JOURNAL OF OUTCOME MEASUREMENT, *1*(2), 87-113

Physical Disability Construct Convergence Across Instruments: Towards a Universal Metric

William P. Fisher, Jr.
Louisiana State University Medical Center

Objectives. This study examines the stability of a physical disability construct across instruments and samples. The purpose is not to report a formal equating of instrument calibrations, but to indicate whether such an effort would be likely to succeed. Theory. The economics transforming health care from its orientation toward crisis-driven disease reactions to population- and evidence-based preventive health management and individualized disease management demand general scale-free measures of functional independence. Methods. A new method, pseudo-common item equating, is demonstrated. Similar, but not identical items, from different instruments, calibrated on different samples, are compared. Data. More than 30 articles presenting Rasch analyses of physical functioning scales were reviewed. Four instruments provided data from ten of these articles, for eleven different calibrations (two instruments are both included in one article). Results. The final overall average correlation disattenuated for error is .93, with an average of 7 pseudo-common items, and an average p-value of .01, meaning that measures based on these calibrations should be linearly transformable versions of the same metric. Scientific importance. The quantitative stability of different areas of physical functional independence across instruments and samples suggests that the development and deployment of a universal metric is a realizable goal.

Requests for reprints should be sent to William P. Fisher, Jr., Louisiana State University Medical Center, 1600 Canal St., New Orleans, LA 70112.

OBJECTIVES

Researchers are increasingly appreciating the advantages of conducting what L. L. Thurstone (Thurstone 1959, p. 228) called the "crucial experimental test" of the hypothesis that an instrument's measures are mathematically sound, i.e., not affected by the abilities or attitudes of the particular persons measured, nor by the difficulties of the particular survey or test items used to measure. Georg Rasch independently derived a similar, more efficient, and less cumbersome form of the same experimental test, employing Ronald Fisher's (Fisher, 1922) "formalization of sufficiency [to] nail down the conditions that a model must fulfill in order for it to yield an objective basis for inference" (Rasch, personal communication recorded in Wright, 1980, p. xii).

Although application of Rasch's models is growing steadily, little or nothing has yet been done to try taking full advantage of the objective bases for inference established in this work. What remains to be done is to explore the extent to which instruments passing the "crucial experimental test" can support metrological systems in which the concept of scale-free measurement is pushed to its limit. Metrology to date has focused on the calibration and maintenance of physical weights and measures, but if scale-free measurement is what it appears to be, there should be no reason why universal metrics cannot be established for the measurement of constructs accessed via tests and rating scales.

This study examines the possibility that a single construct presents itself as a stable phenomenon across instruments intended to measure physical disability, and across samples of persons experiencing physical disability. No formal equating of the instruments is attempted; the goal is only to indicate whether such an effort would be likely to succeed. Common-sample equatings (Wright & Stone, 1979, p. 107-112; Masters, 1985) of instruments intended to measure physical disability, or motor skills (Fisher, Harvey, Taylor, Kilgore & Kelly, 1995; Heinemann, Segal, Schall & Wright 1996), and this comparison of independently conducted calibration studies, suggest that the construct is stable across several instruments and diagnostic groups. Whether the construct retains its definition across health care providers, geographic regions, and other sources of possible variation remains to be seen.

Typical methods of equating employ either a common sample of persons (to link different instruments) or common items across tests or surveys (to link different samples of persons measured). Another method

that suggests itself is a pseudo-common item equating, in which the calibrations of similar, but not identical items, from different instruments are compared. Common-sample equating shows that differences between rating scales may cause similar items from different instruments to scale in a common order, but not at the same points on the measurement continuum (Fisher, Harvey, Taylor, Kilgore & Kelly, 1995). Thus the results of this study merely suggest directions that future research might take, and do not delineate a common metric for the measures of physical disability studied.

THEORY

One of the reasons why little has yet been done to connect success in scale-free measurement with metrological possibilities is because of insufficient respect for the extent to which the history of measurement is tightly intertwined with the history of economics. The history of British weights and measures, for instance, shows that metrological systems proliferate during periods of economic depression dominated by local trade, whereas widely accepted metrological standards emerge in periods of economic expansion characterized by regional and international trade (Zupko, 1977).

The economic transformation of health care taking place today thus has many historical precedents. We are moving from local economies of disease-crisis management to regional, national, and international economies of population-based, preventive health management. As health care shifts from the disease management of individuals to the health management of populations, the need to compare health status (physical and psychosocial functioning, quality of life and life satisfaction, risk behaviors, and satisfaction with services) across groups, regionally, nationally, and internationally, increases.

The problem is that the transformation of the system cannot take place as long as health status is measured in scale-dependent units that vary in their size and meaning by instrument brand, provider, plan, and patient, and in the presence of missing data (skipped items). The most commonly proffered solution to this problem is to have all users employ the same scale and always supply complete data, which at least offers the appearance that the same thing is measured. A number of data-sharing schemes based on common scale use have emerged (HEDIS, NCQA, UDS, etc.), but these are flawed in a number of ways, including their need for complete data; their incapacity to adapt to the needs of individuals; their insufficient data quality assessment and improvement procedures;

their treatment of ordinal data as interval; their lack of individual error estimates for each measure; and their mutual incommensurability.

Scientific instruments are often described as embodying experiments and theory in a practical portable device (Ackermann, 1985; Heelan, 1983; Ihde, 1991; Schaffer, 1992). Rasch measurement theory works well in this context as a framework for defining the relevant experimental and theoretical structures. In order for physical disability to be counted on as a measurable construct, 1) experimental data must mediate the relationship between a theory of the variable and the instrument embodying it; 2) theory must mediate the relationship between the instrument and the data; and 3) the instrument must mediate the theory-data relationship (Ackermann, 1985; Ihde, 1991). These mediations succeed to the extent that each of them is mathematically invariant enough to support the separation of the parameters associated with each facet of the design (Bachelard, 1984; Kline, 1985).

In other words, the instrument must provide a theoretically transparent window on the data, and the data on the instrument. We must be able to look through the instrument and see the component ratings comprising the measure (Feinstein, 1987), otherwise the summary scores are not sufficient statistics, the parameters will not separate, the instrument is seriously affected in its measuring function by the object of measurement, the crucial experimental test has not been passed, and the hypothesis that the variable is quantitative (Michell, 1990) has been falsified.

In the present case, the successful eleven tests of the hypothesis that physical disability is a quantitative variable provides extensive information on the relationships among theory, data, and instruments in this field. The eleven experiments suggest a theory of physical disability in which gross motor skills are affected by at least two factors: 1) the extent to which upper and lower extremities must work together in coordinated activity, and 2) the amount of strength available in these extremities. In 1991, a cursory glance at the Rasch scalings of two physical disability scales, from the PECS (Silverstein, Kilgore, & Fisher, 1989; Silverstein, Fisher, Kilgore, Harvey, & Harley, 1992) and the FIM (Heinemann, Hamilton, Granger, Wright, Linacre, et al., 1991), suggested the rudiments of the theory tested in the present study, namely, that:

- feeding and grooming, strictly upper extremity functions, are the easiest tasks (with the lowest calibrations) usually found on disability instruments;

- upper extremity dressing and bathing are easier than lower extremity dressing and bathing, but more difficult than feeding and grooming;
- transfers to and from chairs or seats of various kinds, involving some coordination of, and strength in, both upper and lower extremities, are of medium difficulty;
- walking, requiring more extensive coordination of the upper body and lower extremities, as well as more strength, is also of medium difficulty, but of greater difficulty than most transfer activities;
- stair climbing, demanding more coordination and strength than of the activities usually included on physical disability assessments, is of greatest difficulty, with the highest calibrations; and
- motor skills that have something of an involuntary muscle control component to them, such as bowel and bladder management, will vary in the amounts of difficulty they pose, and so will not fit the measurement model as well as the more consistently structured skills.

