Rehabs: A Common Language of Functional Assessment

William P. Fisher, Jr, PhD, Richard F. Harvey, MD, Patricia Taylor, MSW, Karl M. Kilgore, PhD, Carol K. Kelly, PhD

ABSTRACT. Fisher WP, Harvey RF, Taylor P, Kilgore KM, Kelly CK. Rehabit: a common language of functional assessment. Arch Phys Med Rehabil 1995;76:113-22. Probabilistic measurement models offered by Rasch and others can be used to link different functional assessment instruments into a single measurement system. This study assessed 54 subjects (diagnoses: 8 brain injuries, 7 neuromuscular, 22 musculoskeletal, 7 spinal cord, 10 stroke) admitted to a free-standing rehabilitation hospital at admission and discharge using both the Functional Independence Measure (FIM) and the Patient Evaluation and Conference System (PECS). Thirteen FIM and 22 PECS motor skills items were scaled together into a 35-item instrument, providing scale values for all items in the same unit of measurement. Separate FIM and PECS measures produced for each subject correlate .94 and .91 (p < .0001), respectively, with the recalibration measures, and 0.91 (p < .0001) with each other. Either instrument's ratings are easily and quickly converted into the other's using the common unit of measurement, the rehabit (rehabilitation measuring unit). This article argues that the stability of the PECS and FIM item difficulty estimates over thousands of subjects, dozens of hospitals, hundreds of raters, and years of assessment is convincing evidence in support of the widespread use of their recalibrated, common scale values as a functionometric ruler.

© 1995 by the American Congress of Rehabilitation Medicine and the American Academy of Physical Medicine and Rehabilitation

The word rehabit is a contraction of rehabilitation functional assessment measuring units. This is the name we propose for an empirically based, scale-free, and sample-free functionometry—the quantitative study of function. The empirical basis of the measuring system is emphasized because of the possibility that a common language of functional assessment will be decided by committee, making the use of one particular brand of functional assessment instrument mandatory for reimbursement or quality assessment purposes. This need not be the case.

Scale- and sample-free measurement, measurement that does not depend on the brand name of the instrument used, nor on the particular group of persons supplying data, for the size, quality, and meaning of its units, has a history of success in education and psychology. Scale-free measurement dates at least as far back as L. L. Thurstone's work at the University of Chicago in the 1920s, which is also where the Danish mathematician Georg Rasch lectured on some closely associated ideas in 1960. Benjamin Wright concurs: "An acceptable method of scaling must result in a derived scale which is independent of the original scale and of the original group tested." Rasch stated this idea mathematically, in what he called a separability theorem, requiring the data modeled to support comparisons of subjects without regard to the items, and comparisons of items without regard to the subjects.

Loevinger wrote that "Rasch must be credited with an outstanding contribution to one of the two central psychometric problems, the achievement of nonarbitrary measures. . . . When his model fits, the results are independent of the sample of persons and of the particular items within some broad limits. Within these limits, generality is, one might say, complete." Complete generality and results that are independent of the sample of persons measured and the particular items measuring means that a scale's structure remains invariant across all samples of persons for whom the items are relevant and reasonably well-targeted (ie, of a difficulty level that is neither very much greater nor very much less than the persons' ability levels).

The feasibility of obtaining data that achieve these high standards has been noted in hundreds of applications in the last 40 years. Routine applications of Rasch measurement models to test and rating scale data are performed by companies such as the Psychological Corporation, school districts in Portland, Phoenix, Chicago, and New York, and medical school admissions and certification boards, including the National Board of Medical Examiners, the American Society of Clinical Pathologists, the American Dental Association, the American Council of State Boards of Nursing, and others.
The purpose of this article is to transfer the technical procedures of test equating used in the educational applications of Rasch's probabilistic models to the cocalibration of two functional assessment systems, the Patient Evaluation and Conference System (PECS)\textsuperscript{9,10} and the Functional Independence Measure (FIM).\textsuperscript{11,12} Though a table showing rehabili- 

ty for the PECS and the FIM is provided, the data set from 

which the repertoire has been derived may be insufficient to 

support widespread practical application at this time. Future 

research will draw together more assessment instruments 

and sufficient numbers of rehabilitation patients to support 

universal use of common units for the measurement of func-


tional independence.

**REHABITS PRECEDENTS**

All of the physical measures whose empirically derived, 

scale-free, and sample-free properties we take for granted 

emerged from contexts like that of functional assessment 

today, in which measuring units are commonly decided by 

committee, and depend on the particular instrument used, 

where it is used, or the sample of persons measured, for 

their size, quality, and meaning.

