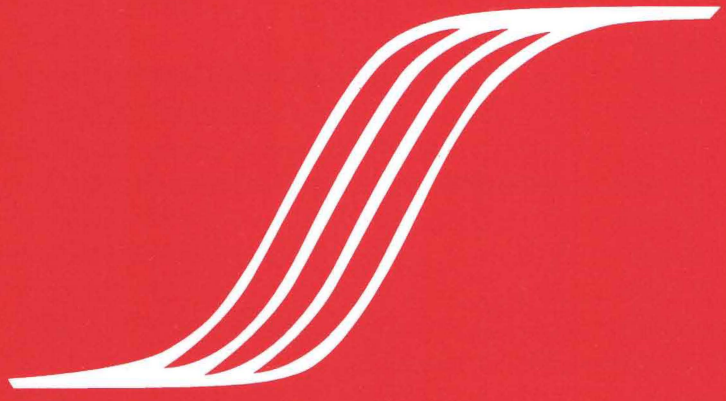
Volume 1, Number 4, 1997 ISSN 1090-655X

Journal of Outcome Measurement®

Dedicated to Health, Education, and Social Science



REHABILITATION FOUNDATION INC.
Est. 1993
Research & Education

JOURNAL OF OUTCOME MEASUREMENT®

Volume 1, Number 4 1997

Indexing/Abstracting Services: JOM is currently indexed in *Current Index to Journals in Education* (ERIC).

REVIEWER ACKNOWLEDGEMENT

The Editor would like to thank the following people who provided manuscript reviews for the Journal of Outcome Measurement during 1996 and 1997.

David Andrich, *Murdoch University*
Betty Bergstrom, *Comission on Dietetic Registration, ADA*
Trevor Bond, *James Cook University*
Ayres D'Costa, *Ohio State University*
Barbara Dodd, *University of Texas at Austin*
George Engelhard, Jr., *Emory University*
William Fisher, Jr., *Louisiana State University Medical Center*
Sarah Gehlert, *University of Chicago*
Carl V. Granger, *SUNY Buffalo*
Thomas Haladyna, *Arizona State University West*
Allen Heinemann, *Rehabilitation Institute of Chicago*
Robert Hess, *Arizona State University West*
Mark Johnston, *Kessler Institute for Rehabilitation*
William Koch, *University of Texas at Austin*
Gene A. Kramer, *American Dental Association*
Anna Kubiak, *Educational Testing Service*
Joanne Lenke, *The Psychological Corporation*
J. Michael Linacre, *University of Chicago*
Geofferey Masters, *Australian Council on Educational Research (ACER)*
David McArthur, *UCLA School of Public Health*
Carol Myford, *Educational Testing Service*
Nambury Raju, *Illinois Institute of Technology*
Tomas E. Rudy, *University of Pittsburgh Medical Center*
Randall Schumacker, *University of North Texas*
Mary Segal, *Moss Rehabilitation Research Institute*
Alan Tennant, *University of Leeds*
Luigi Tesio, *Fondazione Maugeri, Pavia, Italy*
Craig Velozo, *University of Illinois at Chicago*
Benjamin D. Wright, *University of Chicago*
David Zurakowski, *Children's Hospital, Boston*

JOURNAL OF OUTCOME MEASUREMENT®, *1*(4), 259 - 285

Evaluating the FONE FIM: Part II. Concurrent Validity & Influencing Factors

Wei-Ching Chang
University of Alberta

Chetwyn Chan
Hong Kong Polytechnic University

Susan E. Slaughter and Deborah Cartwright
Northern Alberta Regional Geriatric Program, Glenrose Rehabilitation Hospital

The "motor" (activities of daily living) component of the FONE FIM, the telephone version of the Functional Independence Measure (FIM) was evaluated in a cohort of 132 patients who had been discharged to home from a geriatric inpatient assessment and rehabilitation program. In the current study, Rasch person ability measures were derived from telephone assessments 5 weeks after discharge and in-home assessments 1 week later. Concordance between the modes was shown to be satisfactory for the Rasch measures based on intraclass correlation coefficients. However, the telephone mode consistently generated lower estimates than did the observational mode. This was due to the fact that the telephone mode underestimated motor function for the majority of patients who were at higher levels of cognition and motor function, but overestimated for patients who were at lower levels of cognition and motor function. At the item level, concordance, as determined by Kappa statistics, was better when the FONE FIM responses came from the patient rather than proxy respondents, and when the assessments were done by more experienced rather than less experienced raters. Based on these findings, a mixed strategy, the telephone mode for patients capable of responding to the FONE FIM and in-home assessments for those who are incapable, is recommended.

Requests for reprints should be sent to Wei-Ching Chang, Department of Public Health Sciences, Faculty of Medicine and Oral Health Sciences, 13-103 Clinical Sciences Building, University of Alberta, Edmonton, Alberta, Canada T6G 2G3.

An effective treatment program that desires to maintain quality and produce long-lasting results requires monitoring of its outcomes through postdischarge follow-ups. Options for gathering follow-up data include interviews in person (by observation), by telephone, or by mail. Information may be obtained either by the patient or from a proxy respondent who is knowledgeable about the patient's condition. Each method has its strengths and weaknesses in terms of response rate, turnaround time, completeness of data, bias, burden, error of interpretation, and cost (Guyatt, Feeny, & Patrick, 1993; Smith, 1992; Weinberger, Oddone, Samsa, & Landsman, 1996).

In Part I of this study (Chang, Slaughter, Cartwright, & Chan, 1997), the construct validity of the "motor" (or activities of daily living) component of the FONE FIM, the telephone version of the Functional independence Measure (FIM), was examined in relation to the observational mode (OBS FIM) administered in patients' homes. It was shown, from separate Rasch analyses of the 2 modes that the characteristics of the construct were strikingly similar in terms of their item hierarchical structures and misfitting items (Table 1). Their optimal item sets and optimal scale levels were also comparable, although the telephone mode suffered from a lower level of reproducibility than the observational mode. It was also noted that these characteristics were similar to the admission and discharge FIM data extracted from 14,799 records in the Uniform Data System for Medical Rehabilitation (Linacre, Heinemann, Wright, Granger, & Hamilton, 1994). Based on this corroborating evidence, it is proposed that the in-home, observational mode be used in this study as an external criterion for assessing the concurrent validity of the FONE FIM.

The key practical question is whether or not a person's functional ability can be assessed as accurately by the telephone as by the observational mode. This question needs to be addressed for several reasons. First, there may be discrepancies between the telephone and the observational mode of administering the FIM. The telephone mode has been shown to underestimate the functional level in such instruments as the Barthel Index (Shinar, Gross, Bronstein, Eden, Cabrera, Fishman, Roth, Barwick, & Kunitz, 1987), the Mini-Mental State Examination (Roccaforte, Burke, Bayer, & Wengel, 1992), the Functional Status Index (Jette, 1987) and the SF-36 (Weinberger et al., 1996). However, other studies showed overestimation with respect to the Barthel Index (Korner-Bitensky, Wood-Dauphinee, Siemiatycki, Shapiro, & Becker, 1994; Korner-Bitensky & Wood-Dauphinee, 1995) and large non-systematic differences with respect to the SF-36

Table 1
Rasch Analysis of 138 Motor Items: FONE FIM vs. OBS FIM

	FONE FIM			OBS FIM		
FIM Items	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)	Item Logits(Error)	INFIT MnSq(Std)	OUTFIT MnSq(Std)
MOTOR ITEMS (n=132)						
Eating	-.90(.13)	1.6(2)	1.4(1)	-1.03(.14)	1.0(0)	1.0(0)
Grooming	-.80(.13)	1.5(1)	1.1(0)	-.81(.13)	1.0(0)	.7(-1)
Toileting	-.53(.11)	1.1(0)	.8(-1)	-.39(.11)	1.0(0)	.7(-.1)
Dressing, Upper	-.47(.11)	.9(0)	.9(0)	-.33(.11)	1.1(0)	.9(0)
Bed Transfer	-.29(.10)	.5(-3)	.5(-3)#	-.52(.11)	.6(-2)	.5(-3)#
Dressing Lower	-.23(.10)	1.0(0)	.8(-1)	-.08(.10)	1.3(1)	.9(0)
Bowl Management	-.18(.10)	.9(0)	1.0(0)	-.28(.11)	1.0(0)	1.2(0)
Bladder Management	-.15(.10)	2.2(5)	1.9(3)*	-.28(.11)	2.4(5)	2.0(4)*
Toilet Transfer	-.10(.10)	.4(-4)	.5(-3)#	-.12(.10)	.5(-3)	.7(-1)
Walking	.28(.09)	.9(0)	.8(-1)	.29(.09)	1.0(0)	1.0(0)
Bathing	.88(.07)	1.4(2)	1.3(1)	.94(.08)	1.1(0)	1.1(0)
Tub Transfer	.94(.07)	.9(-1)	1.1(0)	1.00(.08)	.9(0)	1.1(0)
Climbing Stairs	1.56(.07)	1.6(3)	1.7(3)*	1.60(.07)	1.4(2)	1.7(3)

Model fit Statistics:	Root Mean-Square Std. Error	Adjusted Std. Dev.	Separation	Reliability	# of Strata
Item Statistics:					
FONE FIM	.10	.69	6.82	.98	9.4
OBS FIM	.10	.72	6.98	.98	9.6
Person Statistics:					
FONE FIM	.40	1.03	2.55	.87	3.7
OBS FIM	.42	1.06	2.52	.86	3.7

* Misfitting items; # Muted items

(Weinberger, Nagle, Hanlon, Samsa, Schmader, Landsman, Uttech, Cowper, Cohen, & Feussner, 1994). Second, the validity of the FONE FIM may depend on whether the information is provided by the patient or proxy respondent: the proxy may exaggerate the patient's disability due to perceived burden in caring for the patient (Magaziner, Simonsick, Kashner, & Hebel, 1988) or due to a common tendency to give more weight to negative than positive information when forming impressions of others (Epstein, Hall, Tognetti, Son, & Conant, 1989; Weinberger, Samsa, Schmader, Greenberg, Carr, & Wildman, 1992). On the other hand, under-reporting by proxies is known to be the major source of disagreement between the patient and proxy reports of patients' health problems (Clarridge & Passagli, 1989). Discrepancies between telephone and observational methods may result from varying degrees of reliance on proxy responses, which may constitute up to 30% of the sample (Korner-Bitensky et al., 1994, 1995). Third, telephone responses may be based more on actual habit (including the assistance of another person) than the patient's ability to perform a given function (Shinar et al., 1987). Fourth, the environment, physical disability (including diminished hearing) and cognitive status may have been responsible for the differences between the two modes (Magaziner et al., 1988; Edwards, 1990; Korner-Bitensky, et al., 1994, 1995; Rothman, Hedrick, Bulcroft, Kickam, & Rubenstein, 1991; Rubenstein, Schairer, Wieland, & Kane, 1984; Sager, Dunham, Schwantes, Mecum, Halverson, & Harlowe, 1992). Fifth, the rater does not have a chance to meet with the patients and develop a better impression when administering the FONE FIM, thus encouraging more conservative ratings (underestimation) of function. Sixth, the discrepancies may be attributable to the training and experience of the raters, since inter- and intra-rater reliability tended to positively correlate with the raters' training and experience with respect to the FIM (Fricke, Unsworth, & Worrell, 1993) and the Glasgow Coma Scale (Rowley & Fielding, 1991). These issues need to be studied in relation to the FONE FIM.

As in Part I of this paper, this study focused exclusively on the "motor" component of the FONE FIM. The Cognitive component was excluded due to its ceiling effect, which was particularly bothersome in relation to follow-up assessments. Whereas the construct validity of the FONE FIM at the item level was examined in Part I, the concurrent validity at the subject level was the main focus of Part II. The hypotheses tested for the motor component of the FONE FIM in this paper are the following:

1. There is a high degree of concordance between the two modes.
2. The FONE FIM is associated with lower estimates of patients' motor functioning than the OBS FIM.

3. Both the concordance and the difference between the 2 modes depend on the following factors:
   3.1. respondent status
   3.2. the rater's experience
   3.3 the patient's cognitive status
   3.4. the patient's motor function

METHODS

Subjects

The study took place in Edmonton, Alberta, Canada during September to December, 1993, and was repeated in the same 4 months of 1994. The study group consisted of 132 subjects, a subgroup of the 315 patients discharged from the Northern Alberta Regional Geriatric Program to their homes during the study period. Non-participation was due to residing outside of Edmonton (n=47); refusal (n=38); (re-)admission to an inpatient bed (n=29), day hospital (n=20) or continuing care programs (n=2); missing discharge FIM (n=28); death (n=1); a non-geriatric case (n=1); and unsuccessful follow-ups (n=17).

Functional Independence Measure

The FIM was developed by a task force sponsored in 1984 by the American Academy of Physical Medicine and Rehabilitation and the American Congress of Rehabilitation Medicine to assess an individual's level of functional independence (Granger, Hamilton, Linacre, Heinemann, & Wright, 1993; Smith, Hamilton, & Granger, 1990). The FONE FIM was designed as its telephone version and has the same 18 items as the FIM. The FIM has been shown to consist of at least 2 dimensions: 13 "motor" and 5 "cognitive" items (Heinemann, Linacre, Wright, Hamilton, & Granger, 1993; Lina-

cre et al., 1994). Each of these items is designed to measure an aspect of functional independence on a 7-point scale: 1 and 2 for "complete dependence," 3-5 for "modified dependence", and 6-7 for "independence." Only the "motor" dimension is discussed in this paper.

Procedure

Five weeks after discharge, the participating patients or their significant others (if the patients were unable to respond) were contacted by telephone to administer the FONE FIM and arrange for a home visit in the following week. Those patients who responded to the FONE FIM were those who were able to communicate in English, did not have aphasia or significant hearing loss, and scored >17 on the Mini-Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975). The 1993 data were collected by a graduate student studying for a master's degree in occupational therapy and two RN research assistants, and the 1994 data by three OT students doing their practicums. All raters went through FIM training and testing. During the data collection phase, each rater assessed the same patients through telephone and home interviews. The research assistants, whenever possible, scheduled home visits at times when most FIM activities would normally occur (early morning or evening). If this was not possible, the raters asked the subject to simulate the FIM activities. As a last resort, a patient report was accepted or a "1" (Not Testable) was recorded.

Data Analysis

Validation of the motor component of the FONE FIM was based on Rasch rating scale analysis (Wright & Masters, 1982), which generated item difficulty and subject ability measures on a common interval scale (in logits, the natural logarithm of odds). Both the FONE FIM and the OBS FIM were subjected to Rasch analysis.

To obtain consistent Rasch person ability measures between the two modes, a combined Rasch analysis was performed to obtain "generalized item measures" by treating each subject measured on different occasions/modes as distinct individuals (Chang & Chan, 1995). The resulting ability measures from the combined analysis, instead of those obtained from separate calibrations of the FONE FIM and the OBS FIM data, were used to assess the degree of agreement between the 2 modes. This was done in

several ways:

1. Concordance in Rasch estimates between the 2 modes was examined in terms of intraclass correlation coefficients (ICCs) for random-effect models, stratified by respondent category, the rater's experience, and the patient's cognitive status and motor functioning. The ICC value higher than 0.75 was considered as an indication of good concordance, and lower than 0.75 as moderate concordance (Portney & Watkins, 1993).
2. The following influencing factors were considered: (a) respondent category, by which the patients were divided into self-reporting (n=104) and proxy-reporting (n=28) groups; (b) the rater's experience, by which the patients were grouped into those who were assessed by the 2 more experienced (n=63) and the 4 less experienced raters (n=69); (c) the patient's cognitive status, by which the patients were classified into "higher" (n=107) and "lower" (n=25) cognition groups depending on whether the total OBS FIM cognitive score was at least 30 or less; and (d) the patient's motor function, by which the sample was split into "higher" (n=81) and "lower" (n=51) functioning groups depending on whether the total OBS FIM motor score was at least 78 or less.
3. The possibility of over- or underestimation associated with the 2 modes was assessed first by the paired t-test. The difference scores in Rasch estimates between the 2 modes were further subjected to a hierarchical analysis of variance (ANOVA), with respondent category, the rater's experience, the patient's cognitive status and motor function as grouping variables, to identify factors influencing these difference scores.
4. A graphical method was also used to demonstrate the presence of over- or underestimation of the FONE FIM relative to the OBS FIM.
5. For clinical purposes, it may be necessary to establish a high degree of concordance also at the FIM item level. This was accomplished by examining the coefficients of agreement and kappa statistics for each item: the agreement was judged to be "excellent" when Kappa >.80, "substantial" when Kappa>.60, "moderate" when Kappa >.40, and "poor to fair" when Kappa <.40 (Portney, & Watkins, 1993).
6. To adjust for any discord between the 2 modes, the OBS FIM motor logit scores were regressed on possible influencing factors such as respondent category, the rater's experience, the patient's age, gender, length of stay, cognitive status, and the FONE FIM motor logit scores. The resulting model was used to generate adjusted FONE FIM logit scores that could be used to improve the concordance between the 2 modes.

The computer program FACETS (Linacre & Wright, 1993) was used to

generate Rasch item difficulty and subject ability estimates and related fit statistics, and SPSS for Windows, Version 6.1, was used to perform other statistical procedures.

RESULTS

Sample Characteristics

Of the 132 subjects in our sample, 68% were female. The average age was 79 years. The average MMSE score was 25 on admission and 26 at follow-up. The lengths of stay averaged 38 days, and 68% used Home Care services. These characteristics were similar to those of the 315 patients who had been discharged to their homes during the study period.

Concordance in Terms of Rasch Person Ability Measures

A key measure of concordance between two or more continuous variables is the intraclass correlation coefficient (ICC). ICCs between the two modes were computed for all subjects and by subgroups: 0.90 for all subjects, and also for both patient and proxy respondent groups; 0.92 and 0.94 for patients assessed by more and less experienced raters; 0.88 and 0.94 for higher and lower cognition patients; and 0.82 and 0.85 for higher and lower motor function groups, respectively. Thus, good concordance of >0.75 between the two modes was observed for the entire sample and selected subgroups. However, concordance between modes as measured by the ICC was not higher for the patient respondents, the more experienced raters, the higher cognition group, and the higher motor function group as was originally expected.

The differences in the person ability estimates between the two modes were compared next, using a paired t-test. The estimates turned out to be significantly lower for the FONE FIM than the OBS FIM (Ms=1.49 and 1.62 logits, $t(131)=-2.88$, $p=0.005$).

To investigate whether the differences in Rasch person ability estimates were influenced by factors such as the respondent's category, the rater's experience, and the patient's cognitive and motor function status, the mean and the standard deviation were first tabulated, stratified by these factors (Table 2). A hierarchical analysis of variance (ANOVA) procedure was applied to the differences between the two sets of estimates, based on a 2^4 factorial design (Table 3). The results confirmed that the

Table 2
Mean (Standard Deviation) of Person Ability Logits by Mode Stratified by Respondent Category, Cognative Status, Motor Function and Rater's Experience

Category	N	FONE FIM Mean(S.D.)	OBS FIM Mean(S.D.)
RESPONDENT CATEGORY:			
Proxy	28	0.84(1.44)	0.92(1.39)
Patient	104	1.66(0.98)	1.80(1.11)
RATER EXPERIENCE:			
Less Experienced	69	1.57(1.10)	1.67(1.17)
More Experienced	63	1.40(1.19)	1.56(1.28)
COGNITIVE STATUS:			
Cognitive Score <30	25	0.85(1.18)	0.75(1.21)
Proxy Respondent		1.03(1.51)	0.87(1.59)
Patient Respondent		0.61(0.53)	0.58(0.46)
Cognitive Score≥30	107	1.64(1.08)	1.82(1.14)
Proxy Respondent		0.66(1.40)	0.97(1.23)
Patient Respondent		1.79(0.95)	1.95(1.07)
MOTOR FUNCTION:			
Motor Score<78	51	0.53(0.79)	0.45(0.70)
Motor Score ≥78	81	2.09(0.88)	2.35(0.86)

Table 3
Hierarchical Analysis of Variance with Respondent Category, Rater Experience, and Motor and Cognative Function as Factor Variables

Source of Variation	Sum of Squares	DF	Mean Square	F	P-value
Main Effects	4.335	4	1.084	4.833	.001
Respondent	.077	1	.077	.344	.559
Rater	.158	1	.158	.706	.402
Cognition	1.978	1	1.978	8.818	.004
Motor	2.122	1	2.122	9.463	.003
2-Way Interactions	1.874	6	.312	1.392	.223
Respondent x Rater	.264	1	.264	1.179	.280
Respondent x Cognition	1.248	1	1.248	5.567	.020
Respondent x Motor	.282	1	.282	1.259	.264
Rater x Cognition	.331	1	.331	1.477	.227
Rater x Motor	.543	1	.543	2.421	.122
Cognition x Motor	.426	1	.426	1.902	.170
Explained	6.209	10	.621	2.769	.004
Residual	27.137	121	.224		
Total	33.346	131	.255		

patient's cognitive status, $F(1,121)=8.82$, $p=0.004$, and motor function, $F(1,121)=9.46$, $p=0.003$, significantly affected the discrepancies in the estimates. In addition, the respondent's category significantly interacted with the patient's cognitive status, $F(1,121)=5.57$, $p=0.020$. The rater's experience, however, did not turn out to be significant.

