarely, however, are calibrated rating scale-based instruments organized to represent the variable of interest quantitatively. When a variable has shown itself to be stable over an adequate period of time and across a wide variety of situations, the quantitative information can be used to lay out a worksheet or map that will inform the user of the relevant scale value, error and fit information as soon as it is filled out [86,87].

Each of the five PECS LifeScales™ [54] have been laid out as PECSMAPs™. Fig. 1 shows the Motor Skills scale in the self-scoring format. The areas in which ratings are assigned are listed on the left side of the page, and the rating scale categories are across from them on the right. The first thing that stands out to the new user is the fact that the rating scale categories are not printed in neat, evenly spaced columns and rows. Instead, they are arranged on the page to illustrate the respective amounts of independence in motor skills each of them represents, item by item. The items, listed on the left, include mobility skills rated by physical therapists and ADLs rated by occupational therapists. The item at the bottom of the list, feeding, is easiest to perform and so is likely to provoke the highest ratings, and the item at the top of the list, home management, is most difficult to perform and consistently results in the lowest ratings received by patients.

As the PECSMAP is filled out, ratings should fall in a column up and down the page, varying by an error or two to the left or right. The measure can be read off the instrument by simply running a line through the ratings and the horizontal

PECS MAP

PATIENT EVALUATION AND CONFERENCE SYSTEM

MOTOR SKILLS SELF-SCORING DIAGNOSTIC PROFILE

DATE _____ THERAPIST _____

PATIENT NAME _____ MEDICAL RECORD NO. _____

INDICATE STATUS/GOALS IS IG SS DS DG
CIRCLE RATINGS FOR GOALS; X RATINGS FOR STATUS

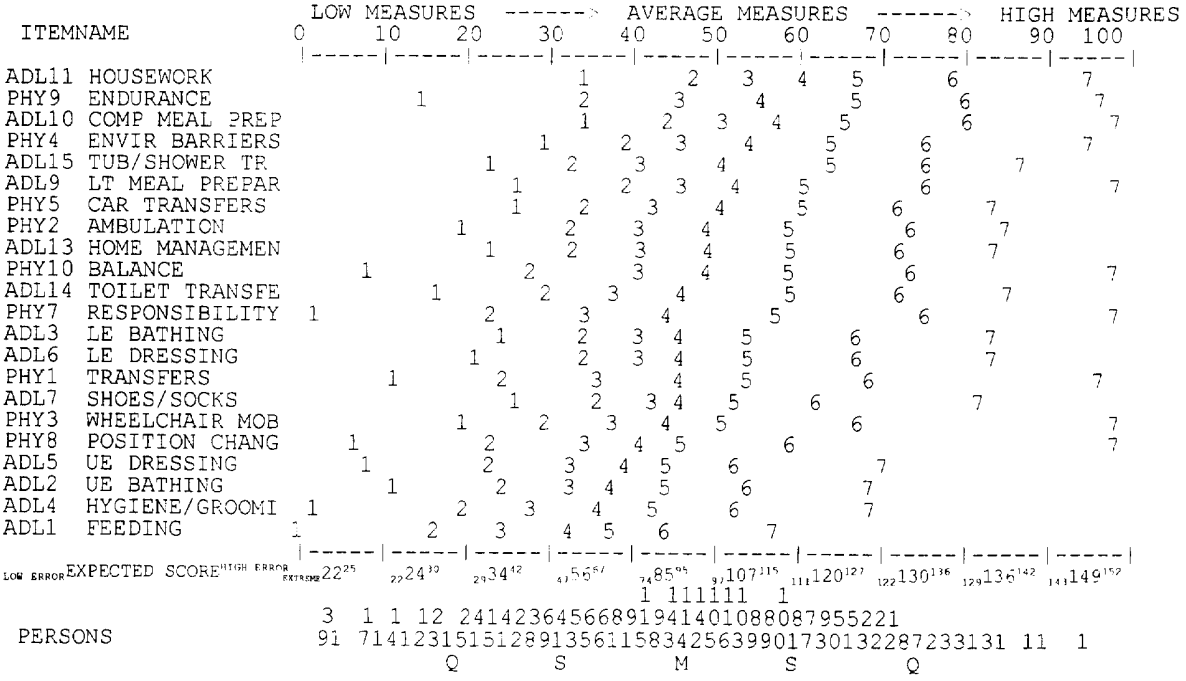


Fig. 1. PECS map.

0–100 logit scale. The error associated with different ranges in the measurement continuum is indicated, but must be inflated for incomplete data sets. The quality of the data is indicated by the extent to which it falls in a vertical line on the page. Ratings that stick out will probably most often indicate special strengths or weaknesses that differentiate a particular patient. Less often, they will indicate areas of patient malingering or extra effort, or of rater error. It is anticipated that worksheets of this kind will result in fewer errors being entered into the records uncorrected.

2.2. Specific points

Functional independence gold standards are within reach; different brands of instruments need no longer measure in their own idiosyncratic units. Rehabits (**Rehab**ilitation measurement **units**) are the result of the co-calibration of two or more functional assessment instruments and provide an empirically based (as opposed to legislatively imposed) currency of exchange for the communication of rehabilitation outcomes [88].

Rehabits are created when items from two or more separate instruments (PECS, FIM, LORS, Barthel, TAMP, AMPS, etc.) that are supposed to measure the same thing are all applied to the same patients, as if they came from one, and not several, instruments. Calibrating this super-scale results in scale values for each item, provided they all contribute mutually consistent data. Once an instrument's items have known positions on

the scale relative to other instruments' items, these values can be used to preset the items' scale values in analyses of patients for whom there are ratings only on that instrument. Because the item scale values producing the new measures are those obtained in relation to the items from the other instrument(s), the resulting measures are hypothesized to be statistically identical with the measures that would result from a similar application of the other instrument(s). These hypotheses would be falsified if (1) the co-calibration produced unacceptably high fit statistics, meaning that the ratings on the two instruments do not produce a consistent ordering of the persons; or (2) the co-calibration does not result in nearly identical measures for the persons as they are measured by the instruments in separate analyses. If the first hypothesis holds, it is likely that the second would as well.

To test the viability of these hypotheses as a basis for creating gold standards, ratings of 53 clients' performance on 13 FIM and 23 PECS motor skills items were made at admission and discharge, making for 106 measures. The FIM ratings were made by persons trained in its use without the knowledge of the therapists making their usual PECS ratings. The clients were consecutively admitted to a freestanding inpatient rehabilitation facility, and fell into a mixture of five diagnostic groups (brain injury (8), neuromuscular (7), musculo-skeletal (22), spinal cord injury (7), and stroke (10)). Prior research indicates that these scale structures remain constant across variations in function introduced by differences among the diagnostic groups, with few notable exceptions [89,90].

Partial credit measurement models [19,72,91,92] *allow the analysis to respect each instrument's distinctive sense of the 1–7 rating continuum.* It would not be reasonable to assume that the PECS's break between dependence and independence at ratings 4 and 5 could be imposed on or in some way shared with the FIM ratings, where this break occurs between categories 5 and 6; nor could the opposite be assumed. In fact, because the PECS ratings are derived from detailed descriptions of the behaviors associated with each category for each item, it is necessary to allow each item to have its own rating scale. Because the FIM has only one generic meaning to the rating categories, it is reasonable to test its data for adherence to this single structure.

The 35 combined PECS and FIM items calibrated along the measurement continuum with a reliability of 0.98; the 106 measures scaled with a reliability of 0.95. In a second analysis, the instruments were calibrated together on admission data only, then on discharge data only; the admission and discharge co-calibrations were then plotted against one another, resulting in a correlation of 0.89 and r^2 of 0.79. When one outlying PECS item and one outlying FIM item were removed, the correlation was 0.93 and the r^2 was 0.86.