For instance, until Newton realized the implications of 

Galileo's experiment at Pisa, the length of an hour varied by 

season, latitude, and the type of clock (sundial or clepsydra, a 

kind of water clock) in use. The ancient Egyptians, for in-

stance, used sundials with equally spaced divisions, meaning 

that midday hours, when the shadows move fastest, were 

shorter than morning or evening hours. The measurement of 

time was further confounded by the fact that the Egyptians 

defined the day as half-light and half-dark, so that there 

were 12 hours of daylight regardless of the season or one's 

distance from the equator, which caused the length of the 

hour to vary by the time of year and by geography, as well 

as by time of day.

These cumbersome systems remained in use in Europe 

until Ibn Yunis (c 1200), realized that, because all weights fall 

at the same speed, all pendula of a given length have the 

same period, which makes them useful for deriving an empiri-

cally based chronometry. A simple pendulum 24.82 cm long 

has a period of almost exactly 1 second no matter where on 

earth one might be. This constancy made it possible to mea-

sure time uniformly across seasons and latitudes. Accurate 

chronometers greatly improved the capacity to navigate open 

seas, making possible landfall within dozens, instead of hun-

dreds, of miles of one's goal. How might accurate func-

tionometry improve the capacity of physical medicine and 

rehabilitation practitioners to navigate the seas of impair-

ment, disability, and handicap?

**PHILOSOPHICAL ISSUES**

The physical sciences are in many ways nothing but mea-

surement and the analysis of measures, but little thought 

went into how measurement systems are devised until re-

cently. In the human sciences, measurement methods are a 

nearly constant source of debate and theoretical interest. The 

problem with rating scale- and test-based measurement is 

that human abilities and attitudes do not seem to have any 

basis for the concatenation operation—the repetition of a 

single physical unit that adds up in the same way as a wood 

block laid end-to-end with itself. Many measurement theore-

ticians, following Campbell,\textsuperscript{13} have asserted that a concrete 

form of concatenation is essential for measurement. Unless 

a variable lends itself to an operation analogous to laying 

blocks of unit length end-to-end, or stacking bars of unit 

weight in a balance beam, many theoreticians consider the 

variable impossible to measure.

Psychologists have exhibited special interest in finding a 

way around this restriction.\textsuperscript{14} Many psychologists of the 

1920s and 1930s believed that their field could not be scienti-


cific without quantitative measurement. Stevens\textsuperscript{15} and Lik-


er\textsuperscript{16} played central roles in defining measurement so that the 

"assignment of numbers to observations according to a 

rule" could be construed as sufficient for maintaining and 

furthering psychology's status as a science. Others, such as 

Thurstone,\textsuperscript{17} realized that scientific status depends on mathe-

matical thinking, and that numbers and the comparisons in 

which they are applied may or may not be mathematical.

Merely to count noses or the answers on a test or seconds of 

reaction time or volume of secretion does not make a study 
either mathematical or scientific. This is not unlike the confusion 

by which arithmetical labor is sometimes called mathematical.\textsuperscript{17} 

Arithmetical manipulations of numbers derived from counts of 

seconds may make functionometry more mathematical or scientific 

than the use of the seconds. Mathematically and 

scientifically rigorous functionometry is a matter of de-

termining a system of associated observable behaviors that 

remain invariantly ordered among the persons exhibiting 

those behaviors, as well as among the items and rating cate-

gories chosen to represent those behaviors. Functionometry 

requires that the functional independence construct be articu-

lated and observed, and not another variable, such as speed 

or distance.

**PRACTICAL BENEFITS OF PROBABILISTIC 

MODELS**

Rating scale data's ordinal status is less of a problem than 

accepting such ratings as measures is, because the latter 
denies research and applications the benefits of mathematical 

thinking.\textsuperscript{4,14,20} One of these benefits is the possibility of co-

calibrating instruments purported to measure a common vari-

able. Cocalibration is simply an extension of test equating,
item banking, and partial credit\textsuperscript{4} principles that have been in use in education for decades. In the same way that banked test items are selectively administered on parallel forms for security or targeting reasons, the items on functional assessment instruments can be treated as belonging to a single pool of items. And in the same way that partial credit testing allows groups of items, or even every individual item, to be calibrated according to the properties of their own particular grading or rating scale structure (true/false, multiple choice, or any number of rating scale points), the differences in functional assessment rating scales can be included in the model specifying what kind of data are expected.