From this data-informed theory of task difficulty, we can more clearly state the expectations concerning the variable's structure that are implemented when data are fit to a Rasch model. Following Thurstone, Fisher, and Rasch, an objective basis for inference requires this clear statement of theoretical expectations, so that unmodeled variation can be detected and so that exactly what is being counted on in the quantification process to produce a mathematically invariant unit of measurement is made as explicit as possible. As is persuasively argued by Andersen (Andersen, 1977), summated scores must necessarily be assumed sufficient statistics for them to be meaningful. Since sufficiency is rarely demonstrated (Michell 1990; Wright 1984; Wright, 1985; Fisher, 1994), objective bases for inference from summated ratings are equally rare.

For a score based on summed ratings to be a sufficient statistic, it must be transparent (Feinstein, 1987), in the sense of allowing one to reproduce from it, and from it alone, the most likely response to every question on the instrument used to measure, whether or not every question was in fact administered. When a construct is stable, in the sense of producing a particular order to the areas addressed by items on scales, an order that remains invariant no matter which scale or which sample is involved, such reproducibility, to use Guttman's (Guttman 1950) term, is obtained. Fit to a Rasch model demonstrates that, given a stable construct, an instrument

embodying it in an invariant item order, and a logit value (the natural logarithm of the rating odds derived from a sufficient statistic), it is possible to reproduce the pattern of most likely responses across the length of the instrument, within an error of measurement.

Guttman's (Guttman, 1950) rigid, deterministic sense of reproducibility is impractical (Wilson, 1989). Rasch's probabilistic approach brings an objective basis for inference within reach by requiring 1) that items obtaining higher proportions of their maximum possible scores always have a higher probability of being succeeded on by any person than items with lower proportions, no matter which scale the items come from; and 2) that persons achieving higher proportions of their maximum possible scores always have a higher probability of success on any item than persons with lower proportions, non matter which scale(s) the persons were measured with.

These requirements make it possible to test explicitly what is usually merely assumed about data, namely, that the abilities of the persons measured and the difficulties of the tasks posed at each rating scale step are the only factors affecting the outcome of the rating process. The basic Rasch rating scale model,

$$\log\left(\frac{P_{nik}}{P_{ni(k-1)}}\right) = B_n - D_i - F_k,$$

is nothing but a mathematical statement of this expectation concerning the relation of person ability (B_n) to item difficulty (D_i) at each rating scale step (F_k), and so constitutes a formalization of what is necessarily assumed whenever summated scores are treated as measures (Andersen, 1977).

After all, the reason why raters are trained to record observations according to a standardized protocol is so that higher scores on the instrument can be inferred to represent greater amounts of the variable. But can summary scores simply be assumed to function as sufficient statistics? What if a particular score is produced by high ratings on tasks that are typically difficult, and low ratings on tasks that are typically easy? The diagnostic utility of such variation evaluated in the context of Rasch measurement has already been documented (Daltroy, Logigian, Iversen, Liang, 1992; Granger & Wright, 1993). Currently, inordinate variation in the meaning of scores is almost always completely unnoticed, but it would

seem to have large consequences for the understanding of the natural history of degenerative diseases, of treatment progress, the quality of care, and the evaluation of treatment effects. A far superior strategy is not to assume that scores are sufficient statistics, but to require that sufficiency be supported by the results of the relevant tests so that congruence with, or divergence from, the relevant theoretical expectations can be observed and evaluated before the "foundations of misinference" (Merbitz, Morris, & Grip, 1989) are laid.

METHODS

Scale-free measurement models (Rasch, 1960; Andrich, 1988; Wright & Masters, 1982) provide the relevant tests. As it becomes increasingly clear that the accountability of educators, psychologists, health care providers, and other professionals cannot remain tied to scale-dependent indicators of unknown or low statistical sufficiency, the practicality, scientific rigor, and mathematical beauty of scale-free measurement will become more widely appreciated. Furthermore, as the computational capacity to implement accountability measures, and the telecommunications capacity to implement wide-scale comparisons, mount hand-in-hand, the intellectual and methodological skills required to use scale-free measurement techniques will become automated. Taken together with the economic demand for accountability and comparisons, these factors will greatly heighten demand for scale-free measurement.

To take full advantage of scale-free measurement's possibilities, all scales designed to measure a single construct, such as physical disability, will need to be calibrated together on a common sample of persons (which is not to say that some missing data could not be tolerated). Since disability measurement scales rarely, if ever, share identical items administered in exactly the same way, common item equating is not a viable option as a means of co-calibration. But in contrast with common sample and common item equating, pseudo-common item equating requires neither shared items across scales nor ratings of a single group of persons' physical disabilities on all the items from all the scales to be studied.

Instead, items that actually differ in the details of their content and administration, but which address conceptually similar areas of functional assessment, are treated as though they are the same for the purposes of exploring the likelihood that a large scale equating study would meet with success. This provisional methodology falls short of providing equated

values for the different scales' items, but has the advantage of being able to make use of calibrations that are readily available in published sources.

The item orders delineated by the scales were examined for mutual congruence by evaluating correlation coefficients and scatter plots of the pseudo-common item values. Scatter plots of the items from the pairs of scales with the lowest correlations were examined in order to find and remove mismatched item pairs, in preparation for calculating another set of correlations. Like common item equating, this pseudo-common item equating method does not require that all respondents be rated on all items, since subsets of items can be used to establish a common metric. Once the common metric is established, however, the items omitted in the equating study would be replaced in each of the respective scales for normal use. The last step of this study was to remove the effect of calibration error from the correlations.

DATA

More than 30 articles presenting Rasch analyses of physical disability scales were reviewed. Four instrument provided data from ten of these articles, for eleven different calibrations (two instruments are both included in one article). The primary criterion for including an instrument was that it address a range of difficulty relevant to assessing the treatment needs of persons with disabilities. The instruments included are the Functional Independence Measure (FIM) (Chang & Chan, 1995; Fisher, Harvey, Taylor, Kilgore, & Kelly, 1995; Grimby, Andrén, Holmgren, Wright, Linacre, et al., 1996; Heinemann, Hamilton, Granger, Wright, Linacre, et al., 1991; Heinemann, Linacre, Wright, Hamilton & Granger, 1993; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994; Mayes, Perez, Mipro, & Fisher, 1996; Pollack, Rheault, & Stoecker, 1996; Qayum, Ortenberg, Morstead, Siddiqui, Mipro, et al., 1996); the Katz ADL Index (Katz) (Teresi, Cross & Golden, 1989); the Levels of Rehabilitation Scale - III (LORS) (Velozo, Magalhaes, Pan & Leiter, 1995); and the Patient Evaluation Conference System (PECS) (Kilgore, Fisher, Silverstein, Harley & Harvey, 1993; Fisher, Harvey, Taylor, Kilgore, & Kelly, 1995).

Though the Assessment of Motor and Process Skills (AMPS) (Fisher, A., 1993; Fisher, A., 1994) and the Tufts Assessment of Motor Performance (TAMP) (Fisher, A., 1992; Haley & Ludlow, 1992; Ludlow & Haley, 1992; Ludlow, Haley & Gans, 1992) meet the criterion of relevance to assessing the treatment needs of persons with disabilities, they were not included

because of their more tightly focused and detailed assessment approaches. Where the FIM, PECS, LORS, and Katz focus on gross motor skills such as dressing, walking, and bathing, the TAMP and the AMPS focus on fine motor skills, such as lifting, grasping, reaching, balancing, etc. Although common sample equating of instruments intended to measure gross and fine motor skills may be possible and highly desirable, it is difficult to conceive of a way in which the pseudo-common item methodology could be applied.