These positive results encouraged a presetting (known as anchoring in psychometrics) of the item positions on the scale (from the original combined admission and discharge analysis) in instrument-specific analyses of the data. The measures resulting from these analyses were then plotted against the co-calibration measures (PECS/Co-calibration: $r = 0.95$, $r^2 = 0.91$; FIM/Co-calibration: $r = 0.95$, $r^2 = 0.90$) and each other ($r = 0.87$, $r^2 = 0.75$). The full report of the results [88] includes measurement method comparisons [93,94] that are more sophisticated than correlations. The full report also includes raw score to measure conversions for the PECS and FIM in which the Rehabits serve as a medium of exchange for the two scales. For instance, a 13-item FIM Motor Skills scale raw score of 64 is associated with a Rehabit of 50.0, and the 22-item PECS Motor Skills scale raw score associated with a Rehabit of 50.0 is 83.

Co-calibration establishes rigorous requirements for validity. A typical procedure performed when establishing the validity of a new instrument is to correlate measures produced by it with those produced by a similar, already established instrument. The question that then arises is that if a high correlation is desired and obtained, what is the distinct use and need for the new instrument? It often happens that correlations are inadequate tests of whether a common construct is measured by each instrument [93,94].

Co-calibration puts a new spin on this situation. If an existing instrument does not measure in the

necessary range of abilities, a new instrument might be called for. Similarly, if an existing instrument, such as the PECS or the FIM, was designed for management, and not clinical or research applications, then it probably has a measurement error too large for the instrument to provide the sensitivity required by the other applications, and a new instrument would be desirable.

Co-calibration would connect the new and the old instrumentation much more concretely than current practice does. Research would have to demonstrate that the new instrumentation measures the same thing as the old, and that it extends the range of measurement, or lowers the error.

Rehabits and other probabilistically modelled gold standards have potentially profound implications for the communication of outcomes, research, and standards for practice. What would it mean to have outcomes reported in a common unit of measurement? What are the implications for accreditation, reimbursement, program evaluation, marketing, treatment planning, and research? These and other questions must be explored.

Co-calibration suggests the possibility of linking grosser, higher-error management-oriented measures of functional independence with the more finely detailed, lower-error measurement needs of clinicians and researchers. The computerized progress note is a natural place to implement more finely tuned functional assessment instruments that are sensitive to perhaps even the smallest possible gains. Given that the computer is already in place in such a system, all of the benefits of computerized instrument administration follow. The primary benefit is that the item banking principles applied in computerized administration allow the measure to be adapted to its application.

A clinician who wants to plan the next stage of a series of ADL treatments will want a more precise measure than the case manager who wants only to know if his or her clients can dress themselves. The fewer items and greater administrative speed associated with managerial applications can tolerate the necessarily larger measurement error, but the clinician's purposes require a tighter focus, with an associated larger number of items and somewhat slower administration, in order to obtain a lower measurement error. Research is underway to co-calibrate low-level, detailed progress note, functional assessment items with standard PECS items in order to enable a simple conversion of the more detailed, lower error measurement information with the more global, higher error measurement information of the usual PECS.

Application of probabilistic models to functional assessment data streamlines program evaluation, quality assessment/improvement, and utilization management. Current uses of functional assessment data for purposes of program evaluation, quality assessment/improvement, utilization management, and communications with referring physicians, case managers, patients, and third-party payors [95–99] are encumbered by the difficulty of interpreting large numbers of data points in the context of length of stay, impairment severity, and age variables that constantly fluctuate. In addition, the data are usually ordinal ratings averaged across patients for individual areas of functional independence, such as ambulation or dressing; the validity of such statistical manipulation of raw rating scale data has been seriously questioned [14,20]. The evaluation of average discharge status, the percentage of treatment goals that were attained, and the percentage of patients discharged at an independent level of functioning faces additional complications from random variation in the amounts of treatment patients receive (usually inferred from length of stay data), the age of the patients, how impaired their independence was when they were admitted for treatment, and other variables, such as the amount of time that had passed since the onset of their impairments.

Fitting data to a probabilistic model makes it possible to aggregate ratings across items, reducing the number of data points that must be interpreted, and placing the measures on interval scales that can be manipulated statistically. The influence of admission levels of functional independence, length of stay, and age on discharge measures can then be gauged via regression techniques on historical data, and these standards used to evaluate current performance [96–98].

Fig. 2, for instance, shows predicted PECS Motor Skills LifeScale discharge measures (predictions generated by fitting a regression model) plotted against the actual measures. The 95% confidence intervals were produced by the 8 years of historical data shown in the figure. To evaluate a new set of outcomes, the predicted and actual data points from the new data would be plotted in relation to these confidence intervals, enabling one to see at a glance how current outcomes compare with those of the past.

Data points falling above the top diagonal line in Fig. 2 represent cases for which the actual outcome measure was higher than expected, given that person's initial status, length of stay, and age. Points falling below the bottom diagonal line indicate cases for which the actual outcome was lower than expected. Such severity adjusted measures and comparisons will be of increasing importance as rehabilitation consumers seek to find the most cost-effective treatment possible. Given the consistency with which initial status, length of stay and age influence discharge status, the multifaceted probabilistic models [44] already mentioned should prove to be relevant and useful in this regard.

3. Implementing probabilistic models

3.1. Software

Personal computer programs that allow the specification of probabilistic models and the testing of the extent to which data fit the model are numerous and varied. Two of the most popular programs are BIGSTEPS [81] and FACETS [44], both available from MESA Press in the Department of Education at the University of Chicago. BIGSTEPS produces 23 tables and figures displaying estimate convergence, sorted observations, scale values, errors, and four kinds of model-fit statistics for any two-faceted (item ×

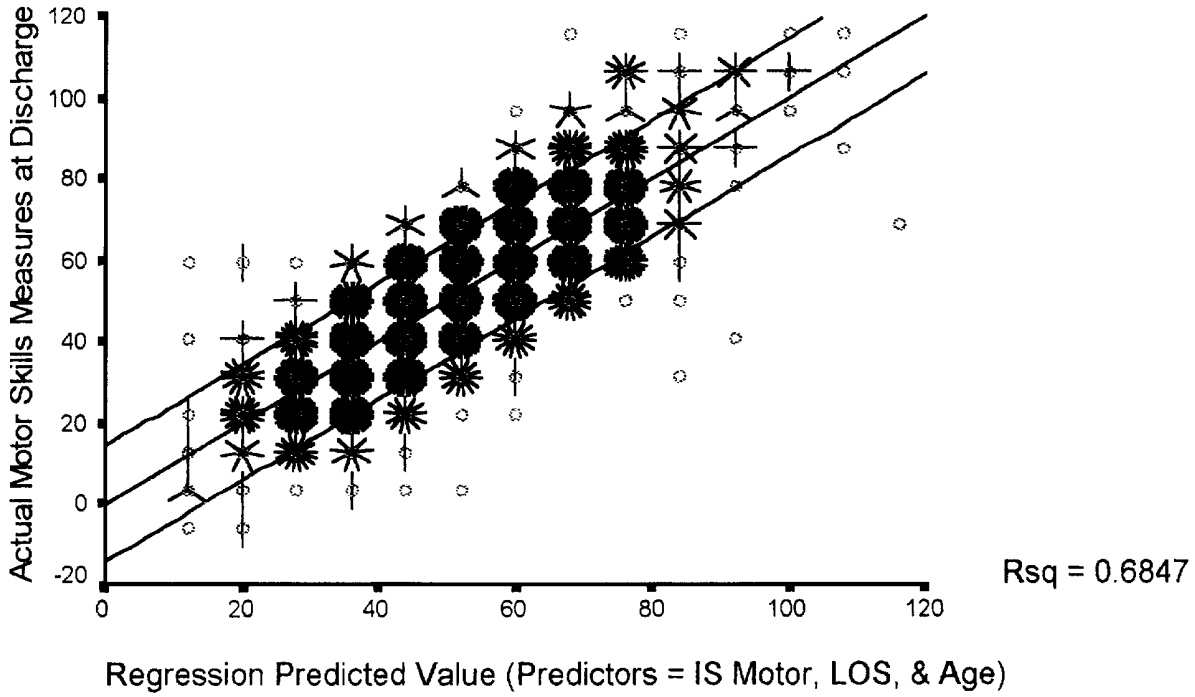


Fig. 2. PEGS Motor Skills Lifescale. Actual and predicted 1985–1993 discharge measures; 6467 subacute, inpatient, and day hospital measures.

person) binomial or rating scale data. FACETS produces much the same output for measurement designs involving three or more facets. A Windows interface for running BIGSTEPS and editing its control, data and output files has recently been made available by MESA Press.