To calibrate a functional assessment instrument in a way that tests the quantitative hypothesis is to ask the question, "Do the results of comparisons of subjects’ levels of functional independence depend on which (relevant and well-targeted) assessment items are administered?" If the answer to this question is yes, to assume that, or proceed as if, measurement is taking place is unethical and unscientific.\textsuperscript{23} Evidence of consistent variation of person abilities across the items administered (whether from one instrument or several), and item difficulties among persons, is fundamental to establishing construct validity, and without construct validity one can hardly purport to know what is being measured.\textsuperscript{24}

Science demands demonstrations and substantiations of the contention that comparisons are based on single, and not double or triple, standards. Comparisons based on raw scores, however, always depend on which items are administered, as any plot of a single group of subjects’ scores from two tests of easy and hard difficulties, or any plot of data from one test taken by two groups of persons of high and low abilities, will show.\textsuperscript{25} In contrast to the popular and simplistic methods offered by Stevens\textsuperscript{15} and Likert,\textsuperscript{16} scale-free measurement requires that conjoint person-item order be demonstrated, not assumed. Rasch’s probabilistic formulation of scale-free measurement easily tolerates missing data and thereby offers many advantages to the comparison of instruments’ relative strengths and weaknesses, and the establishment of a common currency for the exchange of theories about and observations of functional independence.

**METHODS**

This example of scale-free functionometry was first conceived as a possibility when it was noticed that the order of items on the FIM motor scale\textsuperscript{26} was virtually identical to the order of similar items on the PECS motor scale.\textsuperscript{26} This observation led us to obtain ratings of patients’ functional independence on both instruments’ motor skills items. In the first step of the cocalibration, the motor skills items from the two instruments were analyzed together, and in the second step, the theoretically common-unit measures resulting from independent administrations of the instruments were compared.

**Measurement Model**

**Parameter estimation.** A basic mathematical form of scale-free measurement models\textsuperscript{4} specifies the log-odds probability of a given score as:

$$\log \left[ \frac{P_{nk}}{1 - P_{nk}} \right] = B_n - D_i - F_k$$

in which

- $P_{nk}$ = probability of subject $n$ on item $i$ rating in category $k$
- $P_{nk-1}$ = probability of subject $n$ on item $i$ rating one category less than $k$
- $B_n$ = Ability measure of subject $n$
- $D_i$ = Difficulty measure of functional independence item $i$
- $F_k$ = Difficulty calibration of rating scale category $k$

A conjoint, iterative procedure uses $B$, $D$, and $F$ each in turn to produce estimates of the other parameters’ values. Probabilities are calculated according to the number of items, rating scale categories, and subjects sampled. An item’s maximum possible score depends on the number of rating categories and the number of subjects, and a subject’s maximum possible score depends on the number of categories and the number of items. Actual scores are compared with the maximum possible scores (given the number of items rated, for subjects, or the number of subjects rated, for items) to determine the probabilities associated with each item, subject, and rating category interaction for each score group. A quick approximation of ability is used to estimate initial item and rating scale difficulties, which in turn are used to estimate another set of abilities, and so on.

**Parameter convergence.** If the differences between the subsequent and prior estimates shrinks with each iteration, the process is said to converge, and a fundamental hurdle in the test of the quantitative hypothesis has been cleared, because convergence will occur only when some semblance of consistency exists in the data. Measurement error is estimated as a function of the number of items and rating scale points for subjects, and as a function of the sample size for the items.

**Data consistency and fit to the model.** Fit statistics are derived from the difference between observed and expected ratings, expectations being derived from the consistency of the data. For instance, a low rating on an easy task for an able person (someone with high ratings on other more difficult tasks) will produce a high fit statistic. Standardized fit statistics have an expected range of -2.0 to 2.0 (2SD below the mean to 2SD above the mean, which accounts for 95% of the variation). Item calibrations or subject measures with fit statistics outside this range are not sufficiently contributing to nor cooperating with the measurement effort, and require special examination to determine the cause of the inconsistency.

**Parameter separation.** The scale-free principle is built into the mathematical model in such a way that comparisons of two subjects’ abilities can be made irrespective of the particular item on the scale facilitating that comparison. For instance, two subjects’ abilities ($b_n$ and $b_m$) can be compared such that the difficulty ($d_i$) of the functional assessment item providing the basis of comparison disappears from the equation:

$$g_{est} = b_n - d_i$$

$$-g_{est} = b_m - d_i$$

$$g_{est} - g_{est} = b_n - b_m$$

Arch Phys Med Rehabil Vol 76, February 1995
Table 1: Reliability Statistics

<table>
<thead>
<tr>
<th>SD/Err</th>
<th>Alpha</th>
<th>Strata*</th>
<th>Variation (%)</th>
<th>Not Due to Error/ Due to Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>.500</td>
<td>1.67</td>
<td>50/50</td>
<td></td>
</tr>
<tr>
<td>1.53/1</td>
<td>.700</td>
<td>2.37</td>
<td>60/40</td>
<td></td>
</tr>
<tr>
<td>2/1</td>
<td>.800</td>
<td>3.00</td>
<td>67/33</td>
<td></td>
</tr>
<tr>
<td>3/1</td>
<td>.900</td>
<td>4.33</td>
<td>75/25</td>
<td></td>
</tr>
<tr>
<td>4/1</td>
<td>.941</td>
<td>5.67</td>
<td>80/20</td>
<td></td>
</tr>
<tr>
<td>5/1</td>
<td>.962</td>
<td>7.00</td>
<td>83.3/16.7</td>
<td></td>
</tr>
<tr>
<td>6/1</td>
<td>.973</td>
<td>8.33</td>
<td>86/14</td>
<td></td>
</tr>
<tr>
<td>7/1</td>
<td>.980</td>
<td>9.67</td>
<td>88/12</td>
<td></td>
</tr>
<tr>
<td>8/1</td>
<td>.985</td>
<td>11.00</td>
<td>89/11</td>
<td></td>
</tr>
<tr>
<td>9/1</td>
<td>.988</td>
<td>12.33</td>
<td>90/10</td>
<td></td>
</tr>
<tr>
<td>10/1</td>
<td>.990</td>
<td>14.00</td>
<td>91/9</td>
<td></td>
</tr>
</tbody>
</table>

* Ranges in the measurement continuum separated by at least three error terms.

The equation represents the difference (gmn) between the ability (bn) of subject n and the difficulty (dt) of item t minus the difference (gm*n) between the ability (bm*n) of subject m and the difficulty (dm*t) of item t. Subtracting the negative difficulty (dm*) of item t from itself is the same as adding positive and negative difficulties, which results in zero. The difference between the two subjects' ability measures (bm and bn) can then be calculated without any reference to the particular difficulties (dm) of the items on the instrument facilitating the comparison.

The same principle must hold with respect to comparisons of item difficulty; the difference between the item difficulties must not depend on which subject ability estimate is taken as the basis for the comparison. These principles of item and person parameter separation are the mathematical conditions at the origin of the term scale-free and are the conceptual basis of Rasch's separability theorem.

Reliability statistics. Reliability is a matter of producing statistically significant distinctions among measures. Traditional reliability statistics are derived from a true score approach to measurement, which assumes that measures are composed of an unknowable true score and a random error. Wright and Masters note that the error is not completely random, but that the data's internal consistency should locally localize error. Items that are much too easy or too difficult for a subject to perform should measure with more error and less reliability than items that are approximately as difficult as the subject is able to perform.

The reliability statistic should indicate how many statistically distinct quantitative levels a scale is capable of distinguishing. The number of distinctions that can be made among measures after error has been removed can be derived from the ratio of the standard deviation to the error, which increases linearly as variation increases in proportion to the error (Table 1). Traditional reliability statistics, in contrast, are compressed by a ceiling effect. The linear increase in the separation index G and in the strata communicate increased scale sensitivity in a simple and straightforward manner that traditional reliability statistics do not. This enhanced sense of scale reliability is useful for distinguishing among groups of measures, for instance in determining the likelihood of detecting the presence or absence of a treatment in two groups. As the sensitivity of the scale goes up, so does the power of the experimental design in which the instrument is used.

Computer software. Computer software that performs these analyses, supplies scale value, error, and fit statistics, and aids in testing the quantitative hypothesis for various kinds of data structures is documented by Wright and Linacre and Linacre. 27,28

Observation Model

Instruments. Table 2 lists the names of the PECS and FIM motor skills items. The PECS is a 93-item battery of 5 instruments that measure impairment severity, applied self-care, motor skills, cognition, and community reintegration; ratings for all items are structured by a 7-point scale ranging from the least independence (1) to independence within normal limits (7). 10,29-31 The PECS was originally intended for use as a series of single-item indicators of patient status that would serve as the focus of the patient care team conference. 9 Only in light of the possibilities for scale-free measurement introduced by Rasch measurement have the ratings been summed into patient scores.

The FIM is comprised of 18 items; it also uses a 7-point rating scale for all of its items. The FIM ratings were originally intended to be summed across all 18 items into a total score, 11 but subsequent research has shown that the items encompass two distinct constructs, a 13-item motor skills scale and a 5-item cognition scale.

Rating scales. One of the illusory obstacles to calibrating scale-free units of measurement is that rating scales are rarely structured in exactly the same way from one brand of instrument to the next. When this happens, sums of ratings mean different things depending on which instrument is used.

For instance, the PECS has three ratings (5, 6, and 7) for independent functioning, and the FIM has two (6 and 7). Thus, even if a common number of items addressing identical functions is identified, a patient would have a higher raw score on the FIM than on the PECS.