Functional assessment differed according to the patients' cognitive status. For the higher cognition group, the estimates from the FONE FIM were lower than from the OBS FIM, Ms=1.64 and 1.82, SDs=1.08 and 1.14 (Table 2). Figure 1 also shows the higher cognition group being assessed at lower levels on the FONE FIM compared to the OBS FIM, as shown by more points below the horizontal 0-axis. In the lower cognition group, the FONE FIM resulted in higher estimates than the OBS FIM, Ms=0.85 and 0.75, SDs=1.18 and 1.21. However, this trend is less apparent from the graph in Figure 1.

Functional assessment also depended on the levels of patients' motor function, as shown by a non-horizontal regression line in Figure 2. Among patients with higher levels of motor functioning, the estimates from the FONE FIM were lower than from the OBS FIM, Ms=2.09 and 2.35, SDs=0.88 and 0.86. Figure 2 illustrates the same finding of lower estimates on the FONE FIM from the higher function group, as more points were located below the horizontal 0-axis. For the patients with lower levels of motor functioning, the FONE FIM generated higher estimates than the OBS FIM, Ms=0.53 and 0.45, SDs=0.79 and 0.70. Figure 2 graphically shows this same information with more points located above the horizontal 0-axis.

The interaction between the respondent's category and the patient's cognitive status showed that estimation of patients' functional status depended both on who responded to the telephone assessment and on the patients' cognitive status (Figure 3). For the higher cognition group (indicated by the solid lines), the estimates were lower for the patients who did not respond to the FONE FIM than for those who did, regardless of the modes. However, for the lower cognition group (indicated by the dotted lines), the estimates were higher for patients who did not respond to the FONE FIM than for those who did, again irrespective of the modes. These nonparallel lines pointed to the existence of interactions between the patient's cognitive status and respondent category.

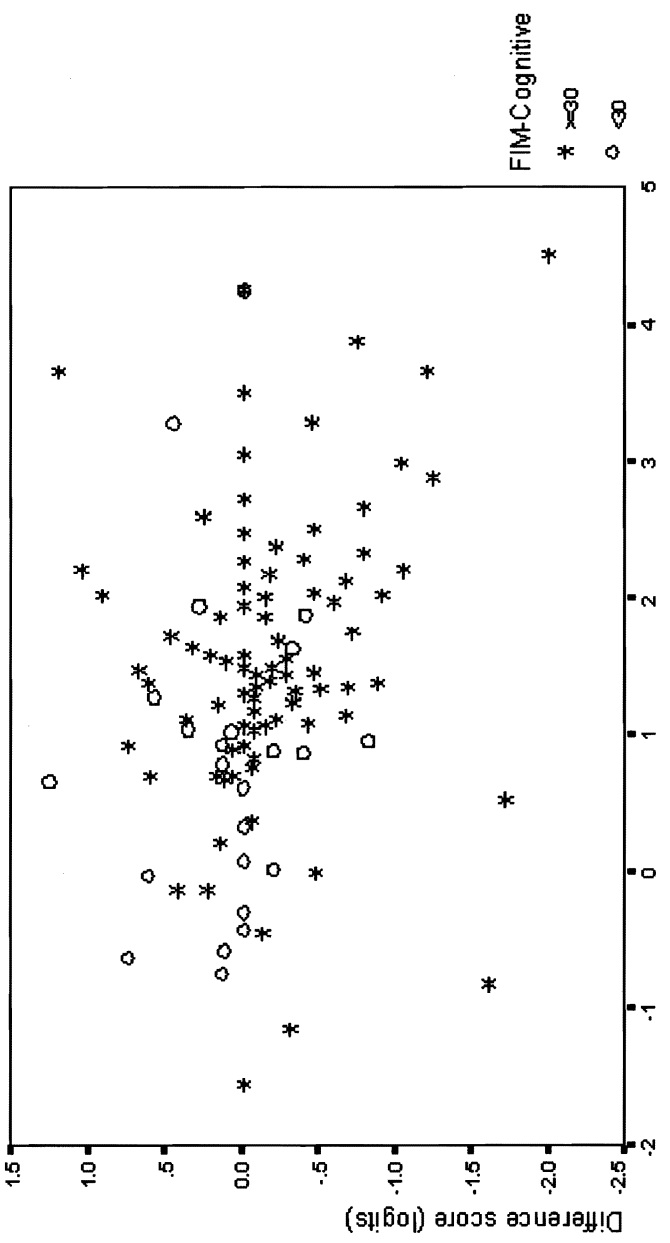


FIGURE 1 Difference (FONE FIM logits - OBS FIM logits) vs. Average scores of the Rasch estimates obtained from the two modes, stratified by the higher (≥=30) and lower (<30) total score for the cognitive component of the OBS FIM. For the higher cognition group, at least, the FONE FIM underestimated the motor function relative t the OBS FIM as shwon by more points below the 0-axis.

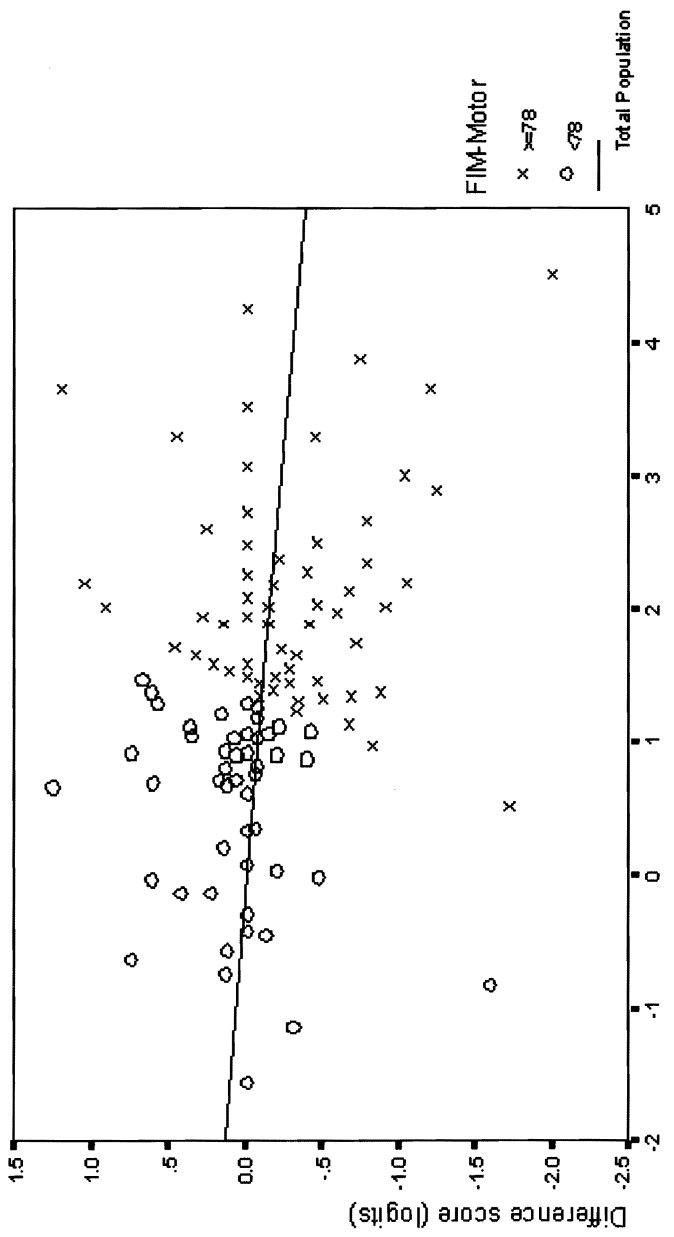


FIGURE 2 Difference vs. Average scores between the FONE FIM and the OBS FIM, stratified by higher (≥78) and lower (<78) motor function groups based on the motor component of the OBS FIM. For the higher function group, the FONE FIM underestimated the patients' motor ability relative t the OBS FIM, as more points were located below th e 0-axis. For the lower function group, more points are located above the 0-axis, indicating that the FONE FIM overestimated the patient's motor function relative to the OBS FIM.

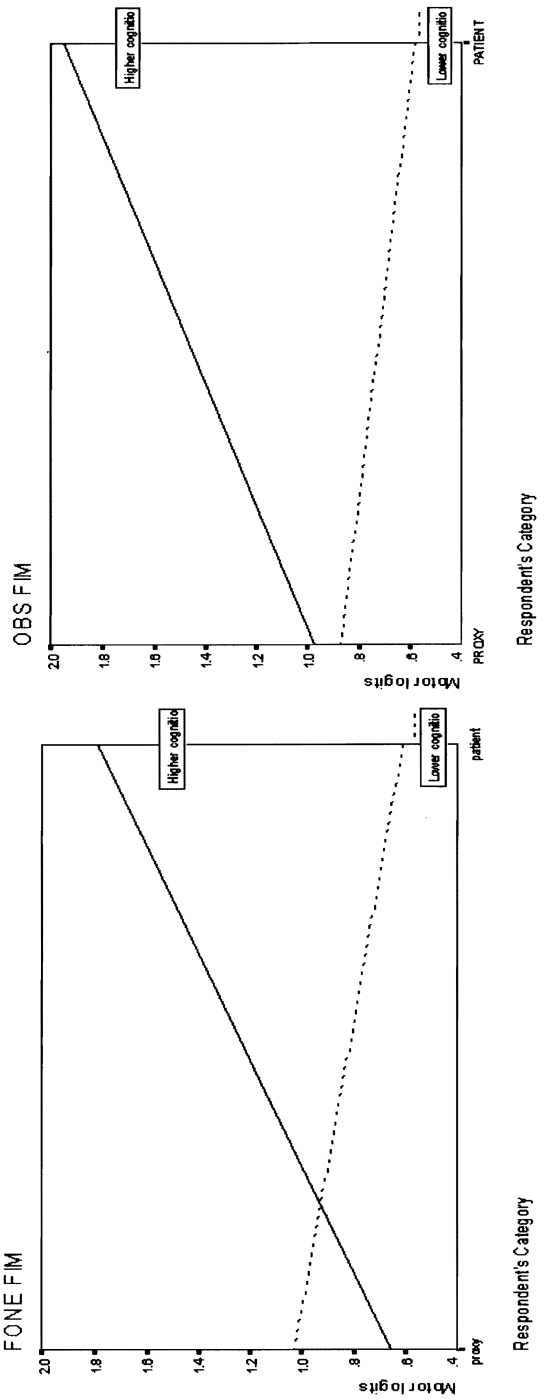


FIGURE 3 Mean motor ability estimates for higher and lower cognition groups from the FONE FIM (left) and the OBS FIM (right), indicating the presence of interaction effects between the patient's cognitive status and respondent category.

Concordance at the Item Level

Concordance between the FONE FIM and the OBS FIM at the item level was examined next. When examining all responses, 9 of the 13 items showed "substantial" levels of concordance when all subjects were examined, with Kappa statistics in the range of 0.60 to 0.80 (Table 4). The remaining 4 items of Bathing, Bed Transfer, Tub Transfer and Climbing Stairs showed "moderate" levels of concordance with Kappa values falling between 0.55 and 0.59. The Kappa value for the total FONE FIM motor score (dichotomized to $\geq$78 and <78) was 0.72, indicating a "substantial" concordance.

To address the concern that the respondent's category may influence functional status measurement, two groups of patients were identified: those who responded to the FONE FIM themselves and those who did not (the patient and the proxy group). At the item level, the agreement between the 2 modes tended to be far better for the patient than the proxy group for most items, despite a comparable level of concordance when the total FIM scores (dichotomized at 78) were used (Table 4). The Kappa statistics for the patient respondent group were 0.60 or higher for 11 items, indicating "substantial" agreements between the modes. Only Tub Transfer and Climbing Stairs had a Kappa value falling below 0.60, but still indicating a "moderate" agreement. In contrast, the Kappa statistic for the proxy group were <0.60 for 10 items, with Toileting, Toilet Transfer, and Tub Transfer as only exceptions.

The rater's experience was examined next. At the item level, the Kappa values were invariably, and sometimes substantially, higher for the more experienced than the less experienced raters for all items (Table 5). The Kappa statistics for the more experienced raters showed an "excellent" level of agreement in 2 items, a "substantial" level in 10 items, and a "moderate" agreement only in Tub Transfer. In contrast, the Kappa statistics for the less experienced raters showed the level of agreement to be "substantial" for 3 items, "moderate" for 9 items, and "poor to fair" for 1 item.

The cognitive factor was examined by stratifying the patients into 2 groups of "high" and "low" cognition, based on the total OBS FIM cognitive dimension score $\geq$30 (n =107) vs. <30 (n = 25). The Kappa statistics showed the level of agreement to be "substantial" for 10 items for the lower cognition group, and for 7 items for the higher cognition group (Table 6). The Kappa statistics for the total motor score (dichotomized again at

Table 4
Coefficients of Agreement (COA) and Kappa Statistics for between the FONE & the OBS FIM: Patient vs. Proxy Responses

	Patient		Proxy		All Patients	
Items	COA	Kappa	COA	Kappa	COA	Kappa
Eating	0.86	0.66+	0.64	0.43*	0.83	0.60+
Grooming	0.89	0.72+	0.54	0.46*	0.84	0.65+
Bathing	0.72	0.62+	0.57	0.48*	0.68	0.59*
Dressing Upper Body	0.87	0.70+	0.68	0.57*	0.83	0.67+
Dressing Lower Body	0.84	0.65+	0.61	0.51*	0.79	0.62+
Toileting	0.88	0.72+	0.71	0.62+	0.84	0.77+
Bladder Management	0.81	0.63+	0.64	0.50*	0.77	0.60+
Bowel Management	0.84	0.71+	0.64	0.52*	0.80	0.66+
Bed Transfers	0.80	0.62+	0.54	0.42*	0.74	0.57*
Toilet Transfer	0.83	0.63+	0.71	0.61+	0.80	0.63+
Tub Transfer	0.66	0.55*	0.64	0.60+	0.65	0.55*
Walking	0.89	0.80++	0.68	0.59*	0.84	0.62+
Climbing Stairs	0.67	0.58*	0.61	0.53*	0.67	0.57*
Total Motor Score	0.87	0.72+	0.86	0.71+	0.86	0.72+

Note. To calculate the kappa statistics, the total motor was dichotomized at a cutoff of $\geq$ 78.

++ "excellent" concordance
+ "substantial" concordance
* "moderate" concordance
** "poor to fair" concordance

Table 5
Coefficients of Agreement (COA) and Kappa Statistics Between the FONE & the OBS FIM by Rater Group

	More Experienced		Less Expereinced	
Items	COA	Kappa	COA	Kappa
Eating	0.89	0.74+	0.77	0.46*
Grooming	0.83	0.70+	0.85	0.59*
Bathing	0.76	0.68+	0.61	0.50*
Dressing Upper Body	0.94	0.90++	0.74	0.49*
Dressing Lower Body	0.86	0.78+	0.71	0.45*
Toileting	0.81	0.69+	0.84	0.61+
Bladder Management	0.78	0.62+	0.78	0.59*
Bowel Management	0.79	0.66+	0.78	0.64+
Bed Transfers	0.89	0.80++	0.62	0.39**
Toilet Transfer	0.84	0.72+	0.76	0.51*
Tub Transfer	0.65	0.56*	0.65	0.52*
Walking	0.86	0.78+	0.82	0.72+
Climbing Stairs	0.71	0.65+	0.60	0.49*
Total Motor Score	0.87	0.75+	0.85	0.70+

Note. To calculate the kappa statistics, the total motor was dichotomized at a cutoff of ≥78.

++ "excellent" concordance
+ "substantial" concordance
* "moderate" concordance
** "poor to fair" concordance

Table 6
Coefficients of Agreement (COA) and Kappa Statistics
Between the FONE & the OBS FIM by Cognitive Status of Patients at Follow-up

	Higher		Lower	
Items	COA	Kappa	COA	Kappa
Eating	0.86	0.54*	0.72	0.60+
Grooming	0.88	0.67+	0.60	0.45*
Bathing	0.68	0.57*	0.72	0.64+
Dressing Upper Body	0.83	0.64+	0.76	0.68+
Dressing Lower Body	0.80	0.58*	0.76	0.70+
Toileting	0.87	0.72+	0.72	0.63+
Bladder Management	0.81	0.63+	0.60	0.50*
Bowel Management	0.79	0.64+	0.76	0.64+
Bed Transfers	0.75	0.54*	0.72	0.63+
Toilet Transfer	0.80	0.62+	0.80	0.67+
Tub Transfer	0.66	0.55*	0.60	0.51*
Walking	0.86	0.75+	0.76	0.66+
Climbing Stairs	0.63	0.53*	0.76	0.71+
Total Motor Score	0.85	0.68+	0.92	0.78+

Note. Cognitive statisitics dichotomized at ≥30. To calculate Kappa staticistics, the total motor score was dichotomized at ≥78.

++ "excellent" concordance + "substantial" concordance

* "moderate" concordance ** "poor to fair" concordance

78) was somewhat higher for the lower cognition group than for the higher cognition group, (0.78 vs. 0.68). However, no firm conclusion can be drawn on the relative level of concordance between higher and lower cognition groups.

Is concordance between the modes affected by the patient's level of Motor function? To answer this question, the respondents were also divided into "higher" and "lower" motor function groups based on the total OBS FIM Motor score of $\geq$78 (n=81) and < 78 (n=51). Concordance, as measured by the coefficient of agreement, was generally better for the higher than the lower function group (for most items except Bathing, Tub Transfer, and Climbing Stairs). However, a "substantial" level of agreement, as measured by Kappa statistics, was indicated for 5 items in the higher function group and 8 items in the lower function group. Again, no firm conclusion regarding the relative level of concordance can be drawn for the higher and lower motor function groups (Table 7).

Statistical Adjustment for Improving Concordance

The discord between the 2 modes may be partially corrected by a regression method. To investigate this possibility, the OBS FIM motor logit scores were regressed on possible influencing factors such as the respondent's category, the rater's experience, and the patient's age, gender, length of stay, the Mini-Mental State Examination scores on admission and at discharge, the FIM cognitive scores at discharge and at follow-up in-home assessment, and the FONE FIM motor logit scores. The FONE FIM motor logit scores were the most important factor, due to a high correlation coefficient of 0.91 with the OBS FIM motor scores. The regression coefficient was 0.98, and the adjusted R^2 was 0.83. The only other statistically significant, independent predictor was the OBS FIM cognitive score, dichotomized into at least 30 or less (Table 8). By adding this factor, the adjusted R^2 was increased marginally to 0.84. These regression equations may be used to adjust the FONE FIM logit scores to render them more consistent with the results of in-home assessment. When the FONE FIM logit scores were adjusted using data from the cognitive component of the OBS FIM, the adjusted difference scores between the FONE FIM and the OBS FIM motor logit scores had a mean of 0, and a slightly smaller standard deviation than the unadjusted difference score, Ms=0.00 and -0.13, SDs=0.043 and 0.044. Regressing the OBS FIM motor logits only on the FONE FIM motor logits resulted in a minimal mean difference

Table 7
Coefficients of Agreement (COA) and Kappa Statistics
Between the FONE & the OBS FIM by Motor Function Level

	Higher Function		Lower Function	
Items	COA	Kappa	COA	Kappa
Eating	0.88	0.87+	0.76	0.63+
Grooming	0.94	0.56*	0.69	0.57*
Bathing	0.68	0.49*	0.69	0.59*
Dressing Upper Body	0.90	0.49*	0.71	0.62+
Dressing Lower Body	0.89	0.44*	0.63	0.54*
Toileting	0.89	0.55+	0.77	0.74+
Bladder Management	0.82	0.47*	0.71	0.62+
Bowel Management	0.82	0.64+	0.74	0.60+
Bed Transfers	0.82	0.49*	0.63	0.38**
Toilet Transfer	0.83	0.65+	0.76	0.50*
Tub Transfer	0.63	0.46*	0.68	0.61+
Walking	0.89	0.79+	0.75	0.65+
Climbing Stairs	0.59	0.45*	0.76	0.67+
Total Motor Score	0.81	n.a.	0.94	n.a.