Pediatric assessment instruments also were not included, more for reasons of relevance to a common population than as a result of methodological considerations.

General measures of health status that have been calibrated via fit to Rasch models, such as the Health Status Questionnaire 2.0 (HSQ) (Radosevich, Wetzler & Wilson, 1994; Fisher, Marier, & Hunter, 1995); the Louisiana State University Health Status Instruments (LSU HSI) (Fisher, Marier, & Hunter, 1995; Fisher, Eubanks, Marier, & Hunter, 1996; Fisher, Marier, Eubanks, & Hunter, 1997); and the Medical Outcomes Study Short-Form 36 (SF-36) 10-item Physical Functioning scale (PF-10) (Fisher, Marier, & Hunter, 1995; Fisher, Eubanks, Marier, & Hunter, 1996; Haley, McHorney & Ware, 1994; Heinemann, Segal, Schall & Wright, 1996; Stucki, Daltroy, Katz, Johannesson, & Liang, 1996) do not meet the inclusion criterion since they are not primarily intended for assessing the functionality of persons with disabilities. Instead, the greater difficulty of their items make them better suited for assessing the physical functioning of persons who may be sick or reaching advanced ages, but who are not in need of rehabilitation services. Common-sample equating of the LSU HSI and the SF-36 PF-10 is nearing completion (Fisher, Eubanks, Marier, & Hunter, 1996). Separate common-sample equatings of the FIM and several cancer quality of life instruments with the SF-36 are underway (Heinemann, Segal, Schall, & Wright, 1996; Cella, Lloyd & Wright, 1996).

A common-sample equating of the physical disability scales from the FIM and the PECS (Fisher, Harvey, Taylor, Kilgore, & Kelly, 1995) shows that differences between the two instruments' rating scales results in a 1-logit (10 rehabit) difference between the scale values for similar items. Other differences of this kind among the instruments studied here cannot be illuminated without further common sample equatings.

Item calibrations obtained for each instrument from the articles reviewed are shown in Table 1. The lowest number of assessment areas shared by two instruments is three, and the most is 13, with the majority sharing 6-8 items. Missing data are not always an indication that the instrument lacks an item

concerning the relevant assessment area as some articles do not provide information on all the items' scale values. Seven of the eleven studies involve the FIM, so the FIM item letters have been included in Table 1 to aid in item identification. The FIM Dressing item is sometimes applied separately to upper and lower extremity functions; when this is the case, the two are averaged for comparison with other studies employing only one Dressing item. The averaged item is removed from comparisons of calibration sets that include the distinction between upper and lower extremities.

The values shown in Table 1 relate the calibrations for items on each instrument to the common core of physical functioning areas assessed. For instance, the LORS Mobility items are associated with the other instruments' Walking items.

As far as can be determined from the articles reviewed, the sample sizes reported indicate the total number of measures. Because most studies incorporate at least two applications of the instrument, this means that the actual number of persons measured is typically half of the sample size reported. The only studies that do not employ two applications of the instrument are the Chang and Chan study of the FIM (FIMCHA), which includes three administrations (admission, discharge, and follow up), and the Grimby, et al., study (FIMGRI), which includes one administration.

Several of the studies do not report a single value for each item. Linacre, et al. (FIMLIN) report only separate admission and discharge calibrations for the FIM; Velozo, et al. (LORS) report separate calibrations for the same LORS-III activities based on ratings from nurses and therapists; Chang and Chan (FIMCHA) reported only separate FIM admission, discharge, and follow up values; and Teresi, et al., (KATZ) reported two Katz calibrations, one for a New York City sample and one for a London sample. In each of these cases, the calibrations have been averaged across raters, times or sites to simplify the analysis.

The Teresi, et al. (KATZ) data are further complicated by four other factors: 1) the items point in the opposite direction from the items on the other scales; 2) the article states that improvements were made to the scale, but the effect of these changes on the scale values is not indicated; 3) an unstated number of original response options were converted to dichotomous responses; and 4) this is the only study to employ a 3-parameter Item Response Theory model for item difficulty estimation. Because of its rating scheme, and unlike the other instruments, low measures on the Katz indicate less disability, and high measures, more, so the resulting negative signs on the correlation coefficients are reversed when these coefficients are averaged.

Table 1
Pseudo-Common Item Calibration Values in Theoretical Order

Item Name	FIM Letter	PECS KILG[1]	FIM LRI[2]	PECS WF[3]	FIM W PECS[4]	FIM LIN[5]	FIM CHA[6]	FIM RST[7]	LOR[8]	KATZ[9]	FIM GRI[10]	FIM POL[11]
Hard												
Stairs	M	1.03	2.24	1.00	1.00	1.58	.95	1.52	.	-.62	.77	2.86
Walking	L	.39	.83	.20	.80	.43	-.44	.39	.86	-.28	-1.30	-.10
Bathtrn	K	.32	.52	.	.	.88	.55	.81	.	.	-.40	1.35
Medium												
Bathing	C	.	.45	.	.	.24	1.01	.23	.89	-1.44	-.15	.12
Toiletin	F	.	.51	.	.	.08	.50	.09	.43	.08	-.90	.
Toiltran	J	.	.29	-.50	.50	.15	.17	.11	.	.	-.80	-.28
BatheLE	C	.	.	-.70	.30	.	.	.	.	.	.	.
DressLE	D	.	.	-.60	.30	.24	.70	.14	.	.	.58	.05
Dressing	D	.02	.01	-1.00	-.20	-.14	.50	-.11	.63	-.11	-.09	-.11
Wheelch	L	-.17	-.25	.	.	.	.	-.07	.	.47	.	.
Gentrans	I	-.39	-.30	-.40	.70	.01	.05	-.07	.	.36	-1.00	-.33
Easy												
DressUE	D	.	.	-1.50	-.80	-.50	.31	-.35	.	.	-.75	-.27
BatheUE	C	.	.	-1.60	-.80	.	.	.	.	.	.	.
Grooming	B	-.32	-1.39	.	.	-.78	-.54	-.61	-.81	.	-1.00	-1.45
Feeding	A	-1.77	-1.65	-1.80	-1.60	-1.25	-1.92	-.96	-1.98	.42	-1.77	-1.34
Unclassified												
Bladder	G	.	-.43	.	.	-.49	-.65	-.42	.	.32	.	-.05
Bowel	H	.	-.28	.	.	-.59	-.69	-.51	.	.65	-.56	.
Sample Size		3,700	250	100	100	29,600	300	1,900	6,000	300	53	98
Max # items		8	13	9	9	13	13	13	6	10	12	11
Stad. Dev.		.81	.98	.90	.89	.74	.83	.64	1.16	.63	.73	1.20
Error		.02	.15	.15	.15	.02	.05	.02	.02	.30	.20	.10
Reliability		1.00	.98	.97	.97	1.00	1.00	1.00	1.00	.82	.93	.99

Table 1 Continued
Pseudo-Common Item Calibration Values in Theoretical Order

Table 1 footnotes:

1. Kilgore, K. M., Fisher, W. P., Jr., Silverstein, B., Harley, J. P., & Harvey, R. F. (1993). Application of Rasch analysis to the Patient Evaluation and Conference System. Physical Medicine and Rehabilitation Clinics of North America: New developments in functional assessment, 4(3), 493-515.
2. Qayum, M., Ortenberg, K., Morstead, R., Siddiqui, F., Mipro, R. C., Jr., & Fisher, W. P., Jr. (1996, June). Measuring functional status in rehabilitation: Louisiana Rehabilitation Institute vs. Uniform Data System [Abstract]. LSU Department of Medicine, Section of Physical Medicine & Rehabilitation, Residents' Research Day. New Orleans, LA: LSU School of Medicine.
3. Mayes, P., Perez, A., Mipro, R. C., Jr., & Fisher, W. P., Jr. (1996, June). Comparison of Functional Independence Measure data from the Uniform Data System and the Louisiana Rehabilitation Institute [Abstract]. LSU Department of Medicine, Section of Physical Medicine & Rehabilitation, Residents' Research Day. New Orleans, LA: LSU School of Medicine.
4. Fisher, W. P., Jr., Harvey, R. F., Taylor, P., Kilgore, K. M., & Kelly, C. K. (1995). Rehabits: A common language of functional assessment. *Archives of Physical Medicine and Rehabilitation, 76*, 113-122.
5. Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 75*(2), 127-132.
6. Chang, W., & Chan, C. (1995). Rasch analysis for outcomes measures: Some methodological considerations. *Archives of Physical Medicine and Rehabilitation, 76*(10), 934-939.
7. Heinemann, A. W., Hamilton, B. B., Granger, C. V., Wright, B. D., Linacre, J. M., Betts, H. B., Aguda, B., & Marnott, B. D. (1991). Rating scale analysis of functional assessment measures [NIDRR Innovation Grant Award Final Report]. Chicago, Illinois: Rehabilitation Institute of Chicago.
8. Velozo, C. A., Magalhaes, L. C., Pan, A., & Leiter, P. (1995). Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. *Archives of Physical Medicine and Rehabilitation, 76*(8), 705-712.
9. Teresi, J. A., Cross, P. S., & Golden, R. R. (1989). Some applications of latent trait analysis to the measurement of ADL. *Journal of Gerontology: Social Sciences, 44*(5), S196-204.
10. Grimby, G., Andrén, E., Holmgren, E., Wright, B., Linacre, J. M., & Sundh, V. (1996, November). Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: A study of individuals with spina bifida. *Archives of Physical Medicine and Rehabilitation, 77*(11), 1109-1114.
11. Pollak, N., Rheault, W., & Stoecker, J. L. (1996, October). Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. *Archives of Physical Medicine and Rehabilitation, 77*(10), 1056-1061.

The Grimby, et al., (FIMGRI) study is the only one in which the FIM's seven-point rating scale is modified; perhaps because of the particular problems experienced by persons suffering with cerebral palsy and spina bifida, the middle three categories were collapsed into a single one. Grimby, et al., omit the Bladder and Bowel items; Pollack, et al., (FIMPOL) omit the Toileting item.

Where Walking and Wheelchair use are rated separately, both calibrations are used; where only one calibration is reported, it is included in the Walking item.

One FIM study (FIMRST) was taken from an unpublished source (Heinemann, Hamilton, Granger, Wright, Linacre, et al., 1991) because the relevant published source (Heinemann, Linacre, Wright, Hamilton, & Granger 1993) reports separate calibrations for each of several different samples of diagnostic-related groups of patients, and these only in the form of scatter plots. The unpublished source was consulted because it provides lists of each diagnostic group's calibration values, but the same data are presented in the publication. The FIM calibrations based on assessments of patients suffering right hemisphere strokes were selected from the available diagnostic groups because of the sample size and as contrast with the Chang and Chan study (FIMCHA), which includes only right stroke patients. Both the 1991 Heinemann, et al. (Heinemann, Hamilton, Granger, Wright, Linacre, et al., 1991) and the 1993 Heinemann, et al. (Heinemann, Linacre, Wright, Hamilton, & Granger 1993) studies concluded that there were no important variations in the instrument structure across diagnostic groups, a conclusion supported by a similar study of the PECS (Kilgore, Fisher, Harvey, & Silverstein, 1993).

The FIMLRI data are from a retrospective study of about 100 medical records (Mayes, Perez, Mipro, & Fisher, 1996; Qayum, Ortenberg, Morstead, Siddiqui, Mipro, et al., 1996). The assessments included are ratings made at admission, during treatment, and at discharge; the patients involved are a diagnostically mixed group of persons presenting themselves for treatment at a state public hospital affiliated with a state university medical center.

Finally, two studies (FIMLIN and FIMRST) are of samples drawn from the same database. Because the sample size of the earlier study (FIMRST) is less than 6.5 percent of the later study (FIMLIN), including both in these comparisons helps make the point concerning the stability of the construct across wide differences in sample size.

Table 2
Pseudo-Common Item Scale Correlation Coefficients

	PECS KILG	FIM LRI	PECS WF	FIMW PECS	FIM LIN	FIM CHA	FIM RST	LORS	KATZ	FIM GRI
FIM	.89									
LRI	(8)									
	P= .00									
PECS	.92	.96								
WF	(5)	(6)								
	P= .01	P= .00								
FIMW	.89	.83	.91							
PECS	(5)	(6)	(9)							
	P= .02	P= .02	P= .00							
FIM	.91	.95	.96	.88						
LIN	(7)	(12)	(7)	(7)						
	P= .00	P= .00	P= .00	P= .00						
FIM	.88	.76	.60	.69	.77					
CHA	(7)	(12)	(7)	(7)	(13)					
	P= .00	P= .00	P= .08	P= .04	P= .00					
FIM	.89	.94	.95	.82	.99	.75				
RST	(8)	(13)	(7)	(7)	(13)	(13)				
	P= .00	P= .00	P= .00	P=.01	P= .00	P= .00				
LORS	.95	.94	.85	.94	.97	.87	.96			
	(4)	(6)	(3)	(3)	(6)	(6)	(6)			
	P= .03	P= .00	P= .18	P=.11	P= .00	P= .01	P= .00			
KATZ	-.80	-.60	-.82	-.64	-.61	-.70	-.61	-.63		
	(6)	(10)	(5)	(5)	(9)	(9)	(10)	(5)		
	P= .03	P= .03	P= .04	P= .12	P= .04	P= .02	P= .03	P= .13		
FIM	.79	.73	.57	.51	.69	.77	.68	.70	-.54	
GRI	(7)	(11)	(7)	(7)	(12)	(12)	(12)	(6)	(8)	
	P= .04	P= .01	P= .09	P= .12	P= .01	P= .00	P= .02	P= .13	P= .17	
FIM	.80	.90	.83	.63	.92	.67	.94	.92	-.46	.75
POL	(7)	(10)	(7)	(7)	(11)	(11)	(11)	(5)	(7)	(10)
	P= .02	P= .00	P= .01	P= .07	P= .00	P= .01	P= .00	P= .03	P= .30	P= .01

(Coefficient / (Cases) / 1-tailed Significance)