But in contrast with the summed ratings approach, Rasch's scale-free measurement models view data in terms of probabilities. The scale values of the functional assessment items depend on the probability that one rating or another is assigned. We expect the FIM items to have lower scale values than the PECS, because lower scale values are

<table>
<thead>
<tr>
<th>Table 2: Rehabs Motor Skills Item Bank</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIM Component—13 items</td>
</tr>
<tr>
<td>A—eating</td>
</tr>
<tr>
<td>B—bathing</td>
</tr>
<tr>
<td>C—dressing</td>
</tr>
<tr>
<td>D—bladder menwrap</td>
</tr>
<tr>
<td>E—UE dressing</td>
</tr>
<tr>
<td>F—toilet</td>
</tr>
<tr>
<td>G—bowl menwrap</td>
</tr>
<tr>
<td>H—C/W transfers</td>
</tr>
<tr>
<td>I—toilet</td>
</tr>
<tr>
<td>J—walking/autonomous wheelchair</td>
</tr>
<tr>
<td>M—stair</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PECS Component—22 items</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADL1—feeding</td>
</tr>
<tr>
<td>ADL4—hyg/groom</td>
</tr>
<tr>
<td>ADL7—shoes/socks</td>
</tr>
<tr>
<td>ADL11—housework</td>
</tr>
<tr>
<td>ADL15—tub/shower</td>
</tr>
<tr>
<td>PHY3—chair</td>
</tr>
<tr>
<td>PHY7—responsible</td>
</tr>
<tr>
<td>PHY10—balance</td>
</tr>
<tr>
<td>ADL2—UE dressing</td>
</tr>
<tr>
<td>ADL5—UE dressing</td>
</tr>
<tr>
<td>ADL9—Urinal</td>
</tr>
<tr>
<td>ADL13—home mobility</td>
</tr>
<tr>
<td>PHY1—transfers</td>
</tr>
<tr>
<td>PHY4—environ barriers</td>
</tr>
<tr>
<td>PHY8—position changes</td>
</tr>
<tr>
<td>ADL3—UE bathing</td>
</tr>
<tr>
<td>ADL6—UE dressing</td>
</tr>
<tr>
<td>ADL10—meal prep</td>
</tr>
<tr>
<td>ADL14—toilet transfer</td>
</tr>
<tr>
<td>PHY2—ambulation</td>
</tr>
<tr>
<td>PHY5—car transfers</td>
</tr>
<tr>
<td>PHY9—endurance</td>
</tr>
</tbody>
</table>
associated with the probability of higher ratings. Higher ratings on an item mean that the function represented is easier to perform than functions on which patients receive lower ratings; therefore, patients with low measures are succeeding primarily on easy tasks, ones with low calibrations. Patients of greater functional independence will garner high ratings on the more difficult tasks as well, giving them higher measures. Having item difficulty and patient ability expressed in the same unit makes it easy to say exactly what a subject with a particular measure can and cannot do. In the same way, cocalibrating the FIM and the PECS will show how much more difficult it is to obtain a rating of a particular value on the PECS than on the FIM.

**Raters.** FIM ratings were assigned to the functional independence of a group of patients by a special group of therapists who were trained in the use of the FIM when they worked at facilities that use it. These therapists observed and rated subjects’ FIM levels at admission and discharge. PECS ratings were assigned by the treating therapists as part of the usual hospital procedures.

**Persons measured.** Fifty-four consecutive patients admitted as inpatients to a free-standing rehabilitation hospital were rated on both the FIM and the PECS. Table 3 shows the programs to which the patients were admitted, and the number in each. Prior research involving tens of thousands of measures made with all five of the PECS subscales and with both of the FIM subscales shows that, with limited exceptions, the scale structures do not vary significantly among diagnostically distinct samples. If this were not the case, it would not be reasonable to expect this mixed-diagnosis group of 54 patients to teach us much about the possibilities that exist for cocalibrating the PECS and the FIM. Should this sample of patients cause the items on these two instruments to calibrate into an order markedly different from the order they take in larger studies, the usefulness of these results will be called into question.

**Analytic Procedure**

The first step in cocalibrating two instruments is to assess a group of patients on all of the items on each. It was necessary to apply the two instruments together, analytically treating the 35 motor-skills items (22 PECS and 13 FIM items), in effect, as one instrument. The names of the items from each instrument are shown in table 2. Once the cocalibration is performed, any future administration of either instrument can be interpreted in terms of the common unit, the rehabit, derived via this analysis. An easy to use table facilitates the conversion of raw PECS scores (the sum of the individual ratings) into raw FIM scores, and vice versa, by means of the rehabits.

Once the two instruments have been analyzed together, the item calibrations are anchored and each instrument is used to generate its own measures to determine how well the theoretical correspondence of the two instruments works in practice. This involves another analysis of the data from each of the two instruments, but this time two analyses are made, one for each instrument. The scale values of the items are preset at their cocalibration values; if the instruments are measuring in the same unit, the patient measures resulting from these two analyses should be virtually identical. The measures from the separate analyses were therefore compared with one another, as well as with the cocalibration measures.