Note. To calculate the kappa statistics, the total motor score was dichotomized at a cutoff of ≥78.

++ "excellent" concordance
+ "substantial" concordance
* "moderate" concordance
** "poor to fair" concordance

Table 8
Coefficients of Agreement (COA) and Kappa Statistics
Between the FONE & the OBS FIM by Motor Function Level

Variable	Coefficient	S.E.	95% Confidence Interval	
FONE FIM motor logits	0.947	0.393	0.870	0.025
OBS FIM cognitive*	0.322	0.114	0.097	0.541
Constant	-0.056	0.104	-0.262	0.150
Multiple R	0.916			
R Square	0.840			
Adjusted R Square	0.837			
Standard Error	0.493			

*Based on the total OBS FIM cognitive score, dichotomized as ≥ 30 or < 30

between the adjusted FONE FIM and the OBS FIM scores without reducing the standard deviation, M=0.001, SD= 0.044.

DISCUSSION

It was reassuring that the Rasch Motor ability measures derived from the FONE FIM and the OBS FIM were shown to be in accord, as measured by ICCs, for all patients as well as for various subgroups as defined by the respondent category, the rater's experience, and the patient's cognitive status and motor function. The ICCs were 0.88 or higher for the entire group and most subgroups, including the patient and proxy groups, the more and less experienced rater groups, and the higher and lower cognition patient groups. The ICC's were 0.82 or higher for the higher and lower motor function groups, still indicating good concordance (ICC > 0.75) between the 2 modes. A similar conclusion was reached at the item level: the lowest Kappa value for all respondents was 0.55 for Tub Transfer, which was higher than 0.45 required by the Uniform Data System for at least 15 of the 18 FIM items to meet the interrater reliability standard (Hamilton, Laughlin, Fiedler, & Granger, 1994). Very few Kappa statistics fell below 0.40 for any subgroups. A notable exception is Bed Trans-

fer, with Kappa statistics of 0.42 for the proxy respondents, 0.39 for the less experienced raters, and 0.38 for the lower motor function group. Those associated with a Kappa value of 0.45 or below included: Eating for proxy respondents, Grooming for lower cognition patients, and Dressing Lower Body for less experienced raters and for lower motor function patients.

The FONE FIM consistently generated lower estimates of Rasch motor measures than the OBS FIM (as shown by a paired t-test), a conclusion also reached by most other research using a variety of functional measures as shown in our literature review. A key to understanding this phenomenon may lie in the patient's cognitive status and motor function, since the FONE FIM overestimated the motor function for the lower cognition and the lower motor function group and underestimated for the higher cognition and the higher motor function group. The issue is further complicated by a significant interaction between the patient's cognitive status and respondent category. Indeed, the patient's low cognitive status was one of the reasons why a proxy responded to the FONE FIM: while 87% of those in the higher cognition group responded themselves, only 44% of those in the lower cognition group did. Hence, functional assessments on the FONE FIM and the OBS FIM of lower cognition patients, Ms=0.85 and 0.75, corresponded more closely to those of proxy respondents, Ms=1.03 and 0.87, than of patient respondents, Ms=0.61 and 0.58, whereas patients' own assessments, Ms=1.79 and 1.95, weighed more heavily than those of proxies, Ms=0.66 and 0.97, when assessing patients in the higher cognition group, Ms=1.64 and 1.82 (Table 2). Interestingly, Korner-Bitensky et al. (1994, 1995) also found less frequent reporting of disability (overestimation) on the telephone by those with moderate and severe disability. Since 81% of the sample belonged to the higher cognition group and 61% were in the higher motor function group, it is not surprising that functional assessments based on the FONE FIM tended to be lower than those based on the OBS FIM.

The discord between the 2 modes was also due to the undifferentiated coding of "Not Testable" and "Total Assist", both coded as a "1" in the FONE FIM and the OBS FIM. This practice particularly affected the items that were difficult to perform, such as Climbing Stairs and Bathing. In assessing Climbing Stairs in the lower motor function group, for instance, the rating of "1" was accorded on the OBS FIM to 5 patients, who otherwise were assessed as "2" for 1 patient, "4" for 2 patients, "5" for 1 patient, and "6" for 1 patient on the FONE FIM. It may be inferred from this that the 4 patients with a FONE FIM score of 4 or above were "Not

Testable" cases on the OBS FIM. Similarly, in assessing Climbing Stairs in the higher motor function group, the rating of "1" was given on the FONE FIM to 5 patients, who were assessed on the OBS FIM as "2" for 1 patient, "5" for 2 patients, and "6" for 2 patients. The last 4 patients were again likely to be "Not Testable" cases on the FONE FIM. Hence, this undifferentiated coding was partially responsible for the discord between the 2 modes, and a revision of coding instructions would likely reduce the discord between the 2 modes (Linacre et al., 1994).

Another important finding is that the accord between the modes shown at the macro level of analysis masked the discord at the item level. It was found that the two modes tended to be more in accord at the item level as measured by Kappa statistics, if the responses were obtained from the patients rather than the proxies, and by the more experienced rather than the less experience raters. The effects of the patient's cognitive and motor function were less consistent and clear-cut. This result was neither reflected in the Kappa statistics for the dichotomized total motor score (Tables 4-7), nor in the ICCs for Rasch scores. Due to the practical importance of concordance at the item level, a reasonable strategy to pursue would be to get as many valid patient responses as possible regardless of their cognitive and physical function, since many of such patients who initially refused to be interviewed on the telephone subsequently participated in face-to-face interviews (Norton, Breitner, Welsh, & Wyse, 1994). Thus, a cost-effective approach would be to use the telephone mode for the majority of patients (79% in our study) who are capable of responding to the FONE FIM themselves, and to rely on home visits to assess other patients. This mixed telephone/face-to-face approach was advocated by some (Crawford, Jette, & Tennstedt, 1997) but rejected by others on the ground that it may lead to biased results (Epstein et al., 1989). In our view, this approach will most likely meet the goal of generating accurate data at a reasonable cost. The use of this mixed strategy would require a careful assessment of the magnitudes of the biases that are likely to be associated with varying levels of cognitive and motor functioning. The proposed regression method would minimize the discord by eliminating the mean difference between the two sets of estimates and reducing the standard deviation. As well, as shown by Fricke et al. (1993), the training and experience of the raters should be a key consideration in this strategy in order to ensure the validity and reliability of functional assessment.

There are several limitations in this study. The validity of the observational method may be questioned, since the rater who visits for a short

time during the day may not be able to actually observe the patient performing all items needed for proper assessment. As well, in-home assessment may be associated more with the patient's peak performance than normal performance. Test-retest reliability and interrater reliability had not been rigorously tested prior to commencing the study, although attempts were made to train all raters and compare their assessments. The order of administering the two modes was not randomized, therefore the raters were aware of the FONE FIM scores prior to doing the home assessment. The discrepancies between the two modes may still persist even with randomization, however, as shown for the SF-36 by Weinberger et al. (1996). Finally, the sample size was relatively small to allow for a more detailed analysis of factors influencing the concordance between modes, based on a factorial design.

Conclusion

The issues of concordance and over- and underestimation between the telephone and in-home observational mode has been examined in this paper. It has been shown that concordance is generally good to excellent for the Rasch motor scale, and moderate to excellent for most items at the item level. However, there was significant under- or overestimation of Rasch motor measures associated with the patient's level of cognition and motor function. Concordance between the 2 modes at the item level was shown to be influenced by the respondent category and the rater's experience. A strategy to minimize discord and improve data quality should, therefore, include recruitment and training of experienced raters, use of the telephone mode to obtain information from patient rather than proxy respondents whenever feasible, reliance on home visits when patient telephone responses are unavailable or of questionable quality, and use of statistical methods to minimize discrepancies by making adjustments to the results of telephone assessment.

ACKNOWLEDGMENT

We would like to thank the reviewer and editor for providing helpful comments and suggestions to improve the quality and presentation of this paper.

REFERENCES

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

Chang, W. C., Slaughter, S. E., Cartwright, D., & Chan, C. (1997). Evaluating the FONE FIM. Part I. Construct validity. *Journal of Outcome Measurement, 1(3),* 192-218.

Clarridge, B.R., & Passagli, M.P. (1989). The use of female spouse proxies in common symptom reporting. *Medical Care, 27(4),* 352-366.

Crawford, S. L., Jette, A. M., & Tennstedt, S. L. (1997). Test-retest reliability of self-reported disability measures in older adults. *Journal of the American Geriatrics Society, 45,* 338-341.

Edwards, M. M. (1990). The reliability and validity of self-report activities of daily living scales. *Canadian Journal of Occupational Therapy, 57,* 273-278.

Epstein, A. M., Hall, J.A., Tognetti, J., Son, L. H., & Conant, L., Jr. (1989). Using proxies to evaluate quality of life. Can they provide valid information about patients' health status and satisfaction with medical care? *Medical Care 27(Suppl.), S91-S98.*

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-Mental State". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,*189-198.

Fricke, J., Unsworth, C., & Worrell D. (1993). Reliability of the Functional Independence Measure with occupational therapists. *Australian Occupational Therapy Journal, 40,* 7-15.

Granger, C. V., Hamilton, B. B., Linacre, J. M., Heinemann, A. W., & Wright, B. D. (1993). Performance profiles of the functional independence measure. *American Journal of Physical Medicine & Rehabilitation, 72,* 84-89.

Guyatt, G. H., Feeny, D. H., & Patrick, D. L. (1993). Measuring health-related quality of life. *Annals of Internal Medicine, 118,* 622-629.

Hamilton, B. B., Laughlin, J. A., Fiedler, R. C., & Granger, C. V. (1994). Interrater reliability of the 7-level Functional Independence Measure (FIM). *Scandinavian Journal of Rehabilitation Medicine, 26,* 115-119.

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 & Rehabilitation, 74,* 566-573.

Jette, A. M. (1987). The Functional Status Index: reliability and validity of a self-report functional disability measure. *Journal of Rheumatology, 14(Suppl. 15),* 15-19.

Korner-Bitensky, N. & Wood-Dauphinee, S. (1995). Barthel Index information elicited over the telephone. *American Journal of Physical Medicine & Rehabilitation, 74(1),* 9-18.

Korner-Bitensky, N., Wood-Dauphinee, S., Siemiatycki, J., Shapiro, S., & Becker, R. (1994). Health-related information postdischarge: telephone versus face-to-face interviewing. *Archives of Physical Medicine & Rehabilitation, 75,* 1287-1296.

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 & Rehabilitation, 75,* 127-132.

Linacre, J. M. & Wright, B. D. (1993). *FACETS, Many-Facet Rasch Analysis with FACFORM, Data Formatter.* Chicago, IL: MESA Press.

Magaziner, J., Simonsick, E. M., Kashner, T. M., & Hebel, J. R. (1988). Patient-proxy response comparability on measures of patient health and functional status. *Journal of Clinical Epidemiology, 41,* 1065-1074.

Norton, M. C., Breitner, J. C. S., Welsh, K. A., & Wyse, B. W. (1994). Characteristics of nonresponders in a community survey of the elderly. *Journal of the American Geriatrics Society, 42,* 1252-1256.

Portney, L. G. & Watkins, M. P. (1993). *Foundations of Clinical Research. Applications to Practice.* Norwalk, CT: Appleton & Lange.

Roccaforte, W. H., Burke, W. J., Bayer, B. L., & Wengel, S. P. (1992). Validation of a telephone version of the Mini-Mental State Examination. *Journal of the American Geriatrics Society, 40,* 697-702.

Rothman, M. L., Hedrick, S.C., Bulcroft, K.A., Kickam, D. H., & Rubenstein, L. Z. (1991). The validity of proxy-generated scores as measures of patient health status. *Medical Care, 29,* 115-124.

Rowley, G. & Fielding, K. (1991). Reliability and accuracy of the Glasgow Coma Scale with experienced and inexperienced users. *The Lancet, 337,* 535-538.

Rubenstein, L. Z., Schairer, C., Wieland, G. D., & Kane, R. (1984). Systematic biases in functional status assessment of elderly adults: effects of different data sources. *Journal of Gerontology, 39,* 686-691.

Sager, M. A., Dunham, N. C., Schwantes, A., Mecum, L., Halverson, K., & Harlowe, D. (1992). Measurement of activities of daily living in hospitalized elderly: a comparison of self report and performance-based methods. *Journal of the American Geriatrics Society, 40,* 457-462.

Shinar, D., Gross, C. R., Bronstein, K. S., Eden, D. T., Cabrera, A. R., Fishman, I.G., Roth, A.A., Barwick, J.A., & Kunitz, S.C. (1987). Reliability of the activities of daily living scale and its use in telephone interview. *Archives of Physical Medicine & Rehabilitation, 68,* 723-728.

Smith, P. (1992). Collecting followup data. *UDS Update, August,* 1-2.

Smith, P., Hamilton, B.B., & Granger, C.V. (1990). *The FONE FIM.* Buffalo, NY: Research Foundation of the SUNY.

Weinberger, M., Nagle, B., Hanlon, J. T., Samsa, G. P., Schmader, K., Landsman, P. B., Uttech, K. M., Cowper, P. A., Cohen, H. J., & Feussner, J. R. (1994). Assessing health-related quality of life in elderly outpatients: telephone versus face-to-face administration. *Journal of the American Geriatrics Society, 42,*1295-1299.

Weinberger, M., Oddone, E. Z., Samsa, G. P., & Landsman, P. B. (1996). Are health-related quality-or-life measures affected by the mode of administration? *Journal of Clinical Epidemiology, 49(2),* 135-140.

Weinberger, M., Samsa, G. P., Schmader, K., Greenberg, S. M., Carr, D. C., & Wildman, D. S. (1992). Comparing proxy and patients' perceptions of patients' functional status: results from an outpatient geriatric clinic. *Journal of the American Geriatrics Society, 40,* 585-588.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago, IL: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, 1(4), 286-304

Post-Hoc Rasch Analysis of Optimal Categorization of an Ordered-Response Scale

Weimo Zhu
Wayne State University

Wynn F. Updyke
and
Cheryl Lewandowski
Indiana University

The purpose of this study was to determine the optimal categorization of a self-efficacy ordered-response scale using the Rasch analysis and compare theperformance of the Rasch statistics and parameter estimates with conventionalstatistics. A 50-item scale to measure psychomotor self-efficacy was administered to a total of 2,022 children, including 1,009 boys and 1,013 girls. The data analysis started by collapsing the original five adjacent categories into two, three, and four categories, and a total of 14 data sets were derived. Each of these data sets, including the original one, was analyzed using the Rasch rating scale model, and a set of Rasch model-data fit, category, and separation statistics and parameter estimates, as well as three conventional statistics, were computed and compared. It was found that, instead of the five-category construct designed, the best order of category meanings of the scale in respondents' perceptions was a three-category construct. The Rasch threshold estimates were sensitive indexes in determining the order of the categorization, and that item separation statistics were useful in determining the optimal categorization after its order was confirmed. The commonly used coefficient alpha was found not helpful at all in determining the optimal categorization. The Rasch analysis was demonstrated to be a useful post-hoc analytic approach in determining the optimal categorization of an ordered-response scale.

Requests for reprints should be sent to Weimo Zhu, Ph.D., Division of Health, Physical Education, and Recreation, Wayne State University, 257 Matthaei Building, Detroit, MI 48202.

Categorization has always been considered an important element in constructing an ordered-response scale. Ordered-response scales include scales having ordinal response categories. Rating scales, Likert scales, and summated scales are just a few familiar examples. Categorization of an ordered-response scale has two very important characteristics. First, while all categories of a scale should measure a common trait or property (e.g., feeling, attitude, or opinion), each of them must also have its own well-defined boundaries, and the elements in a category should all share certain exclusively specific properties. Second, categories must be in an order and numerical values generated from the categories must reflect the degrees or magnitudes of the trait (Andrich, 1997; Guilford, 1954). An optimal categorization is a one that best exhibits these characteristics.

Several factors, such as number of categories (Miller, 1956; Parducci & Wedell, 1986), label and position (Klockars & Yamagishi, 1988; Wildt & Mazis, 1978), and type of anchors (Wedell, Parducci, & Lane, 1990), have been found to affect the categorization of a scale. In the past, determination of the categorization has been based mainly upon scale developers' prior knowledge of those factors. While every effort should be made to try to appropriately construct the categorization of an ordered-response scale before administering it to a large sample, such efforts do not guarantee that the categorization constructed will perform as designed. Therefore, it has been a common practice to examine respondents' responses statistically at both scale and item levels after the scale is administered to a sample.

Among commonly used conventional statistics, coefficient alpha (Cronbach, 1951) is perhaps the most popular one at the scale level. Coefficient alpha is a measure of the internal consistency of a scale, and is a direct function of both the number of items in the scale and their magnitude of intercorrelation. Therefore, either increasing the number of items or raising their intercorrelations can increase coefficient alpha. Further, it generally believed that increasing the number of categories will increase coefficient alpha, but maximum gains will be reached with five or seven scale-points, after which coefficient alpha values will level off (Bandalos & Enders, 1996).

Another commonly used conventional statistical index is the item point-biserial correlation coefficient, which reflects the correlation between responses and respondents' total scores. The point-biserial correlation coefficient is a discrimination index at the item level. Generally, the higher the point-biserial coefficient, the better the discrimination of an item, and a

negative value often reveals a problematic item. While both coefficient alpha and point-biserial coefficient may be used to examine the quality of a scale or an item, neither provides any information on the quality of the categories. Clearly, a new approach is needed, and the Rasch analysis shows good potential.

The Rasch analysis was not originally developed for determining the optimal categorization, but rather as a measurement model[1] (Rasch, 1960/80; Wright & Stone, 1979; Wright & Master, 1982), known as the one-parameter logistic model under the framework of the item response theory (IRT). However, information provided by the Rasch analysis, especially that on categories, make it very useful for such a purpose. Conceptually, the Rasch analysis belongs to a post-hoc (Andrich, 1995), or data-based, approach in which the categories in the collected data can be recombined and the optimal categorization is determined based upon a set of statistics provided by the Rasch analysis. In other words, the Rasch analysis is a statistical method that can be used to ascertain and verify respondents' perceptions of the ordering of category meanings (Lopez, 1996).

Technically, the Rasch analysis starts by combining adjacent categories in a "collapsing" process, in which new categories are constructed. By comparing related statistical indexes, the optimal categorization can then be determined. Three sets of statistics or parameter estimates, including model-data fit statistics (Lopez, 1996; Wright & Masters, 1982), category statistics and parameter estimates (Linacre, 1995; Andrich, 1996a, 1996b), and separation statistics (Lopez, 1996; Wright & Masters, 1982), are provided by the Rasch analysis and can be used for determining the optimal categorization. An optimal categorization, according to the Rasch analysis, should be the one that fits the Rasch model[2], has ordered categories (i.e., numerical values generated from the categories must reflect the increasing or decreasing trait to be measured), and leads to a greater discrimination among items and respondents.