Other programs that apply Rasch's probabilistic models are available from Peter Allerup (Danish Institute for Educational Research, Copenhagen, Denmark); David Andrich (ASCORE, Murdoch University School of Education, Western Australia); Assessment Systems Corporation (RASCAL, St. Paul, MN); Gerhard Fischer (Vienna, Austria); Richard Gershon (Computer Adaptive Technologies, Inc., Chicago, IL); Cees Glas (RPS, ProGAMMA, Groningen, The Netherlands); Ivan Horabin (TestCalc, Durham, NC); Geoff Masters (Quest, Australian Council on Education Research, Hawthorn, Victoria, Australia); Matthew Schulz (MFORM, American College Testing, Iowa City, IA); Richard Smith (Marianjoy Rehabilitation Hospital and Clinics, Wheaton, IL); and Mark Wilson (University of California-Berkeley School of Education, CA).

3.2. Creating national databases

Every scale calibrated to fit a probabilistic measurement model prepares the ground for national databases facilitating the comparisons of outcomes and treatment quality. When the calibration values of an instrument's items are found to be stable across thousands of patients, hundreds of therapists, dozens of hospitals, and years of time, as has been found to be true for both the PECS and the FIM, much of value has been learned about the variable that instrument measures. When similar items from two or more instruments designed to measure the same variable calibrate in the same order and with similar spacing, even though the calibration studies have been conducted on entirely different samples of patients (although they are from the same general population), as has happened with the PECS and the FIM Motor Skills scales, confidence in the structure of the variable is boosted still more.

When analogous instrument structures are independently calibrated and result in scale positions as similar as those found for the PECS and the FIM, the time is ripe for calibrating a gold standard for a national database. At this point, the specific diagnostic groups for whom the calibrations hold must be carefully determined. Data on both instruments on a sample of 100–200 such patients chosen as representative of the relevant diagnoses and levels of independence must be obtained. As long as all of the instruments involved have already been independently calibrated on thousands of patients, and their items already exhibit stable and similar scale values, there is no reason to demand a larger size sample than 100–200 patients for the cocalibration.

At this point a number of comparisons will be needed to ascertain the stability and constancy of the unit of measurement. Basic standards for such measurement method comparison studies exist [93,94]. In order to overcome the inadequacies of correlation coefficients, graphical methods of comparison, involving plots of scales' common item calibrations and of the measures produced on a common sample by each instrument, are recommended.

First, the scales' original independent calibrations must be compared with independent calibrations derived from the new, smaller sample. Significant variation in these independent calibrations would indicate that the samples are not sufficiently representative of the population to proceed further with the cocalibration of the instruments. Calibrations based on subsamples of each of the independent calibrations, such as can be obtained by dividing the data according to assessment status (admission versus discharge) or diagnostic group, should also be compared to further establish construct validity.

Second, each instrument's measures, as these are independently derived from the common cocalibration sample, must be compared. The relative value of patients' measures should not differ by more than an error or two across instruments. If these values differ to a statistically significant degree, the source of the variation must be identified and eliminated before proceeding.

Third, the instruments should be treated as a single super-instrument in an analysis that com-

bines all of their items into a single cocalibration. The measures produced by this cocalibration should then be compared with the measures produced by the instruments' separate, independent calibrations. If the two prior steps have produced adequate results, it is unlikely that any problems would emerge at this point.

Fourth, each instrument's items should be anchored at their cocalibration values and used to generate a new set of measures on the cocalibration sample in independent analyses. The measures resulting from each instrument's anchored values should then be compared with the cocalibration measures, with one another, and with the measures produced earlier from the unanchored independent analyses.

Should extensive data be available for one or both instruments, it may be desirable to add these instrument-specific ratings to the analysis at this point. If data are available for one instrument only, these could be added to the cocalibration analysis and the item scale values examined for variation from the smaller joint sample results. If data on separate patient samples are available for both (or all) instruments, using the cocalibration sample to link the two instruments would generate measures in a common unit of measurement for all of the persons rated in one analysis, eliminating the need to establish cocalibration values in one analysis and then to generate measures for each instrument's independent data sets in two or more separate, anchored analyses.

If all of these comparisons support the hypothesis that the measures are statistically identical (within an error of each other), it should be considered reasonable to anchor each instrument's items at their cocalibration values. The measures produced by these items can then be used for program evaluation, quality assessment and improvement, accreditation, treatment planning, research, marketing, etc., and be understood as representing constant amounts of the variable in question.

For making hospital by hospital comparisons, however, it will be necessary to adjust the measures by the variables that impact and influence gains and discharge functional independence status. A multifaceted measurement design could adjust measures according to the influence that a patient's admission status, age, length of stay, and time from condition onset to treatment have on her or his outcome. Once statistically significant groupings in the influencing variables have been identified, rating scales could be devised to label the groupings for analysis. Since these rating scales will likely vary in the number of rating points and in their substantive meaning, a partial credit model [19,72,91,92] will be required.

The FACETS [44] computer program is capable of specifying and fitting multifaceted partial credit models. FACETS would determine not only the differences between people's levels of functional independence and the difficulties of the functional assessment items, but would also adjust these differences by every other modelled facet (admission status, age, length of stay, and time since onset). The tables in the FACETS output include an all-facets variable map in which the positions of all persons, items, and rating scale categories for each facet are displayed. This table makes it possible to see the quantitative relationships among the facets modelled.

For instance, consider the amount of extra gain in functional independence associated with one rating scale unit of increased length of stay (which would represent some predetermined number of days). This extra gain would be graphically and quantitatively presented as an amount relative to the difference between two particular functional assessment items, or between two rating scale points for a single item. Gains associated with the differences between rating scale points for every other facet would be similarly presented, with the influence on the patients' functional independence attributable to each facet separated from the influence due to the others.

Finally, the validity and reliability of any instrument's calibration, once established, are not permanent. Changes in the patient population, for instance, might cause some areas of functional independence to increase or decrease in difficulty, changing the meaning of the measures. Instrument calibrations should be routinely monitored and recalibrated as a matter of course.

4. Implications for healthcare reform

4.1. The role of functional status in determining reimbursement

New health status ICD9 codes are being researched for the purposes of setting up a prospective payment system for ambulatory care [100]. The Health Care Financing Administration (HCFA), Office of Research and Demonstrations, has supported studies utilizing 27 PECS items, most of which have been modified from their seven-point rating scales to a four-point scale for application at the higher levels of functional independence exhibited by clients of outpatient clinics, who are the focus of the study. Should this system be implemented, will it require the administration of this particular set of PECS self care, ADL, mobility, speech, psychology, and pain items? Could the H-codes also be derived from the FIM or the PECS, or any other instrument that measures the relevant constructs?

With the emergence of the Rehabits concept, the answers to these questions may be surprising. An instrument that measures only motor skills and cognition, such as the FIM, will not, of course, be able to capture information outside of those domains. However, should the motor and cognition dimensions be sufficient for predicting resource consumption in some specific instances, then any instrument that can be shown to measure these constructs through co-calibration with the H-codes will be able to use the resulting Rehabits as a medium for converting scores between scales for those instances. From this perspective, the answer to the question 'is it necessary for HCFA to mandate the use of one particular functional assessment instrument for reimbursement, probably the FIM?' is 'no'.