**RESULTS**

**Reliability and Fit**

Table 4 shows that the persons measured are spread along the measurement continuum with a reliability of 0.95, meaning that the 35 FIM/PECS items have distinguished six statistically distinct levels of functional independence (strata) in these persons’ abilities. Conversely, these 54 subjects have separated the items on the measurement continuum with a reliability of 0.98, meaning that they have distinguished almost 10 strata in the items’ difficulties.

The standard deviation of the subject outfit statistic is a little high, meaning that there are some persons with unexpectedly high ratings at the difficult end of the scale, or unexpectedly low ratings at the easy end of the scale. Examination of the data shows that four of the five subjects with the highest outfits have overall high measures, but a rating of 1 on the easiest item, the FIM feeding item, or overall low measures, but a rating of 7 on difficult items. Because a 1 on the FIM can mean that the subject could not be rated on this item, these data points might better be left as missing.

The fit statistic’s standard deviation is probably thrown off primarily by one subject with a mean square outfit of 9.9; this subject has almost no PECS data, has mostly ratings of 1 and 2 on the FIM items, but has a rating of 7 on difficult items. Because a 1 on the FIM can mean that the subject could not be rated on this item, these data points might better be left as missing.

**Table 4: Rehabs Summary Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Persons</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg measure</td>
<td>45.1</td>
<td>50.0</td>
</tr>
<tr>
<td>Avg error</td>
<td>2.7</td>
<td>1.3</td>
</tr>
<tr>
<td>Adj SD</td>
<td>12.1</td>
<td>8.9</td>
</tr>
<tr>
<td>Separation</td>
<td>4.48</td>
<td>6.83</td>
</tr>
<tr>
<td>Reliability</td>
<td>.95</td>
<td>.98</td>
</tr>
<tr>
<td>Avg MNSQ infit</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>SD MNSQ infit</td>
<td>.6</td>
<td>.5</td>
</tr>
<tr>
<td>Avg MNSQ outfit</td>
<td>1.1</td>
<td>1.1</td>
</tr>
<tr>
<td>SD MNSQ outfit</td>
<td>1.1</td>
<td>.7</td>
</tr>
</tbody>
</table>

Arch Phys Med Rehabil Vol 76, February 1995
Common Item Correspondence

Nine of the FIM and PECS items correspond closely in the type of functions addressed. The plot of their calibration values shown in figure 1 replicates the observations first reported by Silverstein and coworkers\(^{10}\) concerning a common structure to the FIM and PECS motor skills subscales. Silverstein and coworker’s\(^{10}\) comparison of their PECS motor skills item calibrations with Heinemann and colleague’s separately conducted study of FIM motor skills item calibrations\(^{25}\) produced a 0.93 correlation. The calibrations produced by the present study, though derived from a small sample, parallel the PECS and FIM calibrations made on tens of thousands of patients.\(^{10,12,25,26,29-34}\) In every instance, feeding/eating is easiest, stairs/environmental barriers is hardest, and upper extremity functions are easier (have lower calibrations) than lower extremity functions. It seems that motor-skills measures, at least, can be constructed not only to transcend the patients measured, the raters rating them, and the facilities where the ratings are made, but to also eliminate the brand of the instrument as a determinant of the measuring unit.

How Much Harder is the PECS?

Though these data confirm the correspondence of similar assessment areas across these two instrument brands, figure 2 shows that they also bring out an important difference. This difference is the result of the greater ease of earning a higher rating on the FIM. For the patients measured here, the FIM is closer to the average subject measure (45.1), with the average FIM calibration of 43.9 and the average PECS calibration of 53.6. The PECS items, though, vary more completely across the range of subject measures. Excluding those two standard deviations below or above the mean, the person measures range from 20 to 68, with an error of 2.4. The FIM items range across 27.3 rehabits, from 32.6 to 59.9 (error = 1.2), and the PECS ranges across 34.2 rehabits, from 34.2 to 68.4 (error = 1.3), a difference of 6.9 rehabits, or more than 5 times the average item error. Both instruments could stand to measure even lower levels of functioning, but as access to rehabilitation services grows, and the delivery method focuses increasingly on day treatment, outpatient, and home health services, the PECS will be better positioned to measure these patients’ higher levels of functional independence.

Comparing Admission Versus Discharge Calibrations

Another way to assess instrument validity is to divide those measured into groups, calibrate on those groups, and then compare the calibrations. Figure 3 compares a calibration made on data from the initial admission assessment with a calibration made on discharge data alone. The two calibrations correlate 0.89, with an R\(^2\) of 0.79, which supports the contention that the same construct is being measured in both samples.

The PECS item involving taking responsibility for mobility and the FIM item involving ambulation are the only two items that differ to a statistically significant degree in their initial and discharge calibrations. Taking responsibility for mobility becomes more difficult at discharge by about 15 rehabits, or more than 10 errors, probably because patients do not change much between admission and discharge in their propensity to take charge of mobility. Although gains are made in every other area, responsibility ratings stay about the same, causing the item’s position in the calibration order to shift accordingly. Deleting this item boosts Pearson’s R to 0.93 and the R\(^2\) to 0.86.