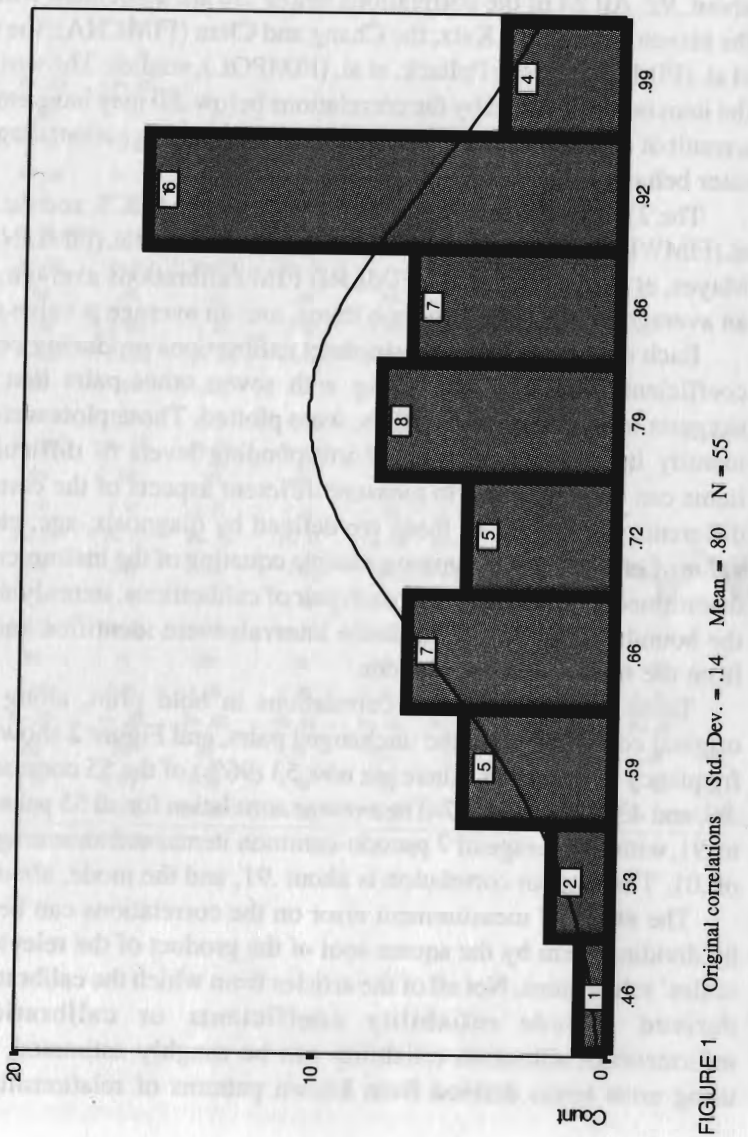

FIGURE 1 Original correlations Std. Dev. = .14 Mean = .80 N = 55

RESULTS

Table 2 shows the correlations, number of items, and p-values for the pseudo-common item comparisons across instruments. Figure 1 shows the correlations' frequency distribution. Thirty-one of the 55 correlations (56%) are over .80, and 23 (42%) are .87 or higher; the range extends from .50 to .99, with a mean of about .79, a median of about .82 and a mode of about .92. All 24 of the correlations below .80 are associated with four of the eleven studies, the Katz, the Chang and Chan (FIMCHA), the Grimby, et al. (FIMGRI), or the Pollack, et al. (FIMPOL), studies. The variations in the item order signified by the correlations below .80 may have emerged as a result of several unidentifiable influences, including patient diagnoses or rater behaviors.

The 21 correlations among the LORS, the two PECS, and the Fisher, et al. (FIMWPECS), Heinemann, et al. (FIMRST), Linacre, et al. (FIMLIN), and the Mayes, et al./Qayum, et al. (FIMLRI) FIM calibrations average .92, with an average of 7 pseudo-common items, and an average p-value of .02.

Each of the 24 pairs of instrument calibrations producing correlation coefficients less than .80, along with seven other pairs that Table 1 suggests have significant outliers, were plotted. These plots were used to identify item pairs not posing corresponding levels of difficulty. Such items can be considered to measure different aspects of the construct for different populations, as these are defined by diagnosis, age, etc., so the value of undertaking a common-sample equating of the instruments can be determined without them. For each pair of calibrations, items lying outside the bounds of the 95% confidence intervals were identified and omitted from the subsequent correlations.

Table 3 shows the new correlations in bold print, along with the original correlations for the unchanged pairs, and Figure 2 shows the new frequency distribution. There are now 53 (96%) of the 55 correlations over .80, and 43 (78%) over .87. The average correlation for all 55 pairs increases to .91, with an average of 7 pseudo-common items, and an average p-value of .01. The median correlation is about .91, and the mode, about .95.

The effect of measurement error on the correlations can be removed by dividing them by the square root of the product of the relevant pair of scales' reliabilities. Not all of the articles from which the calibrations were derived include reliability coefficients or calibration error information.Calibration reliability can be roughly estimated, however, using error terms derived from known patterns of relationships, where

Table 3
Pseudo-Common Item Scale Correlation Coefficients
Outliers Identified and Omitted for Bolded Correlations

	PECS KILG	FIM LRI	PECS WF	FIMW PECS	FIM LIN	FIM CHA	FIM RST	LORS	KATZ	FIM GRI
FIM	.89									
LRI	(8)									
	P= .00									
PECS	.92	.96								
WF	(5)	(6)								
	P= .01	P= .00								
FIMW	.89	.83	.91							
PECS	(5)	(6)	(9)							
	P+ .02	P= .02	P= .00							
FIM	.91	.95	.96	.88						
LIN	(7)	(12)	(7)	(7)						
	P= .00	P= .00	P= .00	P= .00						
FIM	**.96**	**.85**	**.71**	**.80**	**.84**					
CHA	**(6)**	**(10)**	**(6)**	**(6)**	**(11)**					
	P= .00	**P= .00**	**P= .06**	**P= .03**	**P= .00**					
FIM	.89	.94	.95	.82	.99	**.82**				
RST	(8)	(13)	(7)	(7)	(13)	**(11)**				
	P= .00	P= .00	P= .00	P= .01	P= .00	**P= .00**				
LORS	.95	.94	.85	.94	.97	**.99**	.96			
	(4)	(6)	(3)	(3)	(6)	**(5)**	(6)			
	P= .03	P= .00	P= .18	P= .11	P= .00	**P= .01**	P= .00			
KATZ	**-.91**	**-.83**	**-.96**	**-.96**	**-.86**	**-.88**	**-.85**	**-.93**		
	(5)	**(9)**	**(4)**	**(4)**	**(8)**	**(6)**	**(9)**	**(4)**		
	P= .02	**P= .03**	**P= .02**	**P= .02**	**P= .00**	**P= .01**	**P= .02**	**P= .03**		
FIM	**.94**	**.91**	**.89**	**.98**	**.90**	**.88**	**.91**	**.92**	**-.96**	
GRI	**(6)**	**(8)**	**(5)**	**(4)**	**(9)**	**(8)**	**(9)**	**(5)**	**(5)**	
	P= .02	**P= .00**	**P= .02**	**P= .01**	**P= .00**	**P= .00**	**P= .00**	**P= .01**	**P= .01**	
FIM	**.92**	**.90**	**.91**	**.79**	**.92**	**.99**	**.94**	**.92**	**-.88**	**.90**
POL	**(5)**	**(11)**	**(5)**	**(6)**	**(12)**	**(8)**	**(11)**	**(5)**	**(6)**	**(8)**
	P= .01	**P= .00**	**P= .02**	**P= .03**	**P= .00**	**P= .00**	**P= .00**	**P= .03**	**P= .01**	**P= .00**

(Coefficient / (Cases) / 1-tailed Significance)