4.2. Levels of care

The need for more careful management of clients' financial resources is leading to a notion of 'cascading', or moving clients through a series of levels of care targeting the intensity, duration, and site of treatment according to the best medical assessment of that person's capacity to benefit. [55,101–105]. As rehabilitation moves away from its traditional discipline-oriented focus toward diagnostic group-oriented product lines, functional assessment instruments calibrated to measure the complete range of impairments, disabilities, and handicaps treated in SNF-based, acute care units, free-standing in patient facilities, day hospital, outpatient, transitional living centers, and home health contexts will be used to monitor the flow and performance of clients through the system [100–102]. Fig. 3 is a hypothetical graphic report that could be used to convey this information; it shows how admission and discharge statuses, gains, and admission and discharge percents independent, for different levels of care, can be compared and related to one another on a single continuum of functional independence measurement. Figs. 4 and 5 show the same measures, for a hypothetical neuromuscular rehabilitation program, on each of the five PECS LifeScales.

Although these graphs can be understood as saying that a person with a measure of 20 has less functional independence than someone with a measure of 40, the numbers do not indicate just what people with these measures can and cannot do. Figs. 3 and 4 need an interpretive aid that connects the measures shown back to the concrete substance of the original ratings, thereby taking advantage of the transparency of the instruments. Fig. 6 provides that information for any measure on the five scales. Because the numerical scales in the graphs are in the same unit as the grid of descriptive text in Fig. 6, it is easy to connect the graphics to the narrative description of abilities.

Fig. 7 is a format that puts more of the information relevant to the interpretation of the measures on one page. The ranges of measures relevant to each box in the grid in Fig. 7 are the same as the skill bands shown on the left side of each measurement continuum in Fig. 6; these are ranges in the scale that differ to statistically significant degrees. Some scales are more sensitive than others, so they have different numbers of skill bands. Fig. 6 can be configured as a report on individual patients for facilitating communication in treatment team conferences, with referring physicians, or with case managers. It could also work as a program evaluation report, showing

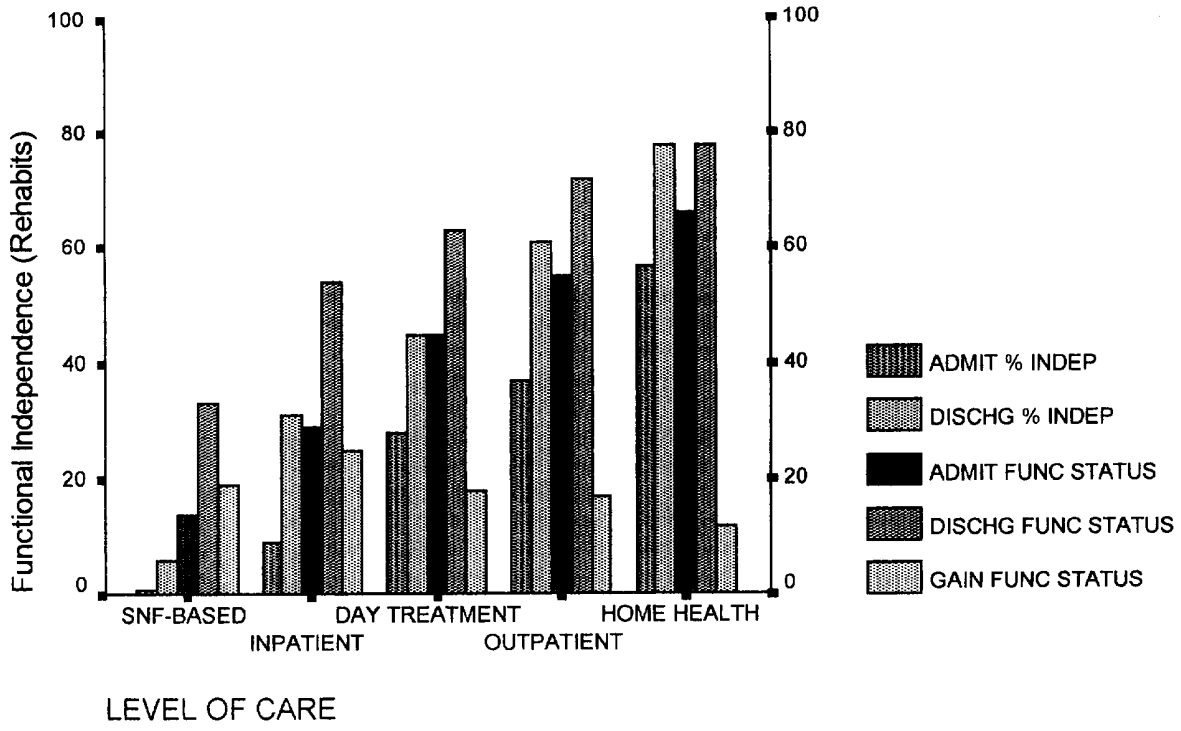


Fig. 3. Idealized change in functional independence across levels of care.

average measures for groups of patients in a specific treatment program during a specific time period. As such, it might also be used to indicate historical standards, such as when a prior fiscal year's performance is used as a basis for evaluating program performance in the current quarter.

If functional status were universally measured in the same units, reports such as these would communicate to each of their readers the information they need to best perform their jobs. The transparency of the instruments would allow clinicians to see the finest levels of detail in a measure, enabling them to know immediately what safety precautions and what form of treatment are relevant. Managers, third party payors, and referring physicians would be able to answer larger questions concerning patients' abilities to dress, feed, or bathe themselves at a glance. Accreditors who travel from facility to facility would no longer be hampered by units of measurement that vary from one program evaluation department to the next. The persons treated will benefit not just from the lower costs that each of the foregoing points will make possible, but from research results that are more easily communicated, evaluated, and implemented.

5. Conclusion

The computerized medical record is expected to evolve from its current centralization within individual treatment sites (hospitals, doctors' and dentists' offices, pharmacies) to being distributed within those sites to being longitudinal and virtual [106,107]. The evolutionary leap from the centralized to distributed record is being driven by the shift from mainframes and minicomputers to client/server architectures. As the infrastructure of the information superhighway grows, the computerized medical record will be decentralized and accessible from any treatment site that the patient chooses. Among the many procedural,