The FIM walking item changes by about 12 rehabits, and seems to have been inordinately influenced by the requirement that patients not tested be rated a 1. Many patients have ratings of 1 at admission, but are discharged with ratings of 5, 6, or 7. The lack of gain associated with taking responsibility for mobility is contrasted here with an excessive amount of gain; patients are too easily moving across the entire 1 to 7 scale. Because the only reason for assigning a 1 for missing data is that raw score summary statistics require it,
perhaps the PECS method of coding functional areas not assessed as 0 will be called for if the full potential of scale-free measurement to inform us about functional independence is to be realized. Deleting the FIM walking item boosts the correlation coefficient to 0.95, with an $R^2$ of 0.90.

**Comparing FIM and PECS Measures**

With the confidence in the stability of the cocalibration values that the comparison of admission and discharge calibrations provides, it is reasonable to use these values to generate measures, using each instrument separately. Measures of functional independence produced by the anchored (preset) FIM items are on the vertical axis in figure 4, and those produced by the anchored PECS items are on the horizontal axis.

The relationship of FIM to PECS rehabs is closer at the top of the measurement continuum than at the bottom. This is largely because of missing data on low ability patients; three to six data sets were deleted from this and the two subsequent graphs because of large amounts of missing data.

The correlation (Pearson's R) of the measures produced by the separately anchored FIM and PECS items is 0.91, and the common variance ($R^2$) of the measures from the two scales is 0.83. The probability of this correlation happening by chance is less than 0.0000; conversely, the probability of obtaining statistically significant results is practically 1.0, given the n of about 100, a p value less than .01, 97 degrees of freedom, an error of 4.5, and one independent variable. Similar statistical power considerations hold in the following two comparisons.

Figures 5 and 6 show plots of the FIM and PECS measures with the original cocalibration measures. The high correlations provide further support for the assertion that the instruments measure the same variable.

Finally, Bland and Altman's suggestion that the type of plot shown in figures 4, 5, and 6 may be misleading when used to assess agreement between two methods of clinical measurement. They provide an example in which a correlation of 0.89 and apparent agreement in a plot of measures from two instruments misses clinically significant lack of agreement. A plot of the mean of each subject's two measures against the difference of each measure from their mean is shown to illuminate this lack of agreement.

Figure 7 is a plot of this type for the FIM and PECS measures. There is not significant lack of agreement in the measures made from the two instruments, because 94 of the 99 measures are not different by more than 9.6 rehabs, or about 2 errors. In fact, 88 (89%) of the measures have a discrepancy of less than 6 rehabs, or about an error and a half. A Kolmogorov-Smirnov test of the normality of the distribution of these differences indicates a statistically significant departure from normality (N = 99, mean difference of the FIM and PECS measures from their mean = -1.16, SD = 4.8, K-S Z = .87, 2-tailed p = .44). Figure 7 suggests a reason for the nonnormality because all of the points outside of the 2SD cut are associated with measures of less than 50.

This association suggests a source for the one discrepancy involving a FIM measure lower than the associated PECS measure. Low FIM measures often include ratings of 1 that are not indications of actual functional independence, but are used to provide a score in the absence of an assessment. Because the PECS does not require complete data for totaling into a summary score, therapists are free to skip irrelevant items. Because of this difference in the use of the instruments' respective rating scales, the one patient with a FIM measure more than 10 rehabs less than her or his PECS measure is likely to have several ratings of 1 on the FIM in association with missing data on the comparable PECS items.

Subject 71, at the top center of figure 7, has a PECS measure of 56 rehabs and a FIM measure of 44, with ratings of 1 on FIM J (toilet transfers), K (tub/shower transfers), and M (stairs), and no data for the associated PECS items, activities of daily living (ADL) 14, ADL 15, and physical therapy (PHY) 4. The FIM ratings on these items come in conjunction with ratings of 4 and 5 on FIM items of similar
difficulty, so it is likely the case that the ratings mean that the patient was not assessed, as opposed to actually functioning at that lowest level. If the ratings of 1 on the FIM were deleted, the proportion of the total raw score to the number of items rated would go up, and the resulting FIM measure would then be in closer proximity to the PECS measure.

The four other person measures outside of the confidence limits in figure 7 are instances in which the FIM measures are higher than the PECS measures. Subject 29, whose FIM measure is almost 20 rehabits more than his or her PECS measure, is a 71-year-old stroke patient who was in the rehabilitation program for only 6 days, and who had a large amount of missing data across all of the items on both the admission and discharge assessments. The anomalous data provided by this patient are clearly not representative of the vast majority of cases and can be safely ignored.