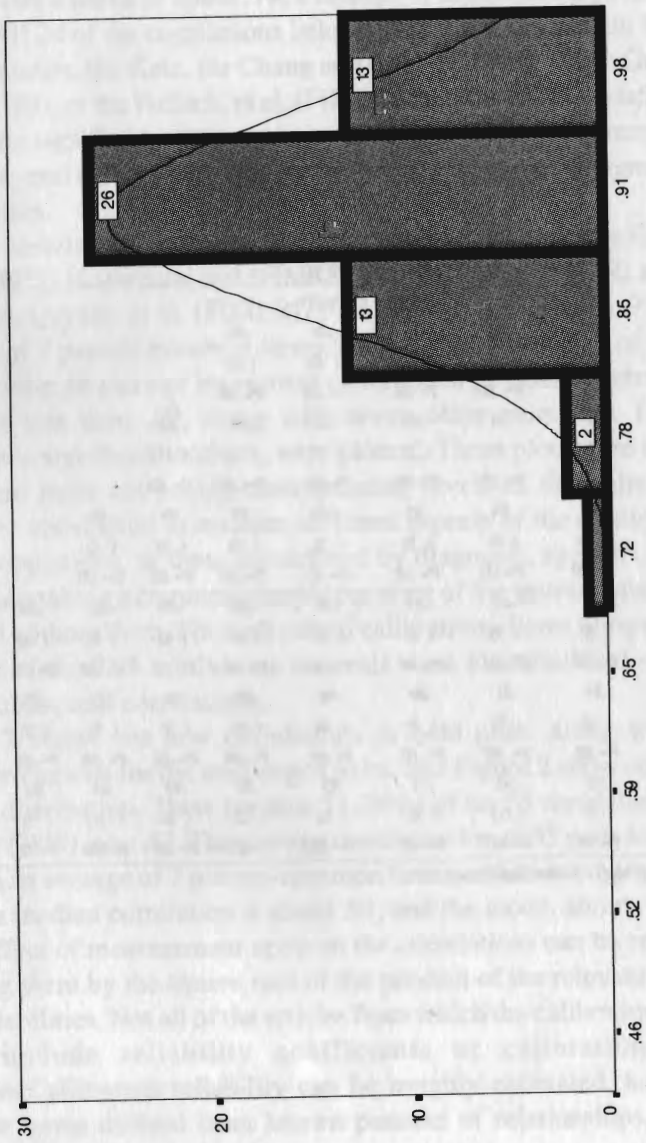

FIGURE 2 Correlations with outliers removed Std. Dev. = .06 Mean = .91 N = 55

error decreases as sample size and the number of rating scale options increase (Wright & Stone, 1979). To be conservative in disattenuating the error from the correlations, it is best to under estimate error and over estimate reliability, or the correlations may be inflated by removing more error from them than actually exists. The correlation of two scales with reliabilities of 1.0 will be unaffected by the disattenuation as there is no error to remove. This calculation can result in numbers higher than 1.0, if the square root of any multiplied pair of reliabilities is greater than its associated correlation. Any results higher than 1.0 were changed to 1.0.

Sample sizes range from 53 to almost 30,000, and numbers of rating options range from the Teresi study's dichotomization of the Katz data, to the LORS's five-point scale, to the Grimby study's five-point FIM, and to the similar seven-point scales used with the FIM and the PECS.

Under-estimated error for the Katz is probably about .30 due to the sample size of 300 and dichotomous data. The Grimby study does not report complete errors for its modified FIM, but with a sample size of 53, error is probably slightly under-estimated at .20. All of the other studies either report error or have sample sizes so large as to make calibration reliability virtually 1.0.

The reliabilities shown in Table 1 were calculated using these errors and the standard deviations of the calibrations shown, using the formula $(SD/Error)^2 / (1+(SD/Error)^2)$ described by Wright and Masters (Wright & Masters, 1982, p. 92). The Dressing item is omitted when calibrations for the UE and LE Dressing items are available.

Talbe 4 and Figure 3 show the effect of error disattenuation on the correlations. The mean disattenuated correlation is .93, with a median of .94, and two modes, with frequencies of 7, at .95 and at 1.00. All but one of the 55 correlations (98%) are now over .80, and 46 of them (84%) are over .87.

These results are evidence in support of the stability of the physical functioning construct across the instruments and samples. Measures based on these calibrations should be linearly transformable versions of the same metric. The trouble and expense of research employing these instruments in common-sample equating efforts are likely to be well-rewarded.

SCIENTIFIC IMPORTANCE

The study results do not falsify the hypothesis that a common unit of measurement could be created for assessments of physical functioning made with different instruments, but on samples drawn from the same or similar

Table 4
Pseudo-Common Item Scale Correlation Coefficients
Effect of Error Removed

	PECS KILG	FIM LRI	PECS WF	FIMW PECS	FIM LIN	FIM CHA	FIM RST	LORS	KATZ	FIM GRI
FIM	.90									
LRI	(8)									
PECS	.93	.98								
WF	(5)	(6)								
FIMW	.90	.85	.94							
PECS	(5)	(6)	(9)							
FIM	.91	.96	.98	.89						
LIN	(7)	(12)	(7)	(7)						
FIM	.96	.86	.72	.81	.84					
CHA	(6)	(10)	(6)	(6)	(11)					
FIM	.89	.95	.97	.83	.99	.82				
RST	(8)	(13)	(7)	(7)	(13)	(11)				
LORS	.95	.95	.86	.95	.97	.99	.96			
	(4)	(6)	(3)	(3)	(6)	(5)	(6)			
KATZ	-1.00	-.93	-1.00	-1.00	-.95	-.97	-.94	-1.00		
	(5)	(9)	(4)	(4)	(8)	(6)	(9)	(4)		
FIM	.98	.95	.93	1.00	.93	.91	.95	.96	-1.00	
GRI	(6)	(8)	(5)	(4)	(9)	(8)	(9)	(5)	(5)	
FIM	.92	.92	.92	.81	.92	1.00	.94	.93	-.97	.94
POL	(5)	(10)	(5)	(6)	(11)	(8)	(11)	(5)	(6)	(8)

(Coefficient / (Cases))

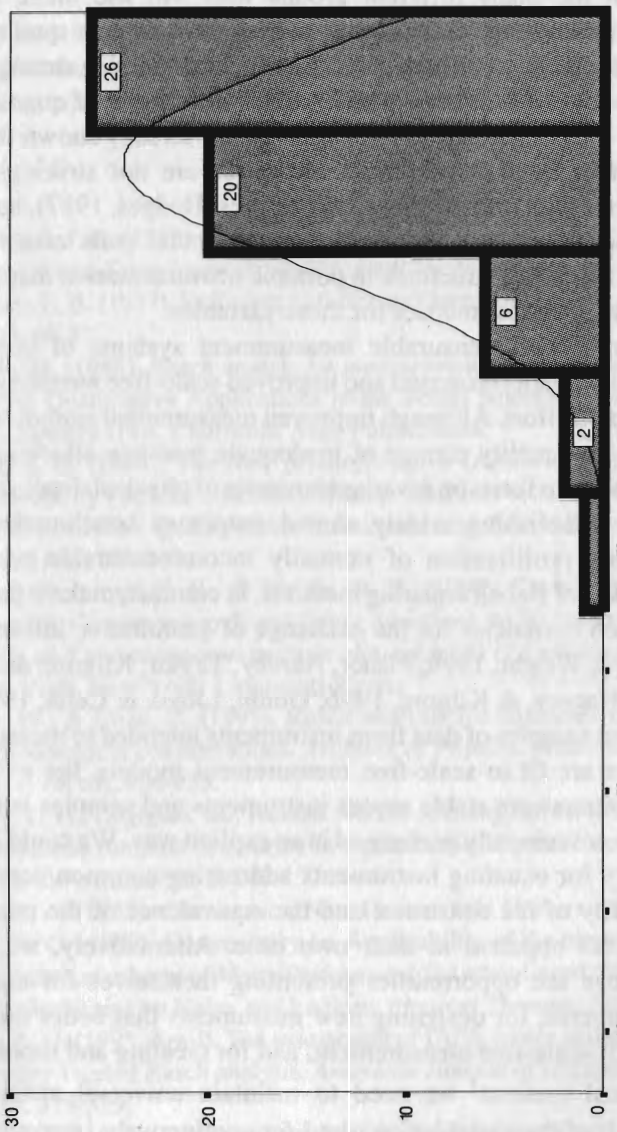

FIGURE 3 Disattenuated correlations Std. Dev. = .06 Mean = .93 N = 55

populations. The theoretical structure embodied in instruments measuring physical disability will be fleshed out as more attention is paid to new calibrations made on new samples, and as new instruments are designed and calibrated with the specific aim of improving the measurement of physical disability.