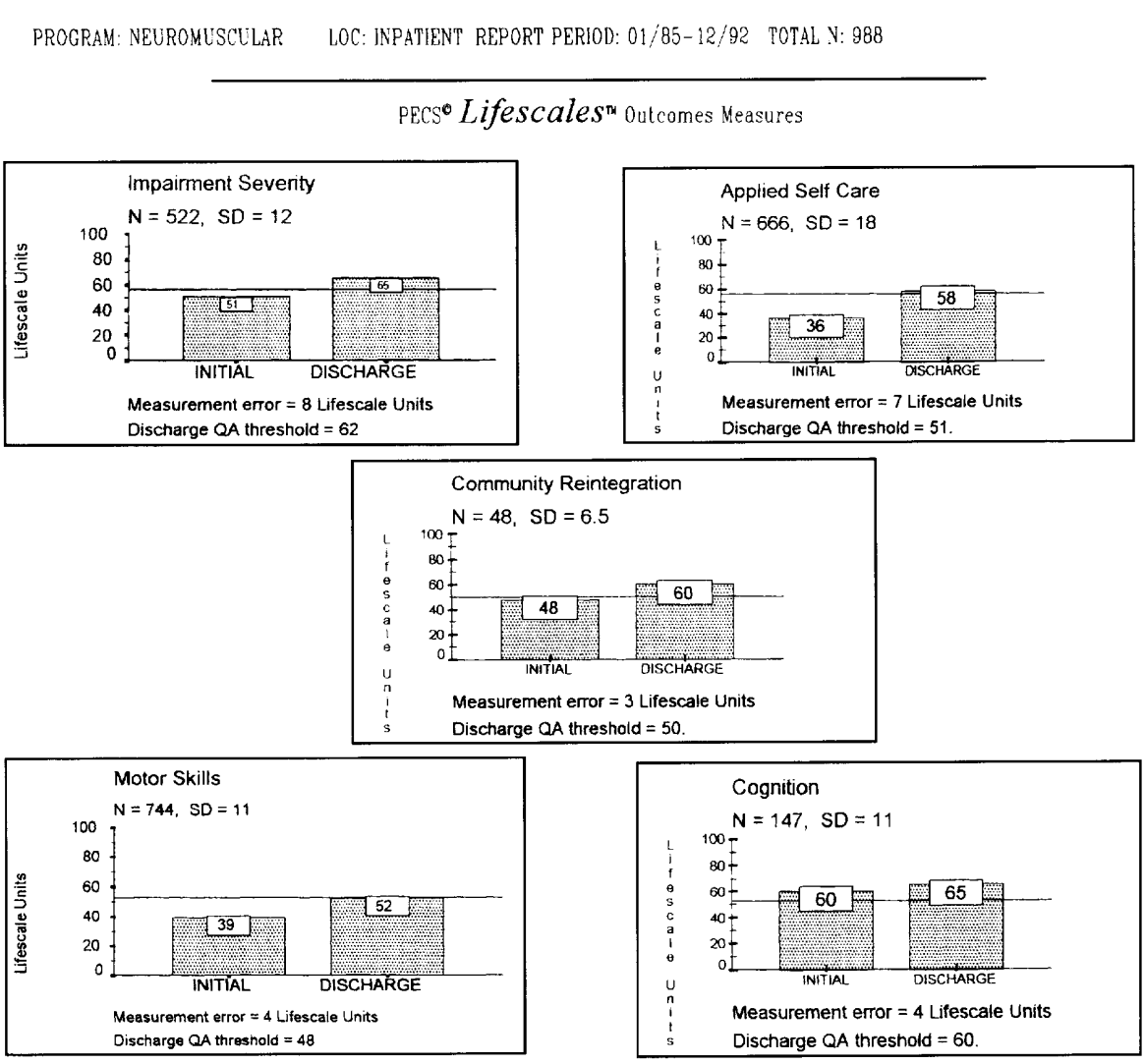


Fig. 4. Quarterly program monitor and evaluation report: initial and discharge functional status. The horizontal line in each chart represents the transition from dependence to independence. Only patients who completed treatment are included.

technical, and legal issues that need to be resolved to make this scenario a reality is the problem of the dozens of data standards that will have to be set and agreed upon.

Standard functional status, health status, quality of life, and other satisfaction, attitude, and ability measures are only a small part of the communications and connectivity problems we face in fostering the growth of the planet's computerized neurology. The problem of gold standards is an imposing one, however, as even a brief consideration of historical metrology [108] shows. The role of measurement standards in the political and economic history of Europe has been

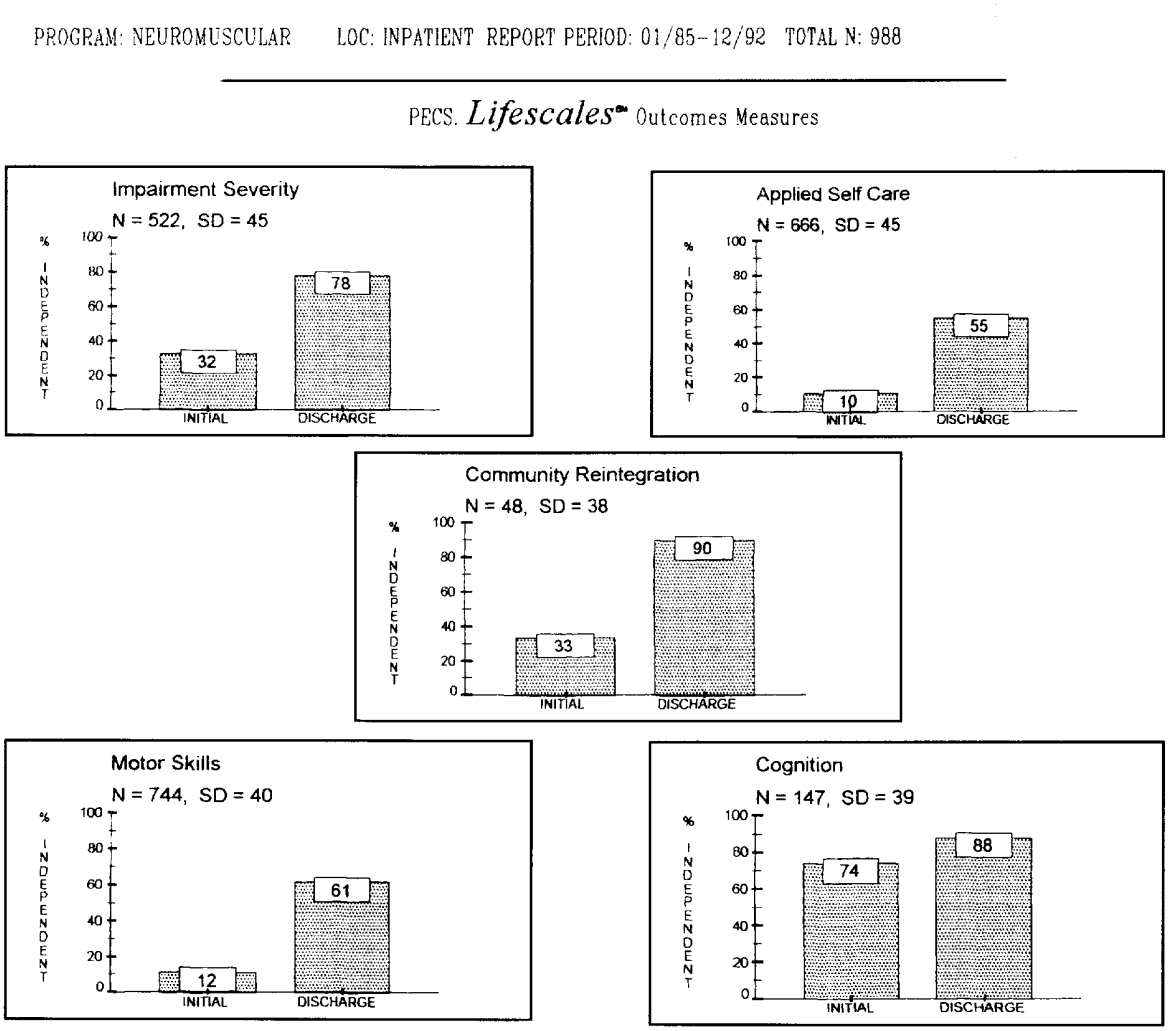
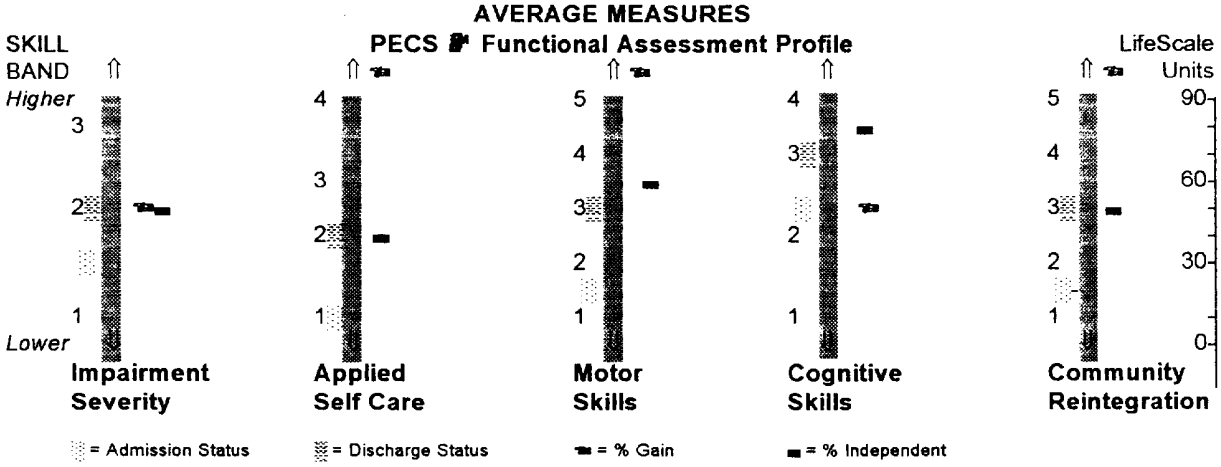