Two of the remaining three measures have inordinately high ratings on the FIM bowel and bladder items, for which there are no PECS counterparts in this scale (they are included in the PECS Applied Self-Care scale). The bowel and bladder items are among the three least well-fitting items on the FIM. Two of the subject measures below the confidence interval in figure 7 include high residuals on these two items. Residuals are calculated by comparing the rating expected for a subject on an item, given the subject’s total score, with the observed rating; a high residual indicates that the observed rating is higher than what the subject’s overall score and the position of the item on the measurement continuum would lead one to expect.

For instance, person 67 has an average PECS-FIM measure of about 32 and a PECS-FIM difference of −12; this subject’s ratings on the cocalibration measure order of the items is

212271711224724 23111132331 1

and subject 83, with an average measure of about 27 and a difference of about −14, has a pattern that looks like this

3243636223222221 131111 111 1 1 1

In both instances, the highest ratings come on the fifth and seventh items from the left, which are the FIM bowel and bladder items, respectively. In addition, subject 67 had a pattern of ratings including another unexpected 7 (on the FIM F hygiene/grooming item) and does not exhibit the expected left to right decline in ratings. Accordingly, this person’s measure is associated with a high fit statistic, which brings the validity of the measure into question. Limiting our concern to the problem at hand, were these especially high ratings on these two items for these two patients omitted from the analysis, their FIM measures would drop into closer accord with their PECS measures.

**DISCUSSION**

**PECS/FIM Rehabs Table**

Table 5 is a sketch of the relationship shared by the PECS and FIM raw scores, as it is mediated by the rehabits. Converting from PECS to FIM and vice versa is as simple as reading a row in the table. A raw PECS score of 23, obtained via ratings of 1 on 21 items and a rating of 2 on one item, corresponds to a rehabit of 0 and a raw FIM score of 14; and a FIM score of 28 measures the same amount of functional independence as a PECS score of 39; and a FIM score of 64 corresponds to a PECS score of 82.
At the bottom of the scale, we see the effect of the PECSSs larger number of items. Having 22 items, the PECSS has a maximum score of 154 (7 × 22), in contrast to the FIM maximum of 91 (7 × 13). This means that the PECSS measures more precisely (with a lower error) and in a wider range than the FIM does.

Besides being easier and more convenient to use than raw scores, rehabs make it apparent that a single amount of functional independence is represented by a variety of raw score differences, depending on which range in the scale is involved. For instance, the 10 rehabit range from 0.0 to 10.0 is associated with a change of one rating point on the FIM and 2 rating points on the PECSS, but the same 10 rehabit span in the middle of the scale, say from 40 to 50, encompasses 19 raw FIM ratings and 25 raw PECSS ratings.

Finally, the column on the far left of table 5, labeled strata, indicates which of four statistically distinct ranges in the measurement continuum each score group occupies. The center of each of these strata is at least three errors away from the center of an adjoining stratum. A span of three errors is an arbitrary and conservative, but high-powered, criterion for determining significant differences; more or less than three errors may be justified in different applications.

This article documents a study showing a unique benefit of scale-free measurement techniques, the cocalibration of instruments that have different items and rating scale structures, but which are intended to measure the same variable. Cocalibration is possible when item response data from both (or all) of the instruments in question vary consistently and meaningfully across persons, and vice versa. Cocalibration opens the door to establishing standards for instrument quality that could greatly improve research in the human sciences by making experimental results more easily comparable and simplifying the general comparison of amounts of ability and attitude.

For instance, whether or not a new instrument measures the same variable as existing instruments would not be determined by means of correlational studies, but through a cocalibration that shows whether the new instrument

- Measures the same variable as the old by means of a test for data consistency across the instruments;
- Improves the quality of measurement by having a lower error or reduction of misfit when compared with the old instrument; or
- Increases the range of the measurement continuum by extending the articulation of the variable to higher highs and lower lows, thereby

Enhancing the capacity of the instrument to measure persons with items targeted at their particular ability levels.

Rehabs, the name of the measuring unit proposed for the scale-free measurement of functional independence in physical medicine and rehabilitation, will open the door to a science of functionometry. Via this science, rehabilitation professionals will discover what the smallest measurable, observable, and meaningful differences in function are, and will have a baseline of standardized criteria for evaluating treatment program effectiveness and performance. Further research will determine to what extent various mobility and ADL scales measure the same variable, and which of them do so with the least error, and with the best fit to the demands of the quantitative hypothesis.

Acknowledgment: The authors would like to thank Wayne A. Gordon, Benjamin D. Wright, Mark V. Johnston, and the anonymous reviewer(s) for their careful readings of the manuscript. The authors would also like to acknowledge the support Burton Silverstein provided this project in its early phases.

References
5. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil 1989;70:857-60.
16. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;140:5-55.


**Suppliers**