In applying the Rasch analysis, as with other measurement models, the model-data fit should be examined first. Two commonly used fit statistics are Infit and Outfit Mean-Square, or simply Infit and Outfit, statistics (Linacre, 1994; Wright & Master, 1982). The Infit statistic denotes the information-weighted mean-square residual difference between observed and expected responses. The Outfit statistic, which is more sensitive to outliers, denotes the usual unweighted mean-square residual. Infit and Outfit, with a value of 1, are considered satisfactory model-data fit, while greater values (e.g., >1.3)[3] or smaller values (e.g., <0.7) are considered misfit. A

greater value often indicates an inconsistent performance, while a smaller value reflects too little variation. Lopez (1996) proposed that an optimal categorization should produce the best fit of data to the model.

If the model and data fit, category statistics and parameter estimates provided by the Rasch analysis, can then be examined. Two such indexes, average measure and threshold, have been proposed. The average measure (Linacre, 1995) is approximately the average ability of the respondents observed in a particular category, averaged across all occurrences of the category. The threshold is the location parameter of the boundary on the continuum between category k and category k-1 of a scale (Andrich, 1978, 1996b; Wright & Masters, 1982). An optimal categorization, according to average measures and threshold estimates, should be ordered (Andrich, 1996b; Linacre, 1995) -- the basic property of the categorization in an ordered-response scale, as described earlier.

Finally, if the average measures and parameter estimates are ordered, the optimal categorization can be determined by selecting the one with the largest separation or discrimination. Two separation statistics provided by the Rasch analysis are item and person separations. Conceptually, the item separation is a measure used to describe how well the scale separates testing items, while the person separation is a measure used to describe how well the scale identifies individual differences (Wright & Master, 1982). Technically, a separation statistic is the ratio of the root mean square standard error for all non-extreme measures to the standard deviation of the non-extreme estimates after removing any measurement error (Linacre, 1994). The greater the separation, the better the categorization because the items will be better separated and the respondents' differences will be better distinguished.

Utilizing the collapsing process and these statistics and parameter estimates, a new and useful post-hoc procedure based upon the Rasch analysis can be proposed to determine the optimal categorization empirically:

(a) Combine adjacent categories in a "collapsing" process, in which new categorizations are constructed;
(b) Select an appropriate Rasch model, apply the Rasch calibra tions, and examine the model-data fit;
(c) If the model-data fit is satisfactory, identify the "candidates" of the optimal categorization whose categories are ordered
(d) Determine the optimal categorization by selecting it from the "candidate"categorizations exhibiting the greatest separation.

Although some of these Rasch statistics had been proposed in deter-

mining the optimal categorization (e.g., Lopez, 1996), they have not been examined and evaluated simultaneously, nor have their performances been compared with conventional statistics. The purpose of this study, therefore, was to determine the optimal categorization of a self-efficacy rating scale using the post-hoc Rasch analysis, and to compare the performance of the Rasch statistics and parameter estimates with conventional statistics.

METHOD

Scale

A 50-item scale to measure psychomotor self-efficacy was developed by Updyke (1992). There are five sub-scales in the scale including sit-ups, running, glide-pull-ups, glide-presses, and pull-ups. An overall question is asked in each sub-scale. For example, "How many sit-ups can you do in one minute?" is asked in the sit-ups sub-scale. Ten items, ranging from easy to difficult (e.g., 20 sit-ups, 25 sit-ups, ..., 65 sit-ups), in each sub-scale are then presented to examinees. The examinees are asked to indicate their confidence in completing the items, using the following response choices: "I know I can," "I think I can," "I am not sure," "I don't think I can," and "I know I can't" (see Table 1). These response categories are coded using numerical values "5," "4," "3," "2," and "1,"respectively.

Participants and Data Collection

The scale was administered to a total of 2,022 children from 15 Midwestern schools in the Fall of 1991. The children included 1,009 boys and 1,013 girls, ranging in age from nine to 13 years of age. One week prior to the self-efficacy tests, a psychomotor-testing battery, including sit-ups, running, glide-pull-ups[4], glide-presses, and pull-ups, was administered to the children so that they had a basic understanding of their own abilities in these tasks.

Data Analysis

Category collapsing. The data analysis started by recombining the original five adjacent categories into two, three, and four categories. For two-category combinations, there were four collapsings, i.e., 11112, 11122, 11222, and 12222. The expression "11112" meant that original category

Table 1
The Sit-ups Subscale of the Self-efficacy Scale

How many sit-ups can you do in 1 minute?

	I know I can	I think I can	I am not sure	I don't think I can	I know I can't
20 sit-ups	☐	☐	☐	☐	☐
25 sit-ups	☐	☐	☐	☐	☐
30 sit-ups	☐	☐	☐	☐	☐
35 sit-ups	☐	☐	☐	☐	☐
40 sit-ups	☐	☐	☐	☐	☐
45 sit-ups	☐	☐	☐	☐	☐
50 sit-ups	☐	☐	☐	☐	☐
55 sit-ups	☐	☐	☐	☐	☐
60 sit-ups	☐	☐	☐	☐	☐
65 sit-ups	☐	☐	☐	☐	☐

"1" was retained as "1," but original categories "2, "3," and "4" were collapsed into category "1," and category "5" collapsed into category "2." For three-category combinations, there were six collapsings, i.e., 11123, 11233, 11223, 12223, 12233, and 12333; and for four-category combinations, there were four collapsings, i.e., 11234, 12234, 12334, and 12344. Thus, including the originally designed category (i.e., 12345), there were 15 collected and derived data sets.

Rasch analyses. Each of these data sets was analyzed individually using the Rasch rating scale model (Andrich, 1978, 1996b; Wright & Master, 1982) through the implementation of the FACETS computer program (Linacre, 1994). The Rasch rating scale model was defined as follows:

$$\pi_{nijx} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i - \tau_j)]}$$

where β_n denotes examinee n's ability, δ_i denotes item i's difficulty, *exp* denotes an exponent of the natural constant $e = 2.71828$, and τ_j denotes threshold. The threshold, based upon which step difficulties can be determined, is a relative location value to the item difficulty.

It should be pointed out that the Rasch partial credit model (Master, 1982), which is another commonly used Rasch model, can also be used for analyzing outcomes recorded in more than two ordered response categories. The reason for selecting the rating scale rather than the partial credit model is that the same set of ordered response categories (from "I know I can" to "I know I can't") was employed in the psychomotor self-efficacy scale, and therefore, the relative difficulties of the steps within each item would not vary greatly from item to item. The rating scale model, therefore, would be more appropriate theoretically. Interested readers are referred to Wright and Masters (1982) for more information about the similarities and differences between these two models.

Two facets -- Examinee and Item -- were defined in the analyses, and the examinee facet was set as "non-center." The facet examinee was positively measured (i.e., the higher the score, the larger the logit values). The convergence was set at 0.5 and 0.01 and maximal iterations were set at 200.

Rasch statistics and parameters. The model-data fit statistics (including Infit and Outfit), category statistics (including average measures) and parameters (including threshold estimates), and separation statistics (including item and person separations) were computed and provided by the FACETS computer program.

Conventional statistics. Besides the Rasch statistics and parameter estimates, coefficient alphas and point-biserial correlation coefficients were also computed for each of these data sets. FACETS provided both person (correlation between a person's responses on various items and total scores of the items) and item (correlation between persons' responses on an item and their total scores) point- biserial correlation coefficients. Because the distributions of these coefficients were skewed, the medians were used as summary statistics. Together, these statistics and parameter estimates were used to determine the optimal categorization and to compare them with each other.

RESULTS

Rasch Statistics and Parameter Estimates

Overall, the model fit the data well and the means of Infit and Outfit were all close to one, except for the two-category collapsings. The means of Outfit of these collapsings were at or above 1.3. The means of Infit and Outfit statistics were illustrated in Figure 1, where "1" denotes the collapsing of "11112," 2 for "11122," ..., and 15 for "12345."

Category statistics4 and parameter estimates are summarized in Table 2. Across collapsings, all of the average measures were ordered. Thus, no optimal categorization can be determined based upon this statistic, indicating that it was not very sensitive to the change in the categorizations. In contrast, the results of threshold estimates, which are also summarized in Table 2, indicated that some of these collapsings, including the original intended category (i.e., "12345"), were not ordered. In fact, only three of the three-category collapsings, i.e., "11223," "12223," and "12233," were ordered in threshold estimates, which made them candidates for the optimal categorization.

Finally, separation statistics are illustrated in Figure 2. Overall, there was a trend that, as the number of categories increased, both item and person separations increased. This was not surprising since the more categories, the finer the measure and, therefore, the better the discrimination and separation. The difference between the separations of the original five-

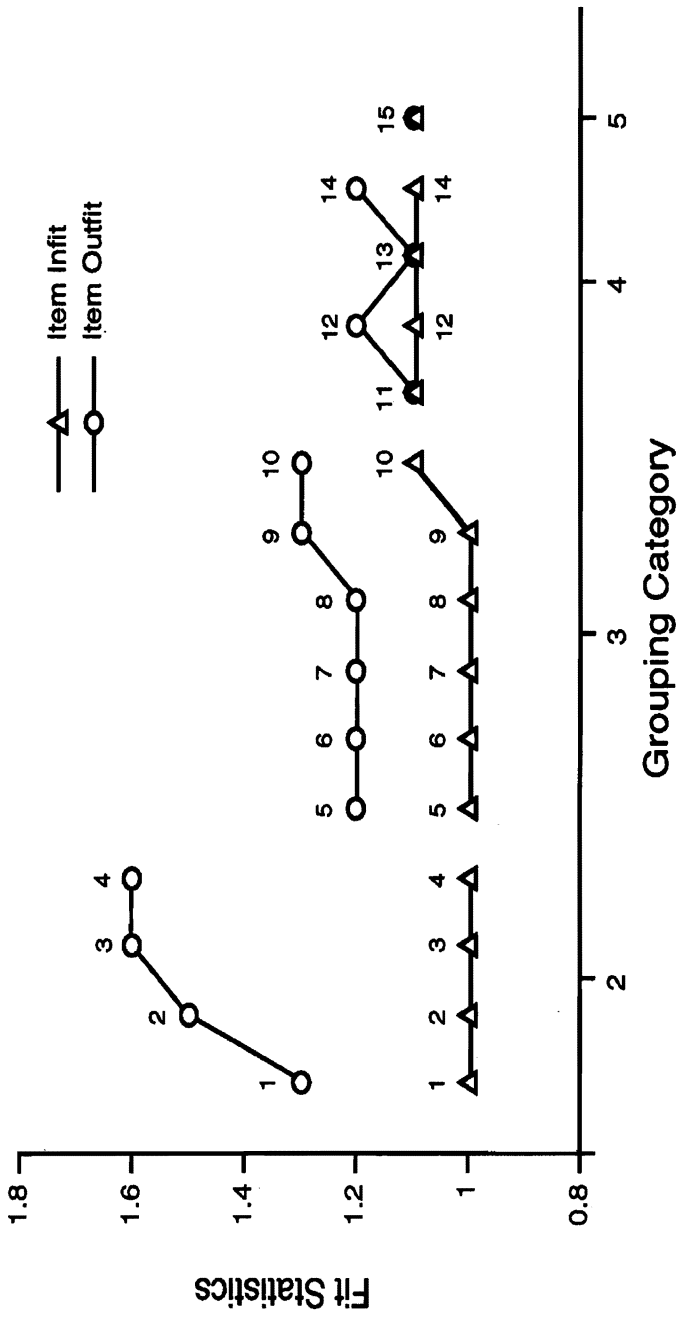


FIGURE 1. Summary of mean Rasch item fit statistics (where1=collapsing "11112," 2= "11122," 3= "11222," 4= "12222," 5= "11123," 6= "11233," 7= "11223," 8= "12223," 9= "12233," 10= "12333," 11= "11234," 12= "12234," 13= "12334," 14= "12344," 5=original five-category "12345"). The "7," "8," and "9" collapsings are the ones with ordered thresholds.

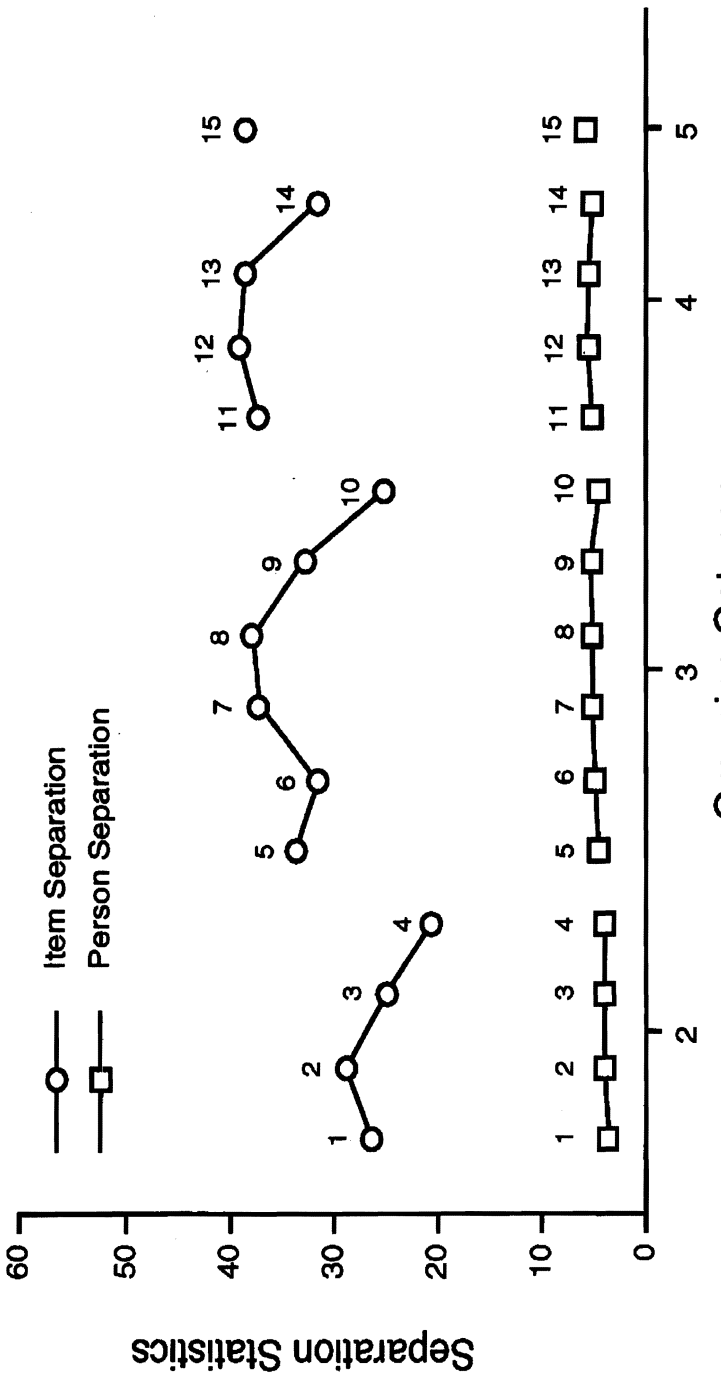


FIGURE 2. Summary of Rasch separation statistics (Note: As the notations used in Figure 1, 1= "11112," ... 15=. "12345"). The "7," "8," and "9" collapsings are the ones with ordered thresholds.

Table 2
Summary of Category Frequency and Rasch Statistics and Parameters

Reference Code	Collapsing	Category Frequency (%)					Average Measure					Threshold Estimate			
		1	2	3	4	5	1	2	3	4	5	1	2	3	4
1	11112	67	33				-3.2	1.8				-			
2	11122	55	45				-2.5	2.4				-			
3	11222	44	56				-2.0	2.8				-			
4	12222	35	65				-1.6	3.2				-			
5	11123	54	13	32			-2.1	-.2	1.7			.24	-.24		
6	11233	42	13	45			-1.6	0.0	2.0			.37	-.37		
7	11223	42	26	32			-2.4	-.1	2.4			-.79	.79[a]		
8	12223	32	36	32			-2.5	0.0	2.9			-1.40	1.40		
9	12233	32	23	45			-1.9	0.1	2.7			-.55	.55		
10	12333	33	11	56			-1.3	0.2	2.2			.61	-.61		
11	11234	42	13	13	32		-1.7	-.4	0.3	1.7		.09	-.18	.09	
12	12234	32	23	13	32		-1.9	-.5	0.5	2.1		-.85	.49	.36	
13	12334	32	10	26	32		-1.7	-.6	0.4	2.3		-.01	-1.06	1.06	
14	12344	32	10	13	45		-1.3	-.2	0.4	1.9		.34	-.20	-.14	
15	12345	32	10	13	13	32	-1.4	-.5	0.0	0.5	1.7	.15	-.60	.21	.24

[a]*Shaded boxes indicate collapsing of those with ordered thresholds.*

category categorization and that of several three- and four-category ones in both item and person separations, however, were small. In comparison to the person separation, the item separation was much more sensitive to the change of the categorization. Among three "candidates for the optimal categorization," the highest separation, including both item and person separations, was the collapsing No.8, "12223," suggesting that this categorization was the optimal one.

Conventional statistics

Conventional statistics were summarized in Figure 3. Although there was a trend for the internal consistency (as representing by coefficient alphas) to increase as the number of categories increased, the overall changes were very small, indicating that this statistic was also not sensitive to the change in categorization. In contrast, person point-biserial correlation coefficients expressed in medians varied dramatically along with different collapsings. In fact, the change patterns of these coefficients were very similar to the item separation, although they alone cannot be used to determine the optimal categorization. Some variations were also found in item point-biserial correlation coefficients, which were also expressed in medians, but they were not as dramatic as the person point-biserial correlation coefficients.

DISCUSSION

The most important finding of this study is that the determination of an optimal categorization should be empirically based and that the Rasch analysis is a very useful post-hoc approach for such a purpose. While the originally intended categorization of the self-efficacy scale was a five-category construct, it was found that, in the respondents' perceptions, the order of category meanings was indeed a three-category construct. Without the Rasch analysis, however, this perceived categorization might not have been detected. Further, the differences among collapsings could be very large even when the same number of categories was employed (see, e.g., large variations among 3-category collapsings). Again, without the Rasch analysis, differences might not have been detected.

A more careful look at what happened in the originally intended (i.e., "12345") and the new optimal (i.e., "12223") categorizations may help in better understanding why the optimal categorization identified by the Rasch analysis is better than the original one. The percentages of category

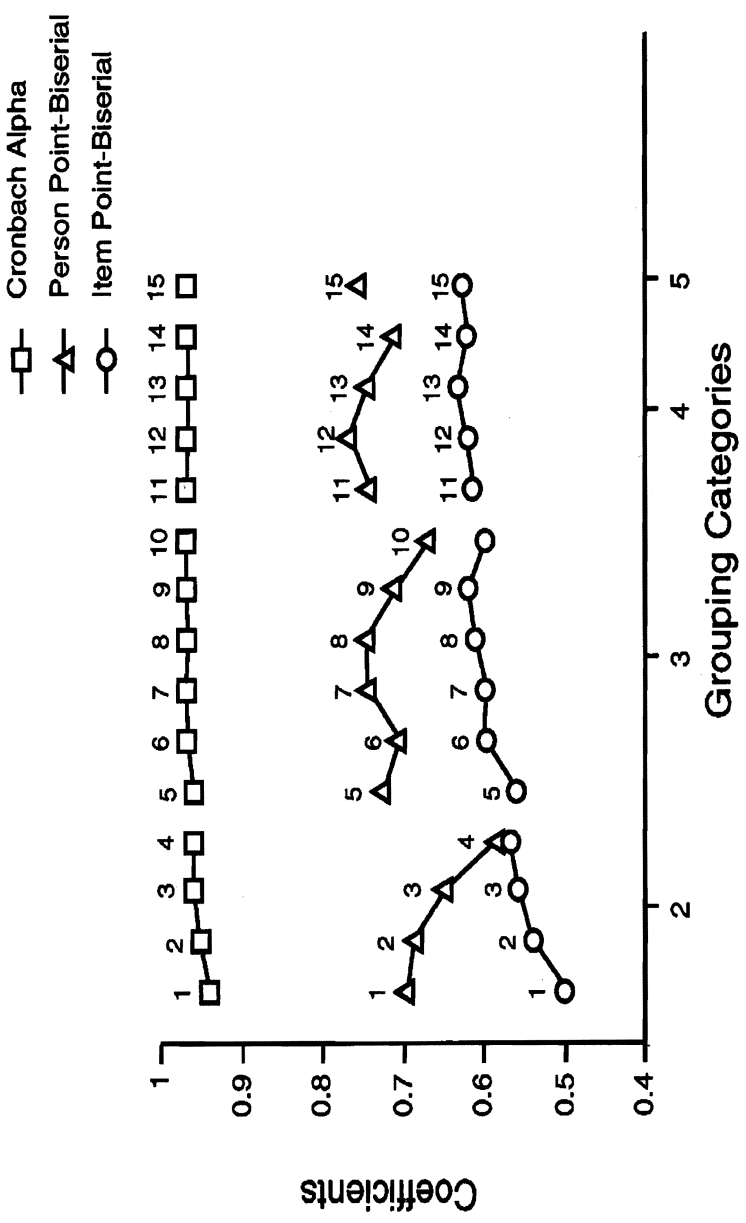


FIGURE 3. Summary of coefficients alphas and median person and item point-biserial coefficients (Note: As the notations used in Figure 1, 1= "11112," ... 15= "12345"). The "7," "8," and "9" collapsings are the ones with ordered thresholds.

frequencies for the original categories one to five were 32, 10, 13, 13, and 32, respectively (see Table 2). According to the threshold estimates (see also Table 2), there was a tendency indicating that respondents were more likely to select the category "I am not sure" (coded as 3) than "I don't think I can" (2) if they passed the category "I know I can't" (1). As a result, the threshold between "I don't think I can" (2) and "I am not sure" (3) became "easier" than that between "I know I can't" (1) and "I don't think I can" (2), and the thresholds became disordered. Although the rest of the thresholds (i.e., from "I am not sure" [3] to "I think I can" [4] to "I know I can" [5]) were ordered, there was also a trend for respondents, who passed the category "I am not sure" (3), to be more likely to rate themselves "I know I can" (5) than "I think I can" (4). There are two possible explanations for these observations.

The first one is that the words "I am not sure" are more familiar to respondents in expressing uncertainty. Studies (e.g., Bradburn & Sudman, 1979) have found that respondents preferred to use their familiar words when responding to a survey. The disordered threshold categories might represent such a familiarity preference. The second explanation is that in respondents' perceptions, the categories "I don't think I can," "I am not sure," and "I think I can" were basically the same when expressing their "uncertain" confidence with respect to their abilities to execute a physical task. The optimal categorization identified by the Rasch analysis in this study supported this explanation. After combining all three "uncertain" categories together, the percentages of category frequency became more balanced (32%, 36%, and 32% for categories one, two, and three, respectively) and the categorization became ordered as reflected in both ordered average measures and threshold estimates (see Table 2). Clearly, the empirical examinations and the rich information provided by the Rasch analysis make a much more thorough understanding of the categorization of the scale possible.

Another interest of this study is to compare and evaluate the performances of Rasch statistics and parameter estimates and conventional statistics in determining the optimal categorization. Consider the Rasch statistics and parameter estimates first. Although there were some variations among model-data fit statistics, they were relatively inefficient in identifying the optimal categorization. The values of Infit statistics, for example, were basically kept the same across all the collapsings, and all were within the "fit" range (i.e., 0.7 and 1.3; see Figure 1). Slightly more variation among collapsings was found among the Outfit statistics, but, again, the

optimal categorization is not necessarily the one with the best model-data fit statistics. Thus, the results of this study did not support Lopez's claim that the optimal categorization "produces the best fit of data to model" (Lopez, 1996, p. 482). Therefore, we concluded that, while model-data fit is a necessary condition in applying the Rasch analysis, the fit statistics themselves provide limited information in identifying the optimal categorization.

While both category statistics and parameter estimates have been found helpful in understanding the changes of the categorization, it was the information of threshold estimates that permitted the identification of the "candidates" for the optimal categorization. Linacre (1995) proposed that a criterion for category utility is that the average measures should be in the same order as the categories. Andrich (1996a), however, argued that the average measure alone is not adequate and that the threshold estimates of the categories must also be ordered. The results of this study supported Andrich's argument. Average measures of all categories across all the collapsings in this study were ordered, which provided no sensitive information on the changes in the categorization, or which categorization was optimal. In contrast, only three collapsings, according to the information of threshold estimates, were found ordinal in their categorization, making them the candidates for the optimal categorization. However, it is unknown whether a similar performance of average measures and threshold estimates would be maintained in other circumstances. Future studies, especially those using simulation data, are needed to further understand the performance of category statistics and parameter estimates in determining optimal categorization.

Since only limited variations in the person separation statistic were found across collapsings, this statistic is not useful in determining the optimal categorization. The item separation was sensitive to the changes in the categorization, although the overall trend was that the more categories, the higher the item separation. As illustrated in this study, along with threshold estimates, this statistic could be used as the final criterion in determining the optimal categorization.

Finally, little variation was found among coefficient alphas, although, again, they increased as the number of categories increased. Thus, this commonly used conventional statistic was not helpful at all in determining the optimal categorization. The person point-biserial coefficient, on the other hand, was sensitive to the changes in the categorization. In fact, its change patten (see Figure 3) was similar to the item separation (see

Figure 2). However, just as the item separation itself cannot be used to determine the optimal categorization, the information of the person point-biserial coefficient alone is also inadequate for determining the optimal categorization. At the same time, this statistic would be useful when sophisticated statistics or parameters, such as those provided by the Rasch analysis, are not available. While some variations were found in the commonly used item point-biserial coefficient, this statistic, like coefficient alpha, provides only limited information about the optimal categorization.

It should be noted that collapsing categories in this study was implemented in a somewhat mechanical way, i.e., we collapsed all possible adjacent categories even if some of the collapsings might be unnecessary. For example, considering that the original categorization was a five-category construct, there should be little chance to derive an optimal categorization from two-category collapsings. To include unnecessary collapsings would create additional computing burdens to scale developers. Therefore, a more intelligent method in which only selected collapsings are used could be employed in practice. The general principles to follow in collapsing categories selectively, recommended by Linacre and Wright (Linacre, 1995; Wright & Linacre, 1992), are: (a) Whatever collapses must make practical sense and be explainable on the same level as the variable being measured, and (b) when combining or deleting categories, the aim should be to balance the category frequencies as much as possible. The results of this study supported these principles. Not only did collapsing three "uncertain" categories make the new optimal categorization more interpretable, the category frequencies also became more balanced. On the other hand, considering today's and the future's computing power, the mechanical way to include all possible collapsings may not be a problem in terms of computing workload, as long as computer programs are equipped with related functions. In this way, no optimal categorization will escape detection because of the scale developers' initial inappropriate selections. It is expected that such functions6 will be included in future versions of Rasch analysis software.

Although an optimal categorization was successfully identified by the Rasch analysis in this study, we should acknowledge that the identified categorization is merely the result of a post-hoc analysis. Therefore, it is still unknown whether the optimal categorization identified by the Rasch analysis would be maintained in subsequent administrations. Applying similar post-hoc analysis to existing longitudinal data may help to provide answers. More importantly, studies are needed to determine whether a

modified categorization which is based upon a previous Rasch analysis maintains its optimal construct in the later measurement practices.

CONCLUSION

In conclusion, the Rasch analysis has been demonstrated to be a very useful and powerful means of determining the optimal categorization of an ordered-response scale. The quantitative information provided by the Rasch analysis clearly provides scale developers with a valuable reference to evaluate the intended categorization, to verify respondents' perceptions of the ordering of category meanings, and to guide reconstruction of the categorization if it is necessary. The practice illustrated in this study, therefore, should be employed routinely in constructing ordered-response scales. More studies to determine the stability of the performance of Rasch statistics and parameters and the identified optimal categorization are needed.

FOOTNOTES

[1]The original Rasch dichotomous model has been extended to include items with multiple response categories, such as the rating scale (Andrich, 1978) and the partial credit models (Masters, 1982).

[2]Depending on the research interests and data characteristics, other IRT models, e.g., Graded Response Model (Samejima, 1996), can also be applied, but the post-hoc analytic paradigm described in this article should still apply.

[3]The model-data fit in the Rasch analysis is often determined by some predetermined criterion values and the criterion "1.3 and 0.7" is commonly used. The criterion values, however, are arbitrarily determined. Interested readers are referred to Linacre (1994, p. 74) for more information about criterion values and their interpretations.

[4]Both glide-pull-ups and glide-presses were performed on an exercise device called "Total Gym," on which children either pulled up or pressed their body weights up and down on the trolley of the equipment.

[5]A rich set of category statistics, such as the expectation measure at category and the category peak probability, were also provided by FACETS. Due to the length constraint of this article, these statistics were not reported and discussed. Interested readers are referred to Linacre (1994) for more detail.

[6]In fact, similar functions to recode category are already available in several Rasch analysis programs (e.g., FACETS, Linacre, 1994), but they have not been completely automatic.

AUTHORS' NOTE

Part of this study was presented at the 1997 annual meeting of the American Educational Research Association, Chicago.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1995). Models for measurement, precision, and the nondichotomization of graded responses. *Psychometrika, 60,* 7-26.

Andrich, D. (1996a). Category ordering and their utility. *Rasch Measurement Transactions, 9*(4), 464-465.

Andrich, D. (1996b). Measurement criteria for choosing among models with graded responses. In A. von Eye & C. C. Clogg (Eds.) *Analysis of categorical variables in developmental research: Methods of analysis* (pp. 3-35). Orlando, FL: Academic Press.

Andrich, D. (1997). Rating scale analysis. In J. P. Keeves (Ed.) *Educational research, methodology, and measurement: An international handbook* (2nd Ed., pp. 874-880). New York: Elsevier Science.

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9,* 151-160.

Bradburn, N. M., & Sudman, S. (1979). *Improving interview method and questionnaire design.* San Francisco: Jossey-Bass.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,* 16, 297-334.

Guilford, J. P. (1954). *Psychometric methods* (2nd Ed.). New York:McGraw-Hill.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels andpositions in rating scales. *Journal of Educational Measurement, 25,* 85-96.

Linacre, J. M. (1994). *FACETS*: Rasch-Model Computer Program (Version 2.7) [Computer software and software manual]. Chicago: MESA Press.

Linacre, J. M. (1995). Categorical misfit statistics. *Rasch Measurement Transactions, 9*(3), 450-451.

Lopez, W. (1996). Communication validity and rating scales. *Rasch Measurement Transactions, 10*(1), 482-483.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika,* 47, 149-174.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of*

Experimental Psychology, 12, 496-516.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research.* [Reprinted, Chicago, IL: University of Chicago Press, 1980]

Samejima, F. (1996). Graded response model. In van der Linden, W. J., & Hambleton, R.K. (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Updyke, W. (1992). *What can I do? A psychomotor self-efficacy scale.* Unpublished manuscript.

Wedell, D. H., Parducci, A., Lane, M. (1990). Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and type of anchors. *Journal of Personality and Social Psychology, 58,* 319-329.

Wildt, A. R., & Mazis, M. (1978). Determinants of scale response: label versus position. *Journal of Marketing Research, 15,* 261-267.

Wright, B. D., & Linacre, J. M. (1992). Combining and splitting categories. *Rasch Measurement Transactions, 6*(3), 233-235.

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.* Chicago: MESA Press.

JOURNAL OF OUTCOME MEASUREMENT®, *1*(4), 305 - 328

The Sexual Experiences Survey: Interpretation and Validity

George Karabatsos
MESA Psychometric Laboratory
The University of Chicago

The Sexual Experiences Survey (Koss, Gidycz, & Wisniewski, 1987) is a commonly used instrument for assessing various degrees of sexual aggression and victimization among male offenders and female victims. Rasch analysis was used to transform qualitative raw score observations into objective linear measures using the responses of a national sample of 6,159 higher education men and women across the United States, aged 18-24. This paper supports the construct validity of the survey through evaluation of the item hierarchy, fit statistics, and separation indices. Findings confirm a "dimensional" perspective on rape, suggesting that sexually aggressive behaviors can be scaled along a single continuum from normal to extreme sexual behavior. The item hierarchy reveals an arrangement of sexually aggressive acts in an order of mild to severe, which compares with the one theorized by the authors of the SES. Identity plots demonstrate the validity of using a common set of SES item calibrations to measure both male and female respondents. For interpretation of person responses to the SES, three conclusions are suggested. First, Rasch analysis must be employed to examine item responses effectively. Second, when the survey is administered to a college sample aged 18-24, the item calibrations obtained in this paper can be used to measure offenders and victims. Third, a total raw score-to-measure conversion is not always sufficient to interpret person measures. Instead, a scalogram method needs be added to the Rasch analysis to separate the measures of offenders and victims who complete the survey. Implications for future research are discussed.

Requests for reprints should be sent to George Karabatsos, MESA Psychometric Laboratory, The University of Chicago, 5835 S. Kimbark Ave., Chicago, IL 60637-1609.

Studies have shown that for every adult rape reported to authorities, 3-10 rapes are unreported (e.g., Law Enforcement Administration [LEAA], 1975). Many female victims do not report out of fear, guilt, or shame. Spousal sexual abuse is rarely reported to authorities. Due to fear of punishment, most male offenders have no motivation to confess their crimes. Therefore, most sexual aggression research has used samples identified through the criminal justice system, treatment facilities, and crisis centers (Koss et al. 1987). This sampling bias not only limits the generalizability of empirical findings, but hinders researchers from learning about the victims and offenders who are "undetected" by the system.

To address this bias problem, Koss and Oros (1982) suggested the use of anonymous sexual aggression surveys to detect hidden victims and criminals. Kirkpatrick, Kanin, and colleagues were one of the first to do so (Kirkpatrick and Kanin, 1957; Kanin, 1957; Kanin and Parcell, 1977). They surveyed males and females regarding the incidence of sexual aggression on university campuses. The items on the surveys involved various degrees of sexual advancements, such as kissing, necking, petting above or below the waist, attempted or completed intercourse, and attempted or completed intercourse using violence. Later, Koss and Oros (1982) introduced the Sexual Experiences Survey (SES). The most recent version of the survey (Koss et al. 1987) is composed of 10 items which asks respondents whether any of four types of sexual advances (sex play, attempted intercourse, completed intercourse, sadistic sex acts) have occurred as a result of four types of force (verbal coercion, misuse of authority, victim intoxication, and physical force).

The SES is based on the idea that rape behaviors represent extreme acts on a continuum of normal sexual behavior, as some research has suggested (e.g., Weis and Borges, 1973). However, the support for a dimensional perspective of sexually aggressive behavior is weak. One reason is that research has used faulty metrics to represent the prevalence of sexual aggression, and determine how different aggressive acts relate to each other. Such metrics include the interpretation of item responses using frequencies and percentages, and the summation of each person's category numeral responses across all items of a survey (total raw score). Arguments and proofs involving the nonlinearity and ambiguity of counting raw observations date back to Thorndike (1904). A second reason is that tests of unidimensionality in sex surveys have involved the interpretation of Cronbach's Alpha and inter-item correlations (a function of Alpha values). Guttman (1977), as well as others, has argued that correlational

analysis of internal consistency does not analyze items:

> It merely attempts to "test" the -challenging!- hypothesis that all inter-item correlations are zero, and usually by an incorrect item-total score correlation technique. It is a way of trying to avoid the basic problem of definition, and involves wishful thinking that correlations should determine content (Guttman, 1977, p. 100).

Frequencies and percentages of item responses have also been used to test unidimensionality. However, tabulating percentages is inadequate to address reliability. Although this simple method has some descriptive utility, it does not determine whether all items of a survey measure the same variable.

The current scoring procedure of the SES is illustrated in Figure 1. The person's score (1-5) depends on the most severe event (or endorsed item) that has occurred. The higher the score, the more severe the sexual act. Respondents who report no victimization are scored 1. For reasons described in the results section of this paper, the two items involving misuse of authority were excluded from the Figure.

Questions arise regarding the validity of this scoring method. Do the acts grouped in each of the five different scores actually separate 1 unit from each other? Are *Sex Play By Force* and *Sex play By Verbal Coercion*, with the same score of 2, equally severe? *Is Attempted Intercourse By Force* (score = 4) really twice as severe as *Sex Play By Force* (score = 2)? This method interprets people's scores on an ordinal scale, which is inconsistent with parametric data analysis.

Another issue surfaces in the hierarchical nature of the scoring procedure. Is the order of items correct? The developers of the survey theorized that all acts involving sex play (2) are less severe than intercourse due to nonviolent and non-intoxicating force (3), which are less severe than attempted intercourse by violence or intoxication (4), and the most severe events all involve completed intercourse through violence or intoxication (5). To determine the exact order of how items arrange on the dimension from less to more severe acts, a thorough quantitative investigation is needed.

In this paper, I take the position that Rasch analysis will be useful to interpret responses of the SES. More specifically, it will:

1. Determine whether all items of the SES contribute to define a single variable, which we will call "rape severity,"
2. Describe how effective the items are in defining this variable,
3. Determine how the items arrange on the variable continuum, and

RAW SCORE	MAP OF ITEMS
5	Intercourse by force Intercourse by intoxication Sex acts by force
4	Attempted intercourse by force Attempted intercourse by intoxication
3	Sexual intercourse by verbal coercion
2	Sex play by verbal coercion Sex play by force
1	Never victimized

FIGURE 1 The current scoring key of the Sexual Experiences Survey.

4. Provide a method to interpret responses to the SES on a *linearized, interval scale.*

The fourth point is important. In order to accurately represent variables and people, objective measurement needs to be made a high priority in sexual aggression research. Raw scores are often subjected to parametric data analysis, even though they are not linear. Raw scores need to be transformed into interval units of measurement to meet the linear data requirement of parametric statistics. Measurement is an important aspect of science, and when its bases are ignored, the validity of the research comes into question:

> Measurement is one crucial hierarchical step in the representation of whatever it is you are talking about and must be seen simultaneously in the light of conceptualization, observation and assignation of quantity as well as relations to other representations. If this was a must in the physical sciences, then it is certainly so in the social sciences. Because the total system has components of a very high degree of uncertainty, apparent accuracy can be misleading, and subsequent

> hypothesis testing irrelevant. Much work in psychology...reveals a spurious search for precise measurement which has manifestly lost contact with the reality claimed to be under study. By and large, researchers in such fields study their own disciplinary, or even sub-disciplinary, output and therefore make little practical contribution to society (Kinston, 1985, p. 102).

Rasch analysis will not only serve as a technical solution for the SES. The item positions on the variable will facilitate theoretical explanations regarding the nature of sexual aggression, as perceived by both male offenders and female victims. In addition, the item hierarchy theorized by the authors (Figure 1) will be compared with the hierarchy that Rasch analysis determines. This paper will also illustrate that the interpretation of SES person measures is not straightforward. Additional steps are needed to separate the people who have been involved with sexually aggressive events of differing severities.

METHODS

Participants

The intent of data collection (1985) was to gather a large sample representative of students attending higher education institutions across the United States. Two steps were taken to reach this goal. First, an attempt was made to recruit higher education institutions which represent every region and school location across the United States. Second, within each recruited institution, classes were randomly selected to recruit subjects.

Ninety-three schools were targeted, of which 19 agreed to participate. An additional 13 institutions were recruited among 60 potential replacements. The 32 institutions which participated in the study represent Alaska, Hawaii, New England, Mideast, Great Lakes, Plain States, Southeast, and Rocky Mountains. There is some overrepresentation in the Northeast and Southwest and underrepresentation in the West. Within each region, location is representative in regards to whether an institution belonged inside or outside a standard metropolitan statistical area (SMSA) (i.e., SMSA > 1,000,000 people; SMSA < 1,000,000 people; outside an SMSA), the minority school enrollment of the institution (above or below the national mean percentage of minority enrollment), the type of power which governs the institution (private secular, private religious, or public sector), the institutional type (university, four-year college, two-year college, and technical/vocational), and the total student enrollment of the in-

stitutions (1,000-2,499 students; 2,500-9,999; more than 10,000 students).

Within each randomly selected classroom, students were asked to complete a large self-report survey titled "National Survey of Inter-Gender Relationships" (about 330 questions, including the SES). Participants were assured of their anonymity, and if they agreed to participate, they were asked to sign a consent form. 98.5% agreed to participate in the study, and the responses of 6,159 students were collected (3,187 women and 2,972 men). This sample is diverse in respect to ethnicity, income, and marital status. At the time of data collection, a recent finding of the Bureau of the Census (1980) concluded that 26% of all people age 18-24 in the United States attended institutions of higher education. This is the group the sample represents.

All of the information contained in Table 1 was obtained from Koss et al. (1987, p. 163-165), which is a good source for readers who need more demographic information, or are interested in learning the finer details of the subject recruitment process.

Survey

The SES asks respondents whether they have been involved in various sexual offenses since the age of 14. There are two forms of the survey, one for men and one for women. They are both provided in the Appendix. Abbreviated versions of the questions are listed below.

1. Sex play by verbal coercion
2. Attempted intercourse by misuse of authority
3. Sex play by force
4. Attempted intercourse by force
5. Attempted intercourse by intoxication
6. Intercourse by verbal coercion
7. Intercourse by misuse of authority
8. Intercourse by intoxication
9. Intercourse by force
10. Other sex acts by force

Men are asked to indicate whether they have committed these 10 offenses, while women were asked to indicate whether they were victims to these crimes. Therefore, both male and female responses represent male offenses to female victims.

The SES has two response formats. For each of the 10 items, respondents are first asked to answer "Yes" or "No." If the answer is "Yes," then

Table 1

Descriptive Statistics of the Sample of Students From the 32 Higher Education Institutions

	Males	Females
N	2,972	3.187
Mean Age	21.0	21.4
Marital Status		
Single	91%	85%
Married	9	11
Divorced	1	4
Race		
White	86%	86%
Black	6	7
Hispanic	3	3
Asian	4	3
Native American	1	1
Family Income (1985)		
$0-15,000	12.6%	13.1%
$15,001-25,000	16.4	17.2
$25,001-35,000	21.2	22.5
$35,001-50,000	22.8	23.0
over $50,000	24.9	21.3
Missing	2.1	2.9

Note. These numbers were inferred from Koss et al. (1987, p. 164-165).

respondents are asked to answer the number of times the offense has occurred (1 = Once, 2 = Twice, 3 = Three Times, 4 = Four Times, 5 = Five or more times).

Fifty-four percent of women claimed to be victimized, while only 25% of men confessed to sexually aggressive behavior (Koss et al. 1987). This is not surprising, since it is easier for a person to confess as a victim than as an offender.

Rasch Analysis

The Rasch model is an application of additive conjoint measurement (Brogden, 1977, 633), a requirement for fundamental measurement (Luce and Tukey, 1964). The dichotomous response model (Wright and Stone, 1979) specifies through log-odds that the probability of person n's response to item i is governed by the measure of the subject (B) and the difficulty of the item (D):

$$\log [P_{ni1} / P_{ni0}] = B_n - D_i$$

where,

P_{ni1} = probability of an endorsed response ("Yes" response to an item of the SES),

P_{ni0} = probability of a non-endorsed response ("No" response to an item of the SES),

B_n = trait parameter (or measure) of person n, and

D_i = difficulty of endorsing item i.

D describes each item's location on the variable line of which they define (item calibrations/measures). B indicates a person's position on that line with respect to the items' locations (person calibrations/measures). When $B_n > D_i$, there is more than a 50% chance of a "Yes" response. When $B_n = D_i$, the chances for a "Yes" response is 50%. When $B_n < D_i$, the probability is less than 50%. In situations where rating scale responses are analyzed, $(- F_j)$ is added to the model (as in: $\log [P_{nij} / P_{nij-1}] = B_n - D_i - F_j$). F_j represents the difficulty of the step from rating scale category j-1 to category j (Wright & Masters, 1982).

Each facet (B, D, F) in the model are separate parameters. The effects of one parameter are free from the effects of the others (Rasch, 1960; Wright and Stone, 1979; Wright and Masters, 1982). This mathematical

property enables "test-free" and "person-free" measurement to occur, two things which Thurstone (1926;1928) required for objectivity. "Test-free" means that person scores do not depend on which items are used to measure them. "Person-free" means that item estimates do not depend on which sample is being measured.

Infit and outfit mean square statistics (MNSQ) determines how well each item contributes to define the rape severity variable. An item which a MNSQ near 0 indicates that the sample is responding to it in an overly predictable and deterministic fashion (little evidence of stochasticity). Item MNSQ values of about 1 are ideal by Rasch specifications, since it indicates local independence. In this case, most items which are easy for a given respondent to endorse ($B_n > D_i$) are answered "Yes," most items which are too difficult for a given respondent endorse ($B_n < D_i$) are answered "No," and items which lies close to a person's ability level ($B_n \approx D_i$) have a combination of "Yes" and "No" responses. A MNSQ value greater than one indicates that the sample's responses to that item are unpredictable. This questions whether the item fits the unidimensional construct. Items and persons with MNSQ values of 1.3 or above will be diagnosed as potential misfits to Rasch model specifications. Two MNSQ statistics are used to detect item misfit. INFIT identifies unexpected responses of items close to the respondents' measure levels. OUTFIT detects unexpected responses to items which are distant from the people's measure levels (i.e., surprising responses to items which are very easy or very difficult for a given person to endorse).

Rasch analysis provides separation indices which indicate the extent to which items and persons identify a useful variable line (Wright and Stone, 1979). The person separation (PSEP) is a standard deviation ratio which describes the number of performance levels the test measures in a particular sample. It equals the square root of the ratio true (unbiased) variance of person measures divided by the error variance due to person measurement imprecision:

$$\text{PSEP} = (\text{True Variance}_N / \text{Error Variance}_N)^{1/2}$$

The item separation (ISEP) index indicates how well items spread along the variable line. It is the square root of the ratio true variance of item measures divided by their error variance:

$$\text{ISEP} = (\text{True Variance}_I / \text{Error Variance}_I)^{1/2}$$

The relationship between reliability and separation is:

$$REL = SEP^2 / (1 + SEP)^2$$

All Rasch analyses were executed using the computer program *BIGSTEPS*, version 2.71 (Linacre and Wright, 1997).

RESULTS AND DISCUSSION

Two initial adjustments needed to be made. First, the two *Misuse of Authority* items (2 and 7) were infrequently endorsed by the sample. Therefore, their positions on the variable line, determined by their item difficulty estimates, were illogical with respect to the other eight items. This led to their removal from further analysis. Second, rating scale analysis of the polytomous response format of the SES showed that the sample only discriminates two response categories, similar to the Yes-No format. There were not enough people in the sample who indicated multiple occurrences of sexual acts to provide the variance necessary for multiple response category discrimination. Thus, all subsequent analyses used the Yes-No response format for the eight items of the SES.

There were 681 males and 1,612 females measurable for analysis (total n = 2,293). A large proportion of males (n = 2,291) and females (n = 1,575) were deleted from the analysis since they either reported no rape history (2,287 males, 1,562 females) or indicated every sexual event (4 males, 13 females).

Identity Plot of Item Measures

Analyses were performed separately for men and women to determine whether item calibrations across these two groups were invariant. The item plot in Figure 2 compares the item calibrations of the male and female samples. It is evident that the estimates are close enough to an identity line to conclude that they are similar across the two groups. However, the small differences do facilitate some plausible hypotheses.

Sexual acts which involve verbal coercion are more likely to be indicated by males than females. This small disparity may be due to gender differences in perception. Men who obtain sex by verbal coercion may realize that they talked unwilling partners into sex. On the other hand, women may perceive this situation as giving consent to sexual contact,

even though they didn't want to engage in something they were pressured into. Because they consented, they could have rationalized that they were willing.

The two items which indicate sexual contact by victim intoxication are also more likely to be indicated by males than females. This suggests that women are not always aware that they were given alcohol to be taken advantage of. The men, since they are doing the offending, realize their motives, and therefore will indicate the occurrence of these events more frequently.

Females are more likely than males to indicate sexual acts which involve force by the offender (4 items). This is not surprising. It is easier for a victim than an offender to admit the occurrence of violent forms of sexual aggression, being that these acts are obviously criminal. Therefore, the veracity of male subjects on force items can be questioned on the grounds that they feared punishment.

Identity Plot of Person Measures

In consideration of the small disparities in item estimates found between the two groups, is there a difference between using male or female item calibrations for measurement? Figure 3 shows that, whether one uses the male or the female "ruler," there is no difference in womens' measures. The same graphical pattern appeared when males were measured. Therefore, the decision was made to combine both samples for further analysis.

SES Item Hierarchy

The combined sample was analyzed to produce the item hierarchy in Figure 4. Items are arranged in less severe to more severe acts. Six of the eight items fit to define a unidimensional variable according to Rasch specifications (INFIT/OUTFIT MNSQ < 1.3). The two items (1,4) which misfit (OUTFIT MNSQ > 1.3) were retained for theoretical considerations, since their positions in the hierarchy are conceptually valid. Figure 4 supports the construct validity of the SES.

Beginning the mild end of the hierarchy is *Sex Play By Verbal Coercion* (-3.34) followed by *Intercourse By Verbal Coercion* (-1.31) after a rise of 2 logits in severity. These two acts are the least severe of the eight because sex by verbal coercion involves some consent by the victim. Once the victim has consented, the offender does not need force to obtain sexual

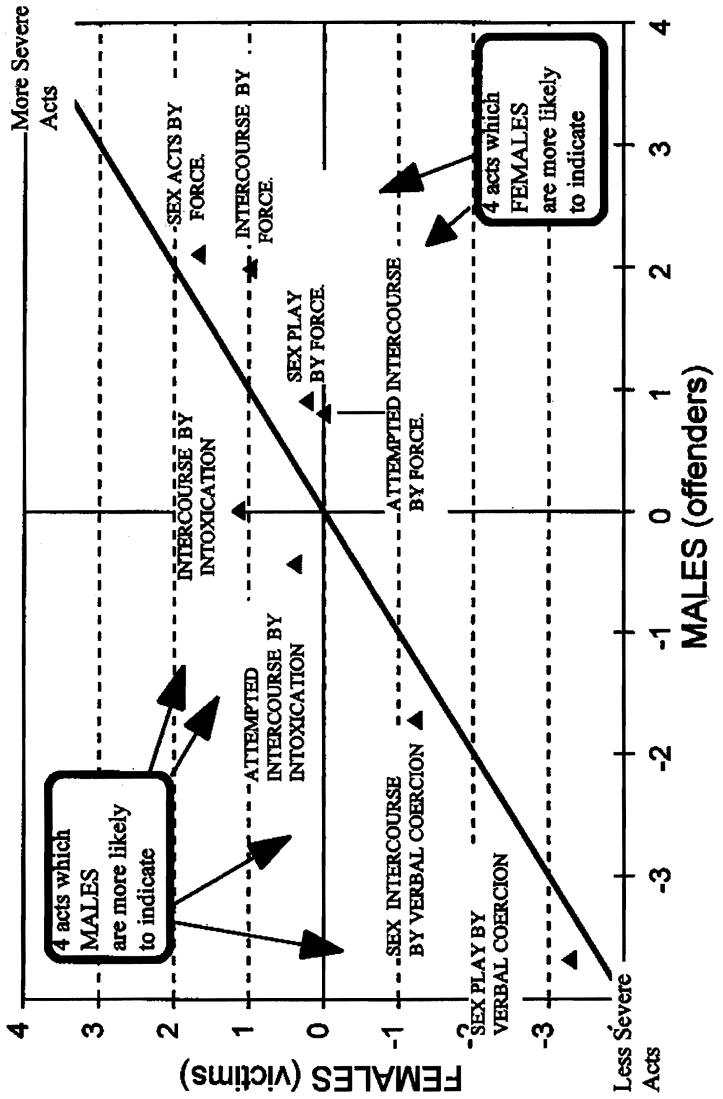


FIGURE 2. An item identity plot comparing the SES item calibrations of the male and female samples.

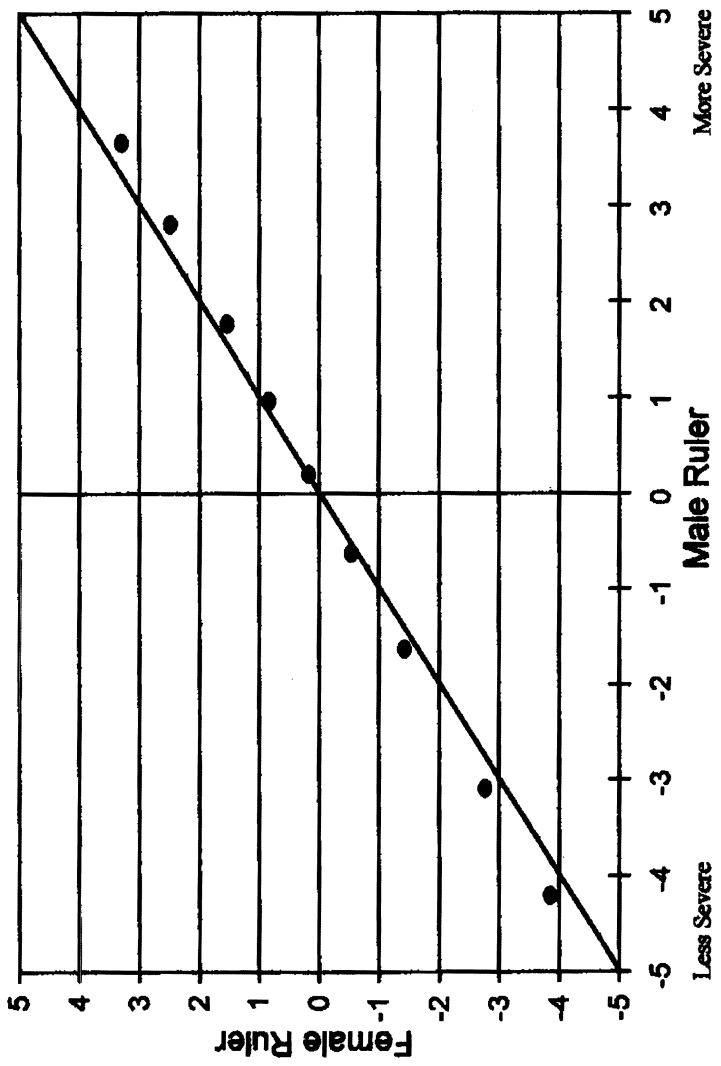


FIGURE 3. Females' rape severity measures using male and female SES item calibrations. The same graphical pattern appeared when males were measured.

contact. The other six items imply less probability of victim consent and higher probability of offender force and victim resistance.

Rising another 1.5 logits in severity are two items which indicate *Attempted Intercourse By Force* (.20) and *Attempted Intercourse By Intoxication* (.22). These items are similar. Both involve acts in which the man forces the woman against her will. *Sex Play by Force* (.40) is slightly more severe than these two items, perhaps because the offenders did achieve some sexual contact.

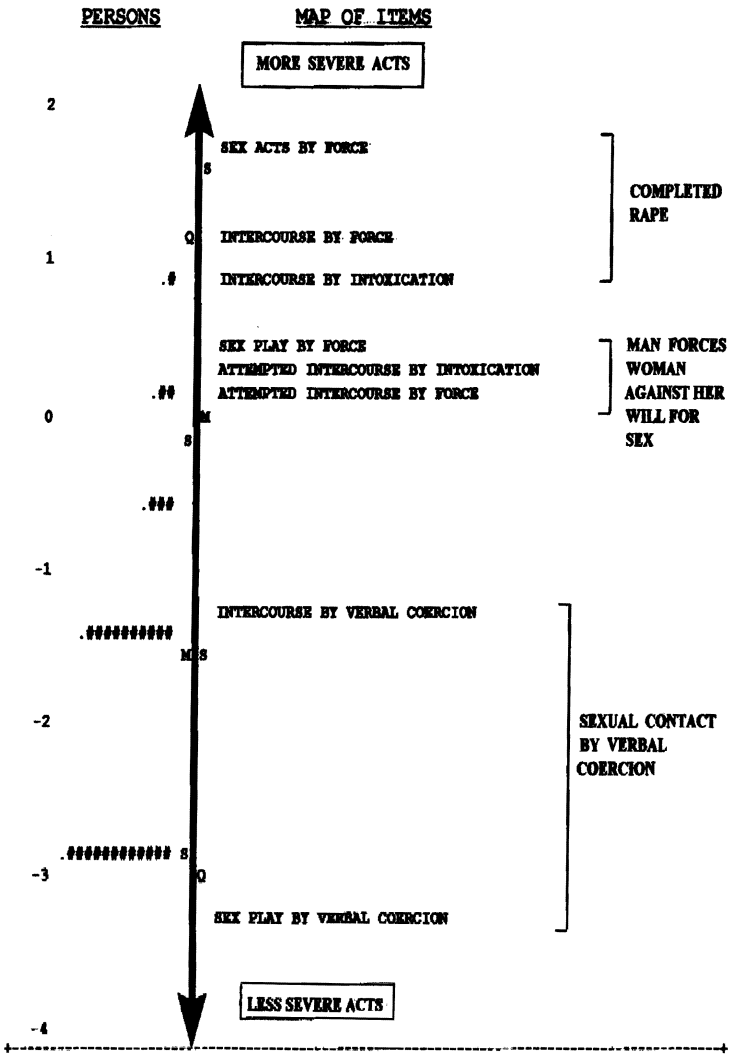
Increasing half a logit further in rape severity, we find three acts in which rape is completed by force or victim intoxication. *Intercourse By Victim Intoxication* (.89) is the least severe of these three. The decrease of inhibition may play a role here. Perhaps the consumption of alcohol or drugs allows the perpetrator to become more aggressive. Also, there is the probability that the victim will consent to sex increases, as her judgment becomes impaired. Intoxication slows and weakens the physical movements of the victim. This decreases the effectiveness of her resistance against the opportunistic perpetrator. As we move up the severity hierarchy, we find two acts involving violence: *Intercourse By Force* (1.17) *and Other Sex Acts By Force* (1.77). *Other Sex Acts By Force* represents the most severe form of rape, as sadistic behaviors are involved (i.e., oral, anal, or object penetration).

Comparison of Item Hierarchies

Now we compare the item hierarchy theorized by the authors of the SES (Figure 1) with the hierarchy determined by Rasch analysis (Figure 4). Figure 5 shows this. For ease of comparison, Rasch item calibrations were rescaled to fit the numerical range (2-5) of the original scoring method. The *No Victimization* statement found in the bottom of Figure 1 was removed, since this is not an item.

In regards to item order, the largest disparity exists for *Sex Play By Force*. The authors' theorized that this item should be considered no more severe than *Sex Play By Verbal Coercion*, since they involve the same degree of sexual advancement. Instead, Rasch analysis clusters this item with other acts involving *Threats/Force*. This supports the idea that the type of force, not degree of sexual advancement, is the more influential determinant in the severity of offenses.

In regards to the measure values of the items, Rasch analysis pushed several items close to each other. From the least severe act (*Sex Play By*



ENTRY NUM	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	ITEMS
28	174	2293	1.77	.09	1.00	.0	.81	-1.5	.28	SEX ACTS BY FORCE
25	266	2293	1.17	.07	.76	-5.3	.82	-5.6	.48	INTERCOURSE BY FORCE
22	320	2293	.89	.07	1.02	.4	.89	-1.3	.27	INTERCOURSE BY INTOXICATION
7	434	2293	.40	.06	.88	-3.3	.87	-1.9	.36	SEX PLAY BY FORCE
13	482	2293	.22	.06	1.21	5.6	1.37	5.4	.10	ATT INTERCOURSE BY INTOX
10	488	2293	.20	.06	1.00	.0	1.10	1.5	.26	ATT INTERCOURSE BY FORCE
16	1025	2293	-1.31	.05	.95	-2.5	.87	-.8	.17	INTERC BY VERBAL COERCION
1	1838	2293	-3.34	.06	1.09	3.2	1.68	6.3	-.07	SEX PLAY BY VERBAL COERCION
MEAN	628.	2293.	.00	.07	.99	-.2	1.03	.3		
S.D.	516.	0.	1.52	.01	.12	3.3	.33	3.7		

FIGURE 4. Item map depicting the sexual aggression hierarchy (analyzed: 2,293 men and women).

Verbal Coercion), there was a significant jump to *Intercourse By Verbal Coercion*, followed by a significant jump to the remaining six items, which then all cluster together. This is a different spacing that was originally theorized. It was presumed that items arrange in four different clusters (Figure 1), equally spaced 1 raw score unit apart.

Interpreting Person Measures

It has been acknowledged that rape behaviors are not a series of interdependent events in which less severe offenses always precede greater ones (Koss & Gidycz, 1985). The high positive and negative fit statistics support this position (located at the bottom of Figure 4), as the sample generally exhibited disjointed strings of responses. Irregular response patterns were typical. Disorders existed in person measurement.

For example, compare two women, each victimized once. Assume that a "Yes" response is 1, a "No" response is 0, and that the eight item response string (e.g., 11001100) arranges items in least difficult-to-most difficult order. Woman A was verbally coerced into sex play (10000000), while Woman B was physically forced into sadistic sex acts (00000001). It is evident that Woman B had a more severe experience than A, and therefore B should have a higher rape severity measure. However, since both persons have the same raw score total of 1, a typical total raw score-to-measure conversion will report that A and B have the same measure of -2.80. Of course, person misfit statistics will indicate their differences in response patterns. But how can we differentiate between A and B in terms of measures?

A quick and simple way this can be achieved is to use the KEYFORM in Figure 6, which displays the most probable measure of a person for a given item response (Karabatsos, 1997). With this device, we refer to the most severe item Women A and B endorse, and obtain their expected measures that way. Now the differences are evident. Woman A has a measure of about -3, while Woman B has a measure of about +2. Woman B, with the more severe experience, receives the higher measure as we wish.

A scalogram method can be used to provide this kind of relevance in person measurement. The method involves identifying zero responses to items which are less severe than the most extreme endorsed response, and reassigning them as missing. Zeroes after the most extreme endorsed response are kept. Therefore, a response string of 01000010 is edited to -1-----10, where "-" represents a missing response. The string 10101100 is

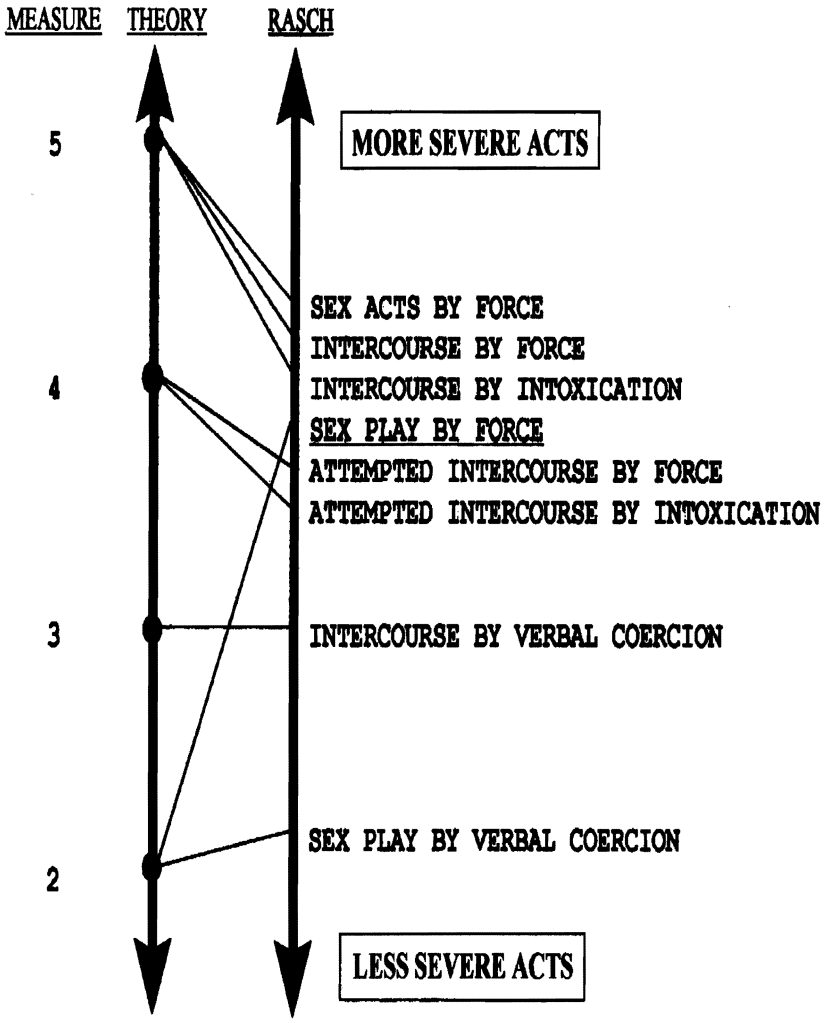


FIGURE 5. A comparison of the theorized item hierarchy versus the hierarchy determined by Rasch analysis. For ease of interpretation, the Rasch item measures were rescaled (2-5) to fit the numberical range the original scoring method (Figure 1).

edited to 1-1-1100, 00000001 would be changed to -------1, and so forth.

Using the item anchors calibrated from the unedited data matrix (i.e., the items' measure values in Figure 4), the edited data were analyzed to estimate new person measures. Since Rasch analysis employs a probabilistic response model using estimated item and person parameters, missing data is not a problem in person measurement. The validity of the scalogram method is presented in Table 2 in measurement order. It contains all possible total raw scores of 0,1, and 2. Respondents are identified by their unedited response strings. If we compare Women A and B again, we can see in Table 2 that Woman A (10000000) has a measure of -2.80, while Woman B (00000001) has a measure of 1.77. From observing the Table in its entirety, it can be inferred that a person's measure not only depends on the total raw score, but also on the severity of the endorsed items.

The scalogram method substantially increased the separation of person measures in the college sample. Using the unedited data, person separation was .40 (rel. = .14) and item separation was 22.59 (rel. = 1). The low person separation suggests that the survey did not target the sample well (this can also be inferred from Figure 4). After the scalogram method was applied, person separation rose to 1.14 (rel. = .56), while item separation remained approximately the same (ISEP = 21.89, rel. = 1). The large increase in person separation indicates that the discrimination of person measures increased, and the test better targeted the sample.

CONCLUSION

Through fit statistics and the investigation of the item hierarchy, a dimensional view of rape is confirmed. We have advanced to the stage where rape outcomes can be assessed with standardized item calibrations. The results of this study are generalizable to a normal population aged 18-24 who respond to the dichotomous Yes-No format of the Sexual Experiences Survey. Anyone who intends to administer the SES to a smaller sample of this kind can use the item benchmarks in Figure 4 to measure people using Rasch analysis. These item calibrations are applicable to both male offenders and female victims. To learn more about the offenders who misuse their authority for sexual contact, items 2 and 7 should be analyzed again using a different sample. Ideally, the sample's responses to these two questions should be reasonably distributed. This would provide logical positions for 2 and 7 on the item hierarchy.

Future research should investigate the degree to which item calibra-

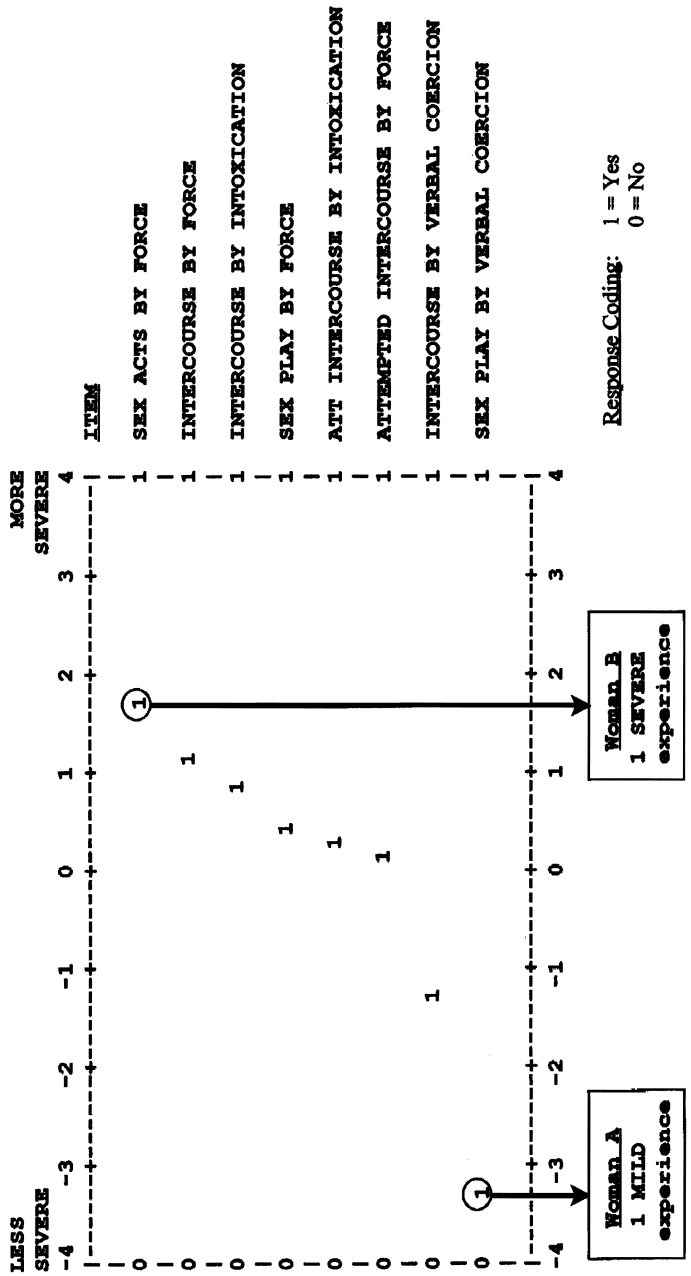


FIGURE 6. An example which uses the KEYFORM to compare two women in their different experiences (analyzed: 2,293 males and females).

Table 2
Person Measures After Data Edited through the Scalogram Method. Respondents Identifies by Response String. Total Sample Represents All Possible Raw Scores of 0,1,2.

PERSON NUMBER	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	PERSON'S RESPONSE STRING	
16	2	2	2.592	1.647	MAXIMUM ESTIMATED MEASURE					00000011	
22	2	2	2.478	1.662	MAXIMUM ESTIMATED MEASURE					00000101	
27	2	2	2.302	1.700	MAXIMUM ESTIMATED MEASURE					00001001	
31	2	2	2.246	1.717	MAXIMUM ESTIMATED MEASURE					00010001	
34	2	2	2.240	1.719	MAXIMUM ESTIMATED MEASURE					00100001	
15	2	3	1.991	1.244	.72	-.7	.67	-.8	.98	00000110	
36	2	2	1.922	1.873	MAXIMUM ESTIMATED MEASURE					01000001	
21	2	3	1.860	1.265	.67	-.8	.61	-.8	.93	00001010	
26	2	3	1.817	1.275	.66	-.8	.59	-.8	.87	00010010	
30	2	3	1.813	1.276	.66	-.8	.59	-.8	.82	00100010	
37	2	2	1.793	1.977	MAXIMUM ESTIMATED MEASURE					10000001	
9	1	1	1.772	2.000	MAXIMUM ESTIMATED MEASURE					00000001	*Woman A*
33	2	3	1.578	1.364	.68	-.8	.52	-.4	.78	01000010	
35	2	3	1.486	1.419	.73	-1.1	.50	-.2	.76	10000010	
8	1	2	1.470	1.430	.74	-1.3	.74	-1.3	1.00	00000010	
14	2	4	1.056	1.030	.70	-1.4	.69	-1.4	.74	00001100	
20	2	4	1.014	1.037	.68	-1.4	.67	-1.3	.76	00010100	
25	2	4	1.009	1.038	.68	-1.3	.66	-1.3	.76	00100100	
29	2	4	.732	1.131	.67	-.8	.58	-.7	.75	01000100	
32	2	4	.590	1.219	.76	-.6	.56	-.3	.74	10000100	
7	1	3	.560	1.242	.79	-.5	.74	-.6	.61	00000100	
13	2	5	.450	.944	.66	-1.4	.63	-1.3	.69	00011000	
19	2	5	.446	.945	.66	-1.4	.63	-1.3	.73	00101000	
24	2	5	.147	1.026	.58	-1.1	.51	-1.1	.74	01001000	
12	2	6	.024	.893	.74	-.9	.67	-.9	.69	00110000	
28	2	5	-.052	1.140	.68	-.6	.49	-.6	.74	10001000	
6	1	4	-.102	1.179	.73	-.5	.61	-.6	.53	00001000	
18	2	6	-.272	.958	.62	-1.0	.53	-1.0	.73	01010000	
23	2	6	-.515	1.085	.72	-.5	.51	-.6	.74	10010000	
11	2	7	-.575	.910	.67	-.8	.57	-.8	.69	01100000	
5	1	5	-.586	1.142	.80	-.3	.62	-.5	.53	00010000	
17	2	7	-.845	1.041	.77	-.4	.56	-.6	.71	10100000	
4	1	6	-.937	1.114	.88	-.2	.67	-.4	.53	00100000	
10	2	8	-1.432	1.034	.41	-1.2	.25	-1.1	.65	11000000	
3	1	7	-1.596	1.154	.52	-.8	.28	-.9	.51	01000000	
2	1	8	-2.803	1.335	.31	-1.1	.13	-.8	.45	10000000	*Woman B*
1	0	8	-3.891	1.665	MINIMUM ESTIMATED MEASURE					00000000	

tions are stable when the survey is administered to different age groups, and samples of victims and offenders involved in multiple offenses. It is likely that the latter sample will discriminate more than two response categories. This would allow analysts to separate (in terms of measures) offenders who have offended once versus offenders who have offended multiple times. The same is true in the measurement of female victims.

The investigation has also uncovered a psychometric issue inherent in pathological scales. These instruments do not target the sample as well, since their items probe rare events.Ordered strings of responses (i.e., 11101000 or 11110000) which are typically found in attitude surveys and ability tests are infrequent. In these cases, the scalogram method provides an alternative for person measurement. Table 2 illustrates that the method yielded person fit statistics which were overly predictable (all MNSQ $\leq$ .88). Therefore, however useful the method may be, it is not perfect. Current research is experimenting with different scalogram methods to determine the best way to handle disjointed strings of responses. The aim is to discover a method which generates response strings with MNSQ values near 1.

REFERENCES

Brogden, H. E. (1977). The Rasch model, the law of comparative judgment and additive conjoint measurement. *Psychometrika*, 42, 631-634.

Guttman, L. (1977). What is not what in statistics. *The Statistician*, 26, 81-107.

Kanin, E. J. (1957). Male aggression in dating-courtship relations. *American Journal of Sociology*, 63, 97-204.

Kanin, E. J. & Parcell, S. R. (1977). Sexual aggression: A second look at the offended female. *Archives of Sexual Behavior*, 6, 67-76.

Karabatsos, G. (1997). Omit inconsequential responses. *Rasch Measurement Transactions, 10,* 523.

Kinston, W. (1985). Measurement and the structure of scientific analysis. *Systems Research*, 2, 95-104.

Kirkpatrick, C. & Kanin, E. J. (1957). Male sex aggression on a university campus. *American Sociological Review*, 22, 52-58.

Koss, M. P. & Gidycz, C. A. (1985). Sexual Experiences Survey: Reliability and Validity. *Journal of Consulting and Clinical Psychology,* 53, 422-423.

Koss, M. P. Gidycz, C. A., & Wisniewski, N. (1987). The scope of rape: Incidence and prevalence of sexual aggression and victimization in a national sample of higher education students. *Journal of Consulting and Clinical Psychology*, 55, 162-170.

Koss, M. P. & Oros, C. J. (1982). Sexual Experiences Survey: A research instrument investigating sexual aggression and victimization. *Journal of Consulting*

and Clinical Psychology, 50, 455-457.
Law Enforcement Assistance Administration. (1975). *Criminal victimization surveys in 8 American cities* (Publication No. SD-NCS-C-5). Washington, DC: U.S. Government Printing Office.
Linacre, J. M. & Wright, B. D. (1997). *A user's guide to BIGSTEPS: Rasch model computer program.* Chicago: MESA Press.
Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology,* 1, 1-27.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press. Original edition, the Danish Institute for Educational Research, 1993.
Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Teachers College, Columbia University. 1913. Revised and enlarged edition.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology,* 33, 529-554.
Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology,* 17, 446-457.
U.S. Bureau of the Census. (1980). *Current population reports 1980-1981* (Series P-20, No. 362). Washington, DC: U.S. Government Printing Office.
Weis, K. & Borges, S.S. (1973). Victimology and rape: The case of the legitimate victim. *Issues in Criminology,* 8, 71-115.
Wright, B. D. & Stone, M. H. (1979). *Best Test Design.* Chicago: MESA Press.
Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

APPENDIX A

THE SEXUAL EXPERIENCES SURVEY FORM FOR WOMEN (VICTIMS)

1. Have you given into *sex play* (fondling, kissing, or petting, but not intercourse) when you didn't want to because you were overwhelmed by a man's continual *arguments and pressure*?
2. Have you had *sex play* (fondling, kissing, or petting, but not intercourse) when you didn't want to because a man used his position of *authority* (boss, teacher, camp counselor, supervisor) to make you?
3. Have you had *sex play* (fondling, kissing, or petting, but not intercourse) when

you didn't want to because a man *threatened or used some degree of physical force* (twisting your arm, holding you down, etc.) to make you?

4. Have you had a man *attempt sexual intercourse* (get on top of you, attempt to insert his penis) when you didn't want to by *threatening or using some degree of physical force* (twisting your arm, holding you down, etc.), but intercourse did not occur?
5. Have you had a man *attempt sexual intercourse* (get on top of you, attempt to insert his penis) when you didn't want to by giving you *alcohol or drugs*, but intercourse did not occur?
6. Have you given in to *sexual intercourse* when you didn't want to because you were overwhelmed by a man's continual *arguments and pressure*?
7. Have you had *sexual intercourse* when you didn't want to because a man used his position of *authority* (boss, teacher, camp counselor, supervisor) to make you?
8. Have you had *sexual intercourse* when you didn't want to because a man gave you *alcohol or drugs*?
9. Have you had *sexual intercourse* when you didn't want to because a man *threatened or used some degree of physical force* (twisting your arm, holding you down, etc.) to make you?
10. Have you had *sex acts* (anal or oral intercourse or penetration by objects other than the penis) when you didn't want to because a man *threatened or used some degree of physical force* (twisting your arm, holding you down, etc.) to make you?

APPENDIX B

THE SEXUAL EXPERIENCES SURVEY FORM FOR MEN (OFFENDERS)

1. Have you engaged in *sex play* (fondling, kissing, or petting, but not intercourse) with a woman when she didn't want to by overwhelming her with continual *arguments and pressure*?
2. Have you engaged in *sex play* (fondling, kissing, or petting, but not intercourse) when she didn't want to by using your position of *authority* (boss, teacher, camp counselor, supervisor)?
3. Have you engaged in *sex play* (fondling, kissing, or petting, but not intercourse)

with a woman when she didn't want to by *threatening or used some degree of physical force* (twisting your arm, holding you down, etc.)?

4. Have you *attempted sexual intercourse* (get on top of you, attempt to insert his penis) with a woman when she didn't want it by *threatening or using some degree of physical force* (twisting your arm, holding you down, etc.), but intercourse did not occur?
5. Have you *attempted sexual intercourse* (get on top of you, attempt to insert his penis) with a woman when she didn't want to by giving her *alcohol or drugs*, but intercourse did not occur?
6. Have you engaged in *sexual intercourse* with a woman when she didn't want to by overwhelming her with continual *arguments and pressure*?
7. Have you engaged in *sexual intercourse* with a woman when she didn't want to by using your position of *authority* (boss, teacher, camp counselor, supervisor)?
8. Have you engaged in *sexual intercourse* with a woman when she didn't want to by giving her *alcohol or drugs*?
9. Have you engaged in *sexual intercourse* with a woman when she didn't want to by *threatening or using some degree of physical force* (twisting your arm, holding you down, etc.)?
10. Have you engaged in *sex acts* (anal or oral intercourse or penetration by objects other than the penis) with a woman when she didn't want to by *threatening or using some degree of physical force* (twisting your arm, holding you down, etc.)?

JOURNAL OF OUTCOME MEASUREMENT®, *1*(4), 329 - 362

Equating the MOS SF36 and the LSU HSI Physical Functioning Scales

William P. Fisher, Jr.

Robert L. Eubanks
and
Robert L. Marier
Louisiana State University Medical Center - New Orleans

This study equates the physical functioning subscales of the Medical Outcomes Study Short Form 36 (SF36) and the Louisiana State University Health Status Instruments (LSU HSI). Data from the SF36's 10-item physical functioning scale, the PF10, and the LSU HSI's 29-item Physical Functioning Scale (PFS), were fit to separate and mixed Rasch rating scale models. Data were provided by a convenience sample of 285 patients waiting for appointments in a public hospital general medicine clinic. Difficulty estimates for a subset of similar items from the two instruments were highly correlated (.95), indicating that the items from the two scales are working together to measure the same variable. The measures from the two equated instruments correlate .80 (.86 when disattenuated for error). Of the two instruments, the PFS's error is lower, model fit is better, and reliability coefficients are higher. Both instruments measure physical functioning, and can do so in a common unit of measurement. Conversion tables are provided for transforming raw scores from either instrument into the common metric.

Requests for reprints should be sent to William P. Fisher, Jr., Louisiana State University Medical Center, 1600 Canal Street, Suite 800, New Orleans, LA 70112.

OBJECTIVE

The purpose of this ongoing study is to equate (Masters, 1985; Wilson, 1994; Fisher, Harvey, & Kilgore, 1995; Fisher, Harvey, Taylor, et al., 1995; Cella, et al., 1996; Gonin, et al., 1996; Zhu, 1996) the physical functioning subscales of the Medical Outcomes Study Short Form 36-item health status measure (MOS SF-36, or SF36) (Haley, et al., 1994; McHorney, et al., 1997; Stucki, et al., 1996) and the Louisiana State University Health Status Instruments (LSU HSI) (Fisher, et al., 1997). The instruments' respective measurement properties are evaluated and compared via application of Rasch measurement models (Andrich 1978; Rasch, 1960; Wright & Masters, 1982) en route to the creation of a common metric.

METHOD

Analysis Steps

Data from the SF36's 10-item physical functioning scale (PF10) and the LSU HSI's 29-item Physical Functioning Scale (PFS) were first analyzed separately using BIGSTEPS, a Rasch analysis computer program (Linacre & Wright, 1997), in order to evaluate model fit and person separation reliability for each instrument. Second, the data were analyzed separately again, after maximizing model fit and person separation reliability, to produce initial estimates of physical functioning for the common sample. Third, the equivalence of the two analyses was determined by a) comparing the two instruments' item difficulties, and b) comparing the two instruments' case estimates for the common sample using graphical and correlational methods. Fourth, with equivalence established, a linear transformation for equating the measures was computed from the ratio of their standard deviations and the average difference between them.

This procedure was compared with the results of a co-calibration, where the data from both instruments were first pooled into a single analysis for equating. The item difficulties are then anchored at their co-calibrated values in separate analyses to produce measures from each instrument that can then be compared for statistical identity. There were two reasons for undertaking both kinds of equating.

The primary reason was to assess the effects of the instruments' different numbers of items and rating scale points. With 10 items and 3 rating categories, the PF10 offers respondents 20 (2 per item) possible distinctions in their physical functioning. In contrast, the PFS with 29 items and

6 rating categories offers 145 distinctions (5 per item). There are two main effects of this difference. Both hinge on the fact that the PFS's more detailed structure dominates the co-calibration and forces the PF10 to be evaluated on its terms. First, because the PF10 has fewer rating categories, it can be expected that transitions from category to category will be more rigidly structured and deterministic (Guttman-like) than they are on the PFS. Respondents have the opportunity to make more than seven times the number of distinctions in their physical functioning on the PFS, so it will be more finely tuned and graduated than the PF10. Relative to the PFS, then, the PF10's structure predisposes it to greater statistical inconsistency and worse model fit. Co-calibrating the two instruments on a common sample accentuates that predisposition. Comparing the results of separate and pooled calibrations makes it possible to evaluate how much difference the PFS's frame of reference makes in assessing the PF10's measurement quality and the potential for equating.

The second effect of the different number of distinctions offered by the instruments involves the standard deviations of their measures. To the extent that respondents use the two instruments consistently and report roughly the same amount of physical functioning on each of them, the PF10, having fewer items and rating categories, will produce measures with more variation than the PFS. Different variances can be accounted for in the first equating method by dividing the PFS's standard deviation by the PF10's, multiplying the PF10 measures by this factor, and adding the average difference between each instrument's measures. Equating is then completed by determining the extent to which the measures from each instrument have the same mean and standard deviation, as shown by a paired-samples t-test.

The second reason for undertaking both separate and co-calibrated equatings was that equating measures from the two instruments via separate BIGSTEPS analyses alone does nothing to establish whether the items and rating categories are positioned in meaningful relation to one another on the variable. Co-calibration of the two instruments on their pooled data is required for this investigation. Furthermore, respective instrument measurement and calibration quality was kept in perspective by comparing the results of the separate and co-calibration analyses. The possibly negative effect of the larger PFS on the PF10 model fit statistics was then evaluated by comparing the separate and combined analyses. The measures produced by the anchored item difficulties were similarly adjusted for different variances in the same manner as employed in the equating based on separate

BIGSTEPS analyses. Then the two sets of equated measures were compared with one another for statistical identity.

Instruments

The LSU HSI PFS is shown in Appendix A. Its rating scale includes six response categories: 1) Impossible, 2) Very Difficult, 3) Difficult, 4) Manageable, 5) Easy, and 6) Very Easy. The 29 questions on this scale were devised to span a wide range of difficulty in physical functioning. Previous research demonstrating the consistency of task difficulties across instruments (see Fisher, 1997a, for a review and synthesis) provided useful information about item functioning that was incorporated into the design of the PFS. Most of this previous research involved clinician ratings of client performance. The results of the present study's focus on client ratings will be useful in future studies that examine whether clinician ratings and client self-reports produce mutually consistent item difficulty orders.

The first page of the SF36, containing the 10-item PF10 (questions 3.a. to 3.j.), is shown in Appendix B. The SF36 is a 36-item general health status measure composed of eight subscales: health perceptions, physical functioning, role limitations attributed to physical health, role limitations attributed to emotional health, social functioning, mental health, bodily pain, and energy and fatigue. The PF10 has three response categories: 1) Yes, Limited a Lot; 2) Yes, Limited a Little; and 3) No, Not Limited at All.

Grouped Partial Credit Modeling for Mixed Rating Scales

The second equating method applied to these data using BIGSTEPS involved pooling the data from both instruments into a single analysis, and then anchoring the item difficulties at the estimates derived from this combined analysis in separate studies of each instrument's items.

This kind of analysis takes advantage of the fact that, when groups of items from two or more instruments, or within a single instrument, have different response categories, BIGSTEPS can estimate parameters for multiple rating scale models (Andrich, 1978). Partial credit models (Masters, 1982) typically allow each item to have its own rating scale. Blending partial credit and rating scale models allows items with common rating scales to be grouped together in rating scale fashion. An important theoretical strength of the partial credit and mixed rating scale models is that they should make it possible to connect different item and rating scale

combinations into a common measurement system, based on observations from a single sample.

When items are grouped by their rating scale structures, item difficulty order is determined by choosing a key category in the step structure of each instrument on which the scale can be said to pivot. This category is then used as the basis for ordering the item difficulty estimates. In situations where different instruments with different rating scales are being equated, the pivot category can be used to anchor the step difficulties from the two scales so that they are meaningfully positioned in relation to one another on the measurement continuum (Linacre & Wright, 1997, pp. 56-7). Because success in physical functioning is the variable of interest, the third step, from Manageable to Easy, was chosen as the pivot for the LSU HSI PFS, and the step from Yes, Limited a Lot, to Yes, Limited a Little, was chosen for the SF36 PF10.

Though the meanings of these categories appear to concur, the extent of their concurrence must be evaluated. The quantitative and qualitative meanings of the categories and the items they are associated with cannot be entrusted solely to the measurement software, and neither can similar categories be linked purely on the basis of theory or hunches.

The dangers of not attending to the instruments' combined expression of the construct are 1) that items and rating scale categories with perhaps quite different qualitative meanings and quantitative implications could calibrate near (within an error of) each other; 2) that items and rating scale categories with perhaps quite similar qualitative meanings and quantitative implications could calibrate far (several errors) away from each other; and 3) that the meaning of the measures could be uninterpretable. These problems are overcome by attending to 1) the difficulty order of similar items from the two instruments; 2) the spacing of these items across rating categories on the measurement continuum; 3) the meaning of the rating option labels; and 4) the quantitative differences in the measures produced by each instrument.

Psychometrics has historically been more concerned with equating scores, and less concerned with producing qualitatively meaningful criterion-referenced measures from equated instruments. In this traditional approach, it would be acceptable to focus solely on reducing the quantitative differences in the measures to zero, and to ignore construct-related issues. The approach employed here focuses on the requirements of meaningfulness, and on theory development, following through to the construction of measures, instead of taking the more commonly employed, oppo-

site tack of assuming the numbers to be meaningful, and creating an interpretation from them after the fact of their construction, perhaps from accidental or unintended events.

Study Participants

A convenience sample of 285 persons presenting themselves for care by appointment in a public hospital general medicine clinic filled out the complete SF36 and LSU HSI forms while waiting to see a doctor. About 65 percent of respondents reporting demographic information are African-Americans, 75 percent are female, about 60 percent are between the ages of 30 and 60, 63 percent have a high school education or less, and about 70 percent have annual incomes of $15,000 or less. A variety of medical conditions and comorbidities are present; these primarily involve diabetes, hypertension, obesity, and arthritis.

Hypotheses

Two hypotheses are tested in this study: 1) that the PFS and the PF10 measure the same physical functioning variable; and 2) that they can do so in the same quantitative metric. The hypotheses were tested by comparing the two scales' combined and separate calibrations, measures, errors, mean square and standardized infit and outfit, separations, and reliabilities.

Success in measuring the same variable is determined 1) by showing that similar, pseudo-common (Fisher, 1997a) items calibrate with statistically identical order and spacing, as established by plots and correlation coefficients; and 2) by showing that the measures also scale with statistically identical order and spacing. Determining that the same variable is measured by both scales is necessary but not sufficient for establishing a common metric. If the first hypothesis is not falsified, neither is the second, but the positioning of the measures on the identity line in a plot may require the additional work of linearly adjusting one or the other set of measures.

The SF36 is a widely accepted health status measure, not a new instrument with unknown measurement properties. Thus, initial analyses, not just of the PF10 but of the PFS as well, focused on obtaining the highest possible person separation reliability given the existing items. This approach stands in contrast to situations in which data from a new instrument are analyzed for the first time; in this case, statistically inconsistent

items are examined and modified or removed as needed. For the purposes of this calibration and equating study, cases were removed from each instrument's data in the initial, separate analyses until there was no further improvement in person separation reliability. The remaining cases common to both instruments, and which had data from at least half of each instrument's items, were then used as the basis for the equating.

Errors and Logits

All error terms are inflated by the fit statistics to adjust for inconsistencies in the data. The adjustment is computed (Wright, 1995) by multiplying the modeled error term (Wright & Masters, 1982) by the larger of two values, 1.0 or the square root of the information-weighted mean square model fit statistic (infit) (Wright & Masters, 1982).

Logit person ability and item difficulty estimates typically span a range between -10.0 and 10.0. For more information on logits and algorithms for estimating them, see Wright and Masters (1982) or Ludlow and Haley (1995).

Statistical Procedures

When the item and step difficulties for each instrument were anchored in separate BIGSTEPS analyses, the measurement output from the two analyses was read into a single SPSS file indexed on the record number. The SPSS procedure was written in reusable syntax (Appendices C and D), making it possible to complete the entire sequence, from the BIGSTEPS analysis to final examination of the measurement statistics, in a matter of seconds.

RESULTS

SF36 PF10

Tables 1 and 2 summarize the results of the two separate and single combined analyses. Initial rating scale analysis of the PF10 produced a .97 item separation reliability and .80 person separation reliability. Of the available 285 respondents, 20 attained the maximum score, 13, the minimum score, and 17 lacked responses. Responses were available for about 9.4 of the 10 items, on average. The standard deviation of the outlier-sensitive mean square fit statistic (outfit) for the measures was 1.1; for the item calibrations, it was .67. The number of responses per category, summing

Table 1
PFS and PF10 Item Measure Results

Indicator	LSU HSI PFS	SF36 PF10	Combined
Average calibration	0.0	0.0	0.0
Calibration SD	1.5	2.7	1.73
Number of items	29	10	39
Avg number of responses/item	173	147	105
Avg calib error	.13	.23	.18
Calib separation (SD / err)	11.2	11.3	9.3
Avg calib mean sq outfit	.98	.86	1.00
Calib mean sq outfit SD	.26	.27	.30
Avg calib mean sq infit	1.02	1.03	1.02
Calib mean sq infit SD	.30	.25	.30
Item separation reliability	.99	.99	.99

Table 2
PFS and PF10 Person Measurement Results

Indicator	LSU HSI PFS	SF36 PF10	Combined
Avg measure	.65	-.23	.24
Measure SD	2.3	3.2	2.1
Number of measures	198	153	113
Avg number of item responses/person	25.4	9.6	36.3
Root mean sq error	.4	1.1	.3
Meas separation (SD/err)	5.7	3.0	6.6
Measurement strata[2]	7.9	4.3	9.1
Avg meas mean sq outfit	.95	.85	.98
Meas mean sq outfit SD	.49	.74	.44
Avg meas mean sq infit	1.0	1.0	1.0
Meas mean sq infit SD	.43	.63	.35
Measure reliability	.97	.90	.98

[2]Strata are ranges in the measurement continuum with centers separated by three errors (Wright and Masters, 1982). To calculate the number of strata, multiply the separation statistic (SD / err) by 4; add 1; and divide by 3.

across all of the PF10 items, ranged from 624 to 917. There were 2213 actual observations out of a possible 2350 (10 times 235) possible observations, making for 94 percent completed.

The least consistent cases, those with the highest outfit statistics, were removed over the course of several subsequent analyses, until there was no further improvement in person separation reliability. This process removed 82 cases from the analysis, reducing the sample size to 153. Item separation reliability improved, increasing from .97 to .99, and person separation reliability improved from .80 to .90. The outfit standard deviation dropped from .67 to .27 for the item difficulty estimates, and from 1.1 to .74 for the person ability estimates. The results of the latter analysis are shown in Tables 1 and 2.

Item difficulty estimates range from -8.0 to 7.0 logits across the two steps provided by the three rating categories. The distributions of the estimates on each of the steps include three gaps of more than 2.0 logits each, which amounts to more than two calibration errors of measurement. The majority (seven) of the items are grouped together in the middle of each step, ranging from about -2.8 to -0.2 logits on the first step, a 2.6 logit range, and from about 2.0 to 4.2 logits on the second step, a 2.2 logit range. Thus, of the 15 logit range of the difficulty estimates, 70 percent of the items have a combined range of 4.8 logits across the two steps. Measures can be interpreted in terms of a 50 percent chance of success in the area of physical functioning represented by an item on less than a third of the total range of measurement. There are 43 measures in the range represented by the majority of the items on the first step, and another 29 in the range covered by the items on the second step, for 72 (64 percent) total.

Cursory examination of the item order showed it to meet rough expectations formed on the basis of previous research on this instrument (Haley, et al., 1994; McHorney, et al., 1997; Stucki, et al., 1996), and on other similar instruments (Fisher, 1997a). Simple, less demanding tasks, such as bathing and dressing, or walking short distances, were consistently easier (had lower difficulty estimates) than complex and more demanding tasks, such as climbing stairs or vigorous activities.

LSU HSI PFS

The first BIGSTEPS rating scale analysis of the PFS revealed that its first two categories, Impossible and Very Difficult, were not distinct steps in a progression of increasing physical functioning. The Very Difficult response

option was never the most likely response, so the step structure advanced over it from Impossible directly to Difficult. The PFS data were therefore recoded from a six-category, five-step structure into a five-category, four-step structure. Each of the remaining five categories stands for an increase in the amount of the variable measured.

The initial 5-category analysis of the PFS produced a .99 item separation reliability and a .95 person separation reliability, on a sample of 262 of the original 285 respondents. Of the 23 cases not included in the analysis, 17 attained the maximum score, 3 had the minimum score, 1 had no responses, and 2 were removed from the analysis because the respondents evidently misread the form and reversed their ratings. On average, the respondents answered 25.5 of the 29 questions. The outfit standard deviation was 1.0 for the respondents, and .78 for the items. The number of responses per category, summing across all of the PFS items, ranged from 918 to 1876. There were 6681 actual observations, out of a total of 7598 (29 times 262) possible observations, for 88 percent completion.

After several iterations through the process of removing the least consistent cases, item separation reliability did not improve from the original .99, and person separation reliability improved from .95 to .97, with the respective outfit statistics at .26 and .49, on a sample of 198 (66 of the original respondents removed). The results of the latter analysis are shown in Tables 1 and 2.

Item difficulty estimates range from -5.2 to 6.0 logits across the four steps provided by the five rating categories. The distributions of the estimates on each of the steps include one gap of about 1.0 logits, or about two errors of measurement. Except for this one gap, the items are spread fairly evenly over a range of 5.0 logits on each step. Because the item distributions on each step overlap, the gap on lower, easier steps' distributions is covered by the position of the items on the higher, more difficult steps, making the meaning of almost all of the measures directly related to a 50 percent probability of success on some items on some steps. There are three measures at -6.0 logits, lower than any item calibration, and there are 23 measures above 6.0 logits, leaving 87 (77 percent) within the items' calibration range. Again, item order appeared to meet expectations across these analyses.

Combined instrument analysis

At this point, the items from both instruments were analyzed together, using data from the 113 respondents that 1) were common to the analyses

producing the two instruments' highest reliabilities; 2) did not have maximum nor minimum extreme scores; and 3) had responses to at least half of the items on each instrument. (Comparison of analyses for each instrument using this restricted data set with analyses based on the best data set showed negligible changes in outfit and reliability.) Of the total 39 items, the average respondent used 36.3 of them; 93 percent of the possible responses were completed.

The first analysis of the data from the 39-item combined instrument and 113 measured persons produced satisfactory model fit, converging in 73 UCON iterations to a maximum logit change of .0003, and a maximum score residual of -.01. Using the fit-inflated error, for the 39-item combined scale, item separation reliability was .99, person separation reliability was .98. The average infit and outfit statistics were within .02 of 1.0 for both the persons and items, with the associated standard deviations less than .45 for the persons and .30 for the items. Item difficulty estimates were centered at 0.0, with a standard deviation of about 1.7 and an average fit-inflated error of about .18. Person measure estimates averaged about 0.2, with a standard deviation of about 2.1, and an average fit-inflated error of 0.3. Other results from this analysis are shown in Tables 1 and 2.

Comparison of Item Calibrations from the Initial Analyses

Eight of the PF10 items have corresponding items in the PFS that address similar areas of physical functioning (see Figures 1 to 4 and Table 3). The difficulty estimates for the items from both the separate and combined analyses of the different instruments correlate .95 (both the correlation and the item separation reliability coefficients are too high to make disattenuation meaningful or useful). Figures 1 and 2 show that the items' difficulty estimates from the separate and combined calibrations are not statistically identical. Since the items do not represent identical areas of physical functioning, this result does not deny the possibility of equating the two instruments, but does present an opportunity for understanding more about the effects of the instruments' differing numbers of rating categories and items.

The effect of the PF10's fewer number of categories on its estimates' variation is especially evident in the comparison of the separate calibrations shown in Figure 1. The estimates range from -6.0 to almost 5.0 logits on the PF10, but from only -2.0 to 2.5 for the PFS. (These plots show only the average calibration per item across all category transitions.) The cor-