Fig. 5. Quarterly program monitor and evaluation report initial and discharge percent independent. Only patients completing treatment included.

crucial, with huge amounts of taxes, profits, and commercial value, not to speak of cultural authenticity and identity, riding on measurement consistency. To date, the measurement of functional independence has been akin to the weights and measures of the Middle Ages, when every burg had its own standards. The techniques and tools of rating scale measurement have thus far been confined largely to the academic researches of educators and psychologists. The information superhighway is about to place these endeavors in a far more public forum. Let us work together to create and adopt rating scale measurement standards that will assure uniform data quality, be relatively painless to implement, and be enforceable.

Base Period: FY94 Program: Neuromuscular % Discharged Home: 80%
Level of Care: Inpatient Satisfaction: 6.2 Avg Age: 68 Avg LOS: 26



Skill Levels

NEUROMUSCULAR PROGRAM CLIENTS IN THESE SKILL BANDS TYPICALLY:

Impairment Severity Band 2
- exhibit normal range of motion in least involved extremities and spine, with moderate limitations of involved extremity(s);
- exhibit normal tone/no dystonia of least involved extremities and trunk, with moderate spasticity or dystonia of involved extremity(s).

Applied Self Care Band 2
- 50-74% effective bowel & bladder programs;
- complete or directs completion of at least 50%, but not 100%, of self care activities, and are dependent on cues.

Motor Skills Band 3
- can feed themselves with supervision and cues to stay on task;
- are able to assist with more than 75% of wheelchair activity;
- are able to walk with intermittent physical assistance.

Cognitive Skills Band 3
- are alert within normal limits;
- exhibit selective inattention problems;
- exhibit verbal language skills adequate for daily living situations.

Community Reintegration Band 3
- demonstrate at most a mild loss of behavior control in moderately stressful situations, and regain control quickly;
- have families that consistently assist in the client's adjustment, but have difficulty incorporating understanding of disability in practice
- will occasionally participate in leisure activities, with encouragement

Fig. 6. Rehabilitation outcome reporting system with PECS©.

Skill Band	Impairment Severity	Applied Self Care	Motor Skills	Cognitive Skills	Community Re-entry
		NEUROMUSCULAR CLIENTS IN THESE SKILL BANDS TYPICALLY:			
5			• feed, ambulate, and dress within normal limits, or nearly so; • prepare complete meals and do housework with no supervision, but with adaptive equipment.		• verbally communicate needs; converses; • learn new behaviors with slight external structure • participate fully & gains full benefit from group activities
4		• regularly adhere to bowel & bladder programs; • performs or directs activities safely, and knows medications, with lapses fewer than 25% of the time.	• feed without assistance; • ambulate independently with assistive device • dress and bathe UE independently; • dress and bathe LE independently with adaptive equipment.	• are alert, oriented, & produce speech within normal limits; • read complex material; writes with minimal difficulty; • initiate discussion of needs; seeks guidance.	• communicate needs verbally with difficulty; • are independent in living 50-75% of time; • are emotionally distressed by condition & prognosis.
3	• exhibit mild to no range of motion limits in involved extremity(s); • exhibit mild to no motor loss and involuntary movements in one extremity	• are 100% able to direct bowel/bladder program, with physical assistance; • are > 50% able to perform/direct activities safely w/ cues; • needs cues for 50% knowledge of medications	• must be supervised in feeding & UE bathing; • require minimal physical assistance in UE/LE dressing, LE bathing, & walking; • require moderate assistance in light meal preparation.	• exhibit minimal impairment of alertness; • have mildly impaired long term memory, basic intellectual skills, orientation, and attention; • read phrases & concrete sentences.	• learn with consistent external structure & behavior reinforcement; • have family beginning to enhance adaptation; • are independent in living 25-50% of time; • are aware of condition, do not accept prognosis.
2	• exhibit moderate to severe range of motion limits in involved extremity(s); • exhibit moderate motor loss & severe spasticity/dystonia in 2 extremities/trunk.	• are about 75% able to direct bowel/bladder program, w/ assistance; • are < 50% able to perform/direct activities safely w/ cues; • needs cues for 75% knowledge of medications.	• require minimal assistance in feeding & UE bathing; • require moderate assistance in UE/LE dressing & walking; • require maximal assistance in meal preparation.	• respond to stimulus automatically or inconsistently; • quickly forget information; confabulate; • produce automatic or imitative language; • copy letters, numbers, some words, reads words.	• require repetitive, simple instructions to learn; • can, but do not, express feelings or needs; • have impaired conversational skills, but will answer direct questions; • are 100% dependent.
1	• exhibit severe range of motion limits & postural deviations in all extremities and spine; • exhibit severe motor loss and dystonia or spasticity in all 4 extremities	• are less than 50% able to direct or participate in bowel/bladder program; • may show no awareness of safety issues; • may show no knowledge of medications; • completes no self care activities.	• require moderate assistance in feeding; • require maximal to moderate assistance in UE bathing, bed position changes, & hygiene; • require maximal assistance in all other mobility skills & ADLs.	• have severe to marked impairment of alertness; • are extremely to severely impaired is long term memory, basic intellectual skills, orientation, and attention; • show little or no comprehension of printed or auditory stimuli.	• exhibit uncontrolled behavior; • exhibit minimal or no skills in interpersonal relations; • are unaware of disability.

Fig. 7. PECS© LifeScales' interpretive guidelines.

Acknowledgement

This work was supported in part by a grant from the Donald P. and Byrd M. Kelly Foundation to the Center for Rehabilitation Outcome Analysis.

References

[1] Kuhn TS: The function of measurement in modern physical science. Isis 1961;52(168):161–193.

[2] Kuhn TS: Metaphor in science. In: Ortony A, ed. Metaphor and thought, Cambridge UK: Cambridge University Press, 1979.

[3] Rothbart D: The semantics of metaphor and the structure of science. Philos Sci 1984;51:595–615.

[4] Gerhart M, Russell A: Metaphoric process: the creation of scientific and religious understanding, Fort Worth: Texas Christian University Press, 1984.

[5] MacCormac E: Metaphor and myth in science and religion, Durham NC: Duke University Press, 1976.

[6] Hesse M: Models and analogies in science, Notre Dame: University of Notre Dame Press, 1970.

[7] Black M: Models and metaphors, Ithaca: Cornell University Press, 1962.