Universal implementation of scale-free metrics will require the cooperation of the many different groups who will use them. Social networks for monitoring, maintaining, and improving data quality will need to circulate standard instruments and standard samples among those interested in having a common currency for the exchange of quantitative values related to physical disability. Research has already shown that the quality and stability of psychosocial measures are not strikingly less consistent than results from the physical sciences (Hedges, 1987), but little has been done to embody the crucial experimental tests establishing variables' mathematical structures in portable instrumentation that can be used to support universal metrics for these variables.

Supplanting incommensurable measurement systems of unknown data quality with quality-assessed and improved scale-free metric systems will take time and effort. Although improved measurement cannot reverse the effects of low quality care or of inadequate resource allocation, the enhanced capacity to focus on invariant amounts of physical disability will be crucial to establishing widely shared outcomes benchmarks. The currently raging proliferation of mutually incommensurable scales is building a Tower of Babel; equating methods, in contrast, make it possible to coin common currencies for the exchange of quantitative information (Cella, Lloyd, & Wright, 1996; Fisher, Harvey, Taylor, Kilgore, & Kelly, 1995; Fisher, Harvey, & Kilgore, 1996; Gonin, Lloyd, & Cella, 1996).

As different samples of data from instruments intended to measure the same variables are fit to scale-free measurement models, the extent to which the constructs are stable across instruments and samples is tested. These tests are not currently performed in an explicit way. We could ignore the opportunity for equating instruments addressing common constructs until the stability of the constructs and the equivalence of the measures make themselves apparent in their own time. Alternatively, we could proactively seize the opportunities presenting themselves for equating existing instruments, for designing new instruments that better meet the requirements of scale-free measurement, and for creating and monitoring the metrological systems[1] we need to maintain universal metrics for measuring each of the variables required for continuously improving the

quality of care. It appears that whether or not we deliberately work to create common metrics for measuring physical disability and other variables accessed via rating scales, these metrics may eventually emerge of their own accord. The human and economic values associated with universal metrics demand that the deliberate and proactive approach win out.

ACKNOWLEDGMENTS

The author thanks Robert L. Marier, MD, MHA, and James H. Diaz, MD, DrPH, for their support of this project. Thanks also to the two anonymous reviewers who helped make this a better article.

REFERENCES

Ackermann, J. R. (1985). *Data, instruments, and theory: A dialectical approach to understanding science.* Princeton, New Jersey: Princeton University Press.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*(1), 69-81.

Andrich, D. (1988). *Rasch models for measurement.* Sage University Paper Series on Quantitative Applications in the Social Sciences, vol. series no. 07-068. Beverly Hills, California: Sage Publications.

Bachelard, G. (1984). *The New Scientific Spirit* (Arthur Goldhammer, Trans.; Foreword by Patrick A. Heelan). Boston: Beacon Press.

Bailar, B. A. (1985). Quality issues in measurement. *International Statistical Review, 53*(2), 123-139.

Cella, D. F., Lloyd, S. R., & Wright, B. D. (1996). Cross-cultural instrument equating: Current research and future directions. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials (2d edition)* (pp. 707-715). New York, New York: Lippincott-Raven.

Chang, W., & Chan, C. (1995). Rasch analysis for outcomes measures: Some methodological considerations. *Archives of Physical Medicine and Rehabilitation, 76*(10), 934-939.

Daltroy, L. H., Logigian, M., Iversen, M. D., & Liang, M. H. (1992). Does musculoskeletal function deteriorate in a predictable sequence in the elderly? *Arthritis Care Research, 5*, 146-150.

Feinstein, A. (1987). *Clinimetrics.* New Haven: Yale University Press.

Fisher, A. G. (1992). Commentary on Applicability of the hierarchical scales of the Tufts Assessment of Motor Performance for school-aged children and adults with disabilities by Haley and Ludlow. *Physical Therapy, 72*(3), 202-204.

Fisher, A. G. (1993, April). The assessment of IADL motor skills: An application of many-faceted Rasch analysis. *American Journal of Occupational Therapy, 47*(4), 319-329.

Fisher, A. G. (1994). Development of a functional assessment that adjusts ability

measures for task simplicity and rater leniency. In M. Wilson (Ed.), *Objective measurement: Theory into practice. Vol II* (pp. 145-175). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, A, 222,* 309-368.

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice. Vol. II* (pp. 36-72). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, W. P., Jr., Eubanks, R. L., Marier, R. L., & Hunter, S. M. (1996). Equating the LSU HSI and the SF-36 using a probabilistic measurement model [Measurement Research Reports]. LSU Medical Center Department of Preventive Medicine and Public Health.

Fisher, W. P., Jr., Harvey, R. F., Taylor, P., Kilgore, K. M., & Kelly, C. K. (1995). Rehabits: A common language of functional assessment. *Archives of Physical Medicine and Rehabilitation, 76,* 113-122.

Fisher, W. P., Jr., Marier, R. L., & Hunter, S. M. (1995). Comparing probabilistic calibrations of two health status instruments: The LSU HSI and the HSQ 2.0 [Measurement Research Reports]. LSU Medical Center Department of Preventive Medicine and Public Health.

Fisher, W. P., Jr., Marier, R. L., Eubanks, R., & Hunter, S. M. (1997). The LSU Health Status Instruments (HSI). In J. McGee, N. Goldfield, J. Morton & K. Riley (Eds.), *Collecting Information from Patients: A Resource Manual of Tested Questionnaires and Practical Advice (Supplement)* (p. forthcoming). Gaithersburg, Maryland: Aspen Publications, Inc.

Gonin, R., Lloyd, S. R., & Cella, D. F. (1996). Establishing equivalence between scaled measures of quality of life. *Quality of Life Research,* pp. 20-26.

Granger, C. V., & Wright, B. D. (1993, August). Looking ahead to the use of functional assessment in ambulatory physiatric and primary care. *Physical Medicine and Rehabilitation Clinics of North America: New Developments in Functional Assessment, 4*(3), 595-605.

Grimby, G., Andrén, E., Holmgren, E., Wright, B., Linacre, J. M., & Sundh, V. (1996, November). Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: A study of individuals with spina bifida. *Archives of Physical Medicine and Rehabilitation, 77*(11), 1109-1114.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer & et al. (Eds.), *Studies in social psychology in World War II. volume 4: Measurement and prediction* (pp. 60-90). New York: Wiley.

Haley, S. M., & Ludlow, L. H. (1992). Author response to Fisher's commentary on Applicability of the hierarchical scales of the Tufts Assessment of Motor Performance for school-aged children and adults with disabilities. *Physical Therapy, 72*(3), 204-206.

Haley, S. M., McHorney, C. A., & Ware, J. E., Jr. (1994). Evaluation of the MOS

SF-36 physical functioning scale (PF-10): I. unidimensionality and reproducibility of the Rasch item scale. *Journal of Clinical Epidemiology, 47*(6), 671-684.