[8] Hacking I: Representing and intervening: introductory

topics in the philosophy of natural science, Cambridge UK: Cambridge University Press, 1983.
[9] Heelan P: Natural science as a hermeneutic of instrumentation. Philos Sci 1983;50:181–204.
[10] Ackermann JR: Data instruments and theory: a dialectical approach to understanding science. Princeton: Princeton University Press, 1985.
[11] Ihde D: Instrumental Realism: the interface between philosophy of science and philosophy of technology, In: Ihde D, ed. The Indiana Series in the Philosophy of Technology, Bloomington: Indiana University Press, 1991.
[12] Lakatos I, Musgrave A eds. Criticism and the growth of knowledge. New York: Cambridge University Press, 1970.
[13] Latour B: Science in action: how to follow scientists and engineers through society. Cambridge MA: Harvard University Press, 1987.
[14] Merbitz C, Morris J, Grip JC: Ordinal scales and foundations of misinference. Arch Phys Med Rehabil 1989;70:308–312.
[15] Wilson M: A comparison of deterministic and probabilistic approaches to measuring learning structures. Aust J Educ 1989;33:127–140.
[16] Rasch G: Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogiske Institut, 1960; reprint, Chicago: University of Chicago Press, 1980.
[17] Wright BD: Sample-free test calibration and person measurement. In: Proceedings of the 1967 Invitational Conference on Testing Problems Princeton, NJ: Educational Testing Service, 1968.
[18] Wright BD: Solving measurement problems with the Rasch model. J Educ Meas 1977;14(2):97–116.
[19] Wright BD, Masters G: Rating scale analysis. Chicago: MESA Press, 1982.
[20] Wright BD, Linacre JM: Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989;70:857–860.
[21] Andrich D: Rasch models for measurement, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-068. Beverly Hills: Sage Publications, 1988.
[22] Fischer GH: Einfuhrungen in die Theorie psychologischer Tests. Bern, Stuttgart, Wein: H. Huber, 1974.
[23] Michell J: An introduction to the logic of psychological measurement. Hillsdale, NJ: Lawrence Erlbaum, 1990.
[24] Thurstone LL: The measurement of social attitudes. J Abnl Soc Psychol 1931;26:249–269; reprinted in Thurstone LL, The measurement of values. Chicago: University of Chicago Press, Midway Reprint Series, 1959:287–303.
[25] Thurstone LL: Psychology as a quantitative rational science. Science 1937;85:228–232; reprinted in Thurstone LL, The measurement of values. Chicago: University of Chicago Press, Midway Reprint Series, 1959:3–11.
[26] Guttman L: The basis for scalogram analysis. In: Stouffer SA et al.:, eds. Measurement and prediction. New York: John Wiley and Sons, 1950.
[27] Bradley RA, Terry ME: Rank analysis of incomplete block designs. I: The method of paired comparisons. Biometrika 1952;39:324–345.
[28] Luce RD, Tukey JW: Simultaneous conjoint measurement: A new kind of fundamental measurement. J Math Psychol 1964;1:1–27.
[29] Krantz DH, Luce RD, Suppes P et al.: Foundations of measurement. Vol. 1: additive and polynomial representations. New York: Academic Press, 1971.
[30] Anderson NH: How functional measurement can yield validated interval scales of mental qualities. J Appl Psychol 1976;61:677–692.
[31] Katz S, Downs TD, Cash HR et al.: Progress in development of Index of ADL. Gerontoloqy 1970;10:20–30.
[32] Stewart AL, Ware JE Jr, Brook RH: Advances in measurement of functional status: construction of aggregate indexes. Med Care 1981;19:473–488.
[33] Williams RGA, Johnston M, Willis LA et al.: Disability: a model and measurement technique. Br J Prev Soc Med 1976;30:71–78.
[34] Carter WB, Bobbitt RA, Bergner M et al.: Validation of an interval scaling: the Sickness Impact Profile. Health Services Res 1976;11:516–528.
[35] Kirshner B, Guyatt G: A methodological framework for assessing health indices. J Chron Dis 1985;38:27–36.
[36] Guyatt GH, Kirshner B, Jaeschke R: Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992;45:1341–1345.
[37] Williams JI, Naylor CD: How should health status measures be assessed? Cautionary notes on Procrustean frameworks. J Clin Epidemiol 1992;45:1347–1351.
[38] Guyatt GH, Kirshner B, Jaeschke R: A methodological framework for health status measures: clarity or oversimplification? J Clin Epidemiol 1992;45:1353–1355.
[39] Spitzer WO: State of science 1986: quality of life and functional status as target variables for research. J Chron Dis 1987;40:465–471.
[40] Costich J: Assessment of measurement practices in medical rehabilitation: measure users and their audiences. Phys Med Rehabil Clin North Am 1993:4:587–594.
[41] Fisher WP: Truth, method, and measurement: the hermeneutic of instrumentation and the Rasch model. Dissertation, University of Chicago, 1988.
[42] Fisher WP: Objectivity in measurement: a philosophical history of Rasch's separability theorem. In: Wilson M, ed. Objective measurement: theory into practice. Norwood NJ: Ablex, 1992:29–58.
[43] Ottenbacher KJ, Tomchek SD: Measurement in rehabilitation research: consistency vs consensus. Phys Med Rehabil Clin North Am 1993;4:463–473.
[44] Linacre JM: Many-facet Rasch measurement. Chicago: MESA Press, 1989.
[45] Lunz ME, Wright BD, Linacre JM: Measuring the impact of judge severity on examination scores. Appl Meas Educ 1990;3/4:331–345.

[46] Lunz ME, Stahl JA: Judge consistency and severity across grading periods. Eval Health Professions 1990;13:425–444.

[47] Lunz ME, Stahl JA: The effect of rater severity on person ability measure: a Rasch model analysis. Am J Occup Ther 1993;47:311–317.

[48] Feinstein AR: Clinimetrics. New Haven: Yale University Press, 1987.

[49] Fisher WP, Harvey RF, Kilgore KM et al.: Applying transparency in Rasch measurement reporting [abstract]. Arch Phys Med Rehabil 1993;74:1252.

[50] Fisher WP: Measurement-related problems in functional assessment. Am J Occup Ther 1993;47:331–347.

[51] Harvey RF, Lambert RW: Rating scale analysis of a functional assessment instrument in physical rehabilitation [abstract]. Arch Phys Med Rehabil 1987;68:583–584.

[52] Silverstein B, Kilgore KM, Fisher WP: Implementing patient tracking systems and using functional assessment scales, In: Harvey RF, ed. Center for Rehabilitation Outcome Analysis monograph series on issues and methods in rehabilitation outcome analysis, vol. 1. Wheaton IL: Marianjoy Rehabilitation Center, 1989.

[53] Silverstein B, Fisher WP, Kilgore KM et al.: Applying psychometric criteria to functional assessment in medical rehabilitation: II. Defining interval measures. Arch Phys Med Rehabil 1992;73:507–518.

[54] Kilgore KM, Fisher WP, Silverstein B et al.: Application of Rasch analysis to the Patient Evaluation and Conference System (PECS). Phys Med Rehabil Clin North Am 1993;4:493–515.

[55] Harvey RF, Silverstein B, Kilgore KM et al.: Applying psychometric criteria to functional assessment in medical rehabilitation: III. Construct validity and predicting level of care. Arch Phys Med Rehabil 1992;73:887–892.

[56] Wright BD, Linacre JM, Heinemann AW: Measuring functional status in rehabilitation. Phys Med Rehabil Clin North Am 1993;4:475–491.

[57] Linacre JM, Heinemann AM, Wright BD et al.: The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1904;75:127–132.

[58] Heinemann AW, Hamilton BB, Granger CV et al.: Rating scale analysis of functional assessment measures. Final Report to the National Institute on Disability and Rehabilitation Research. Chicago IL: Rehabilitation Institute of Chicago, 1991.

[59] Heinemann AW, Linacre JM, Wright BD et al.: Relationships between impairment and physical disability as measured by the Functional Independence Measure. Arch Phys Med Rehabil 1993;74:566–573.

[60] Velozo CA, Magalhaes L, Pan AW et al.: Measurement qualities of the Level of Rehabilitation Scale-III (LORS-III) [abstract]. Arch Phys Med Rehabil 1993;74:661.

[61] Velozo CA, Pan AW, Magalhaes MS et al.: Nurse–therapist differences in ADL assessment [abstract]. Arch Phys Med Rehabil 1994;75:720.

[62] Ludlow LH, Haley SM: Polytomous Rasch models for behavioral assessment: the Tufts Assessment of Motor Performance. In: Wilson M, ed. Objective measurement: theory into practice (vol.1). Norwood NJ: Ablex. 1992:121–137.

[63] Kalinowski A: Measuring clinical pain. J Psychopathol Behav Assess 1985;7:329–349.

[64] McArthur DL, Cohen MJ, Schandler SL: Rasch analysis of functional assessment scales: An example using pain behaviors. Arch Phys Med Rehabil 1991;72:296–304.

[65] Gruft J, Fisher WP, Kelly CK et al.: Pain: Subjective or objective assessment? [abstract]. Arch Phys Med Rehabil 1993;74:1264.

[66] Olsen J, Sabroe S: Screening for neuropsychiatric dysfunction. Scand J Soc Med 1984;12:55–63.

[67] Haley SM, Ludlow LH, Coster WJ: Pediatric Evaluation of Disability Inventory: clinical interpretation of summary scores using Rasch rating scale methodology. Phys Med Rehabil Clin North Am 1993;4:529–540.

[68] Msall ME, DiGaudio KM, Duffy LC: Use of functional assessment in children with developmental disabilities. Phys Med Rehabil Clin North Am 1993;4:517–527.

[69] Campbell SK, Osten ET, Kolobe THA et al.: Development of the test of infant motor performance. Phys Med Rehabil Clin North Am 1993;4:541–550.

[70] Fisher WP, Pugliese K: Measuring the importance of pastoral care in rehabilitation [abstract]. Arch Phys Med Rehabil 1989;70:A-22.

[71] Pugliese K, Fisher WP, Kelly CK et al.: Accountability in pastoral care: can spiritual well-being be measured? [abstract]. Arch Phys Med Rehabil 1993;74:1270.

[72] Willmes K: Psychometric evaluation of neuropsychological test performances. In: von Steinbüchel N, von Cramon DY, Pöppel E, eds. Neuropsychological rehabilitation, New York: Springer, 1992:103–113.

[73] Fisher WP: Measuring rehabilitation client satisfaction. Paper presented at the semi-annual Midwest Objective Measurement Seminar, University of Chicago, Dec. 1992.

[74] Fisher AG, Bryze KA, Granger CV et al.: Applications of conjoint measurement to the development of functional assessments. Int J Educ Res 1994;21(6):579–593.

[75] Fisher AG: Development of a functional assessment that adjusts ability measures for task simplicity and rater leniency. In: Wilson M, ed. Objective measurement: theory into practice (vol. 2). Norwood NJ: Ablex. 1994:145–175.

[76] Fisher WP, Fisher AG: Application of Rasch analysis to studies in occupational therapy. Phys Med Rehabil Clin North Am 1993;4:551–569.

[77] Fisher AG: The assessment of IADL Motor Skills: an application of many-faceted Rasch analysis. Am J Occup Ther 1993;47:319–329.

[78] Loevinger J: Objective tests as instruments of psychological theory. Psychol Red 1957;3:635–694.

[79] Messick S: The standard problem: meaning and value in measurement and evaluation. Am Psychol 1975;30: 955–966.

[80] Loevinger J: Person and population as psychometric concepts. Psychol Rev 1965;72:143–155.
[81] Wright BD, Linacre JM: A users' guide to BIGSTEPS Rasch model computer program. Chicago: MESA Press, 1991.
[82] Harvey RF, Jellinek HM: Functional performance assessment: a program approach. Arch Phys Med Rehabil 1981;62:456–461.
[83] Choppin B: An item bank using sample-free calibration. Nature 1968;219:870–872.
[84] Wright BD, Bell SR: Item banks: What, why, how. J Educ Meas 1984;21:331–345.
[85] Lunz M, Bergstrom B, Gershon R: Computer adaptive testing. Int J Educ Res 1994;21:623–634.
[86] Wright BD, Stone MH: Best test design. Chicago: MESA Press, 1979.
[87] Woodcock R: Woodcock reading mastery tests. Circle Pines MN: American Guidance Counseling, 1974.
[88] Fisher WP, Harvey RF, Taylor P et al.: Rehabits: towards a common language of functional assessment. Arch Phys Med Rehabil 1995;76(2): in press.
[89] Heinemann AW, Linacre JM, Wright BD et al.: Relationships between impairment and physical disability as measured by the Functional Independence Measure. Arch Phys Med Rehabil 1993;74:566–573.
[90] Kilgore KM, Fisher WP, Harvey RF et al.: Diagnosis-based differences in Rasch calibrations of functional assessment scales [abstract]. Arch Phys Med Rehabil 1993;74:1254.
[91] Masters GN: A Rasch model for partial credit scoring. Psychometrika 1982;47:149–174.
[92] McArthur DL, Casey KL, Morrow TJ et al.: Partial-credit modeling and response surface modeling of biobehavioral data. In: Wilson M, ed. Objective measurement: theory into practice (vol. 1). Norwood NJ: Ablex. 1992:109–120.
[93] Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i(8476):307–310.
[94] Ottenbacher KJ, Tomchek SD: Measurement variation in method comparison studies: An empirical examination. Arch Phys Med Rehabil 1994;75:505–512.
[95] Sulton LD, Hardisty B, Bisterfeldt J et al.: Computerized databases: an integrated approach to monitoring quality of patient care. Arch Phys Med Rehabil 1987;68:850–-853.
[96] Bisterfeldt J, Fisher WP, Cicero C et al.: Basic program evaluation and quality monitoring. Presented to the Workshop on Rehabilitation Outcome Analysis: State-of-the-Art Management Applications, Wheaton, IL: Marianjoy Rehabilitation Hospital and Clinics, May 1993.
[97] Fisher WP, Bisterfeldt J, Lynam N et al.: Evaluating patient progress and program outcome: Utilization review, program evaluation, and quality assurance. Presented to the Workshop on Rehabilitation Outcome Analysis: State-of-the-Art Applications in Stroke, Wheaton, IL: Marianjoy Rehabilitation Hospital and Clinics, Oct 1992.
[98] Kilgore KM, Silverstein BJ, Fisher WP: Using outcome prediction for quality assurance and program evaluation [abstract]. Arch Phys Med Rehabil 1990;71:779.
[99] Silverstein BJ, Kilgore KM, Fisher WP: A computerized communication system providing case managers and referring physicians with concise and comprehensive patient status and progress information [abstract]. Arch Phys Med Rehabil 1990;71:779.
[100] Averill RF, Goldfield NI, Wynn ME et al.: Design and evaluation of a prospective payment system for ambulatory care. Wallingford CT: 3M Health Information Systems, 1992.
[101] Silverstein BJ, Venzon M, Harvey RF: Cost-effectiveness of rehabilitation levels of care for medicaid patients [abstract]. Arch Phys Med Rehabil 1991;72:790.
[102] Harvey RF: The future of rehabilitation: Delivery of rehabilitation services in the 1990s. Phys Med Rehabil State Art Rev 1987;1:321–329.
[103] Harvey RF, Griffin MD, Ressel T: The future of rehabilitation. Phys Med Rehabil State Art Rev 1993;7:421–430.
[104] Harvey RF: The emergence of rehabilitation levels of care: the concept and the challenge. Welcoming address to the CME Seminar, From Disability to Community: The Rehabilitation Levels of Care. Wheaton, IL: Marianjoy Rehabilitation Hospital & Clinics, May, 1994.
[105] Fisher WP: Outcome measurement strategies for levels of care: You can compare apples and oranges. Paper presented to the CME Seminar, From Disability to Community: The Rehabilitation Levels of Care. Wheaton, IL: Marianjoy Rehabilitation Hospital & Clinics, May, 1994.
[106] Information technology in healthcare: succeeding in a changing market. Waltham MA: Decision Resources, Inc., 1994.
[107] Wallace S: The computerized patient record. BYTE 1994; 19(5):67–76.
[108] Zupko RE: British weights and measures: a history from antiquity to the seventeenth century. Madison WI: University of Wisconsin Press, 1977.