Hedges, L. V. (1987, May). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *Journal of the American Psychological Association, 42*(5), 443-455.

Heelan, P. (1983, June). Natural science as a hermeneutic of instrumentation. *Philosophy of Science, 50,* 181-204.

Heinemann, A. W., Hamilton, B. B., Granger, C. V., Wright, B. D., Linacre, J. M., Betts, H. B., Aguda, B., & Mamott, B. D. (1991). *Rating scale analysis of functional assessment measures* [NIDRR Innovation Grant Award Final Report]. Chicago, Illinois: Rehabilitation Institute of Chicago.

Heinemann, A. W., Linacre, J. M., Wright, B. D., Hamilton, B. B., & Granger, C. V. (1993). Relationships between impairment and physical disability as measured by the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 74*(6), 566-573.

Heinemann, A., Segal, M., Schall, R. R., & Wright, B. D. (in press). Extending the range of the Functional Independence Measure with SF-36 items. In R. M. Smith (Ed.), *Outcome Measurement: Physical Medicine and Rehabilitation STARs, 11*(2).

Hunter, J. S. (1980, November). The national system of scientific measurement. *Science, 210*(21), 869-874.

Ihde, D. (1991). *Instrumental realism: The interface between philosophy of science and philosophy of technology.* The Indiana Series in the Philosophy of Technology. Bloomington, Indiana: Indiana University Press.

Kilgore, K. M., Fisher, W. P., Jr., Harvey, R. F., & Silverstein, B. (1993). Diagnosis-based differences in Rasch calibrations of functional assessment scales (abstract). *Archives of Physical Medicine and Rehabilitation, 74,* 1254.

Kilgore, K. M., Fisher, W. P., Jr., Silverstein, B., Harley, J. P., & Harvey, R. F. (1993). Application of Rasch analysis to the Patient Evaluation and Conference System. *Physical Medicine and Rehabilitation Clinics of North America: New developments in functional assessment, 4*(3), 493-515.

Kline, M. (1985). *Mathematics and the Search for Knowledge.* Oxford: Oxford University Press.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine and Rehabilitation, 75*(2), 127-132.

Ludlow, L. H., & Haley, S. M. (1992). Polytomous Rasch models for behavioral assessment: The Tufts Assessment of Motor Performance. In M. Wilson (Ed.), *Objective measurement: Theory into practice, Volume 1* (pp. 121-137). Norwood, New Jersey: Ablex Publishing Corporation.

Ludlow, L. H., Haley, S. M., & Gans, B. M. (1992). A hierarchical model of functional performance in rehabilitation medicine: The Tufts Assessment of Motor Performance. *Evaluation & the Health Professions, 15,* 59-74.

Motor Performance. *Evaluation & the Health Professions, 15,* 59-74.
Masters, G. N. (1985, March). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82.
Mayes, P., Perez, A., Mipro, R. C., Jr., , & Fisher, W. P., Jr. (1996, June). Comparison of Functional Independence Measure data from the Uniform Data System and the Louisiana Rehabilitation Institute [Abstract]. LSU Deparment of Medicine, Section of Physical Medicine & Rehabilitation, Residents' Research Day. New Orleans, LA: LSU School of Medicine.
Merbitz, C., Morris, J., & Grip, J. (1989). Ordinal scales and the foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70,* 308-312.
Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Pollak, N., Rheault, W., & Stoecker, J. L. (1996, October). Reliability and validity of the FIM for persons aged 80 years and above from a multilevel continuing care retirement community. *Archives of Physical Medicine and Rehabilitation, 77*(10), 1056-1061.
Qayum, M., Ortenberg, K., Morstead, R., Siddiqui, F., Mipro, R. C., Jr., & Fisher, W. P., Jr. (1996, June). Measuring functional status in rehabilitation: Louisiana Rehabilitation Institute vs. Uniform Data System [Abstract]. LSU Department of Medicine, Section of Physical Medicine & Rehabilitation, Residents' Research Day. New Orleans, LA: LSU School of Medicine.
Radosevich, D. M., Wetzler, H., & Wilson, S. (1994). *Health Status Questionnaire (HSQ) 2.0: Scoring comparisons and reference data.* Bloomington, MN: Health Outcomes Institute.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.
Schaffer, S. (1992). Late Victorian metrology and its instrumentation: A manufactory of Ohms. In R. Bud & S. E. Cozzens (Eds.), *Invisible connections: Instruments, institutions, and science* (pp. 23-56). Bellingham, WA: SPIE Optical Engineering Press.
Silverstein, B. J., Fisher, W. P., Jr., Kilgore, K. M., Harvey, R. F., & Harley, J. P. (1992). Applying psychometric criteria to functional assessment in medical rehabilitation: II. defining interval measures. *Archives of Physical Medicine and Rehabilitation, 73*(6), 507-518.
Silverstein, B. J., Kilgore, K. M., & Fisher, W. P., Jr. (1989). *Implementing patient tracking systems and using functional assessment scales.* Center for Rehabilitation Outcome Analysis monograph series on issues and methods in rehabilitation outcome analysis, vol. 1. Wheaton, Illinois: Marianjoy Rehabilitation Center.
Stucki, G., Daltroy, L., Katz, N., Johannesson, M., & Liang, M. H. (1996). Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. *Journal of Clinical Epidemiology, 49*(7), 711-717.

Teresi, J. A., Cross, P. S., & Golden, R. R. (1989). Some applications of latent trait analysis to the measurement of ADL. *Journal of Gerontology: Social Sciences, 44*(5), S196-204.
Thurstone, L. L. (1959). Attitudes can be measured. In L. L. Thurstone, *The Measurement of Values.* Midway Reprint Series. Chicago: University of Chicago Press, 1959, pp. 215-233. (Article originally appeared in *American Journal of Sociology, 1928, XXXIII,* 529-544.)
Velozo, C. A., Magalhaes, L. C., Pan, A., & Leiter, P. (1995). Functional scale discrimination at admission and discharge: Rasch analysis of the Level of Rehabilitation Scale-III. *Archives of Physical Medicine and Rehabilitation, 76*(8), 705-712.
Wilson, M. (1989). A comparison of deterministic and probabilistic approaches to learning structures. *Australian Journal of Education, 33*(2), 127-140.
Wright, B. D. (1980). Foreword, Afterword. In *Probabilistic models for some intelligence and attainment tests,* by Georg Rasch [Reprint; original work published in 1960 by the Danish Institute for Educational Research]. Chicago: University of Chicago Press.
Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 281-288.
Wright, B. D. (1985). Additivity in psychological measurement. In E. Roskam (Ed.), *Measurement and personality assessment.* North Holland: Elsevier Science Ltd.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.
Zupko, R. E. (1977). *British weights and measures: A history from antiquity to the seventeenth century.* Madison, WI: University of Wisconsin Press.

[1] The creation and maintenance of standard measures and of data quality standards via metrological systems are important parts of the missions of standards organizations, such as the American Society for Testing and Materials (ASTM), the American National Standards Institute (ANSI), and the International Standards Organization (ISO). A standard procedure for implementing scale-free measurement principles in metrological systems based on rating scale instruments is currently being drafted under the auspices of the ASTM E-31 Committee on the Electronic Health Record. When completed, the standard will be submitted to ANSI for approval and submission to the ISO, as all ASTM standards are. Interested parties are invited to contribute their expertise to this project. Contact Teresa Cendrowska, staff manager for Committee E-31 at ASTM (610 832-9718) for further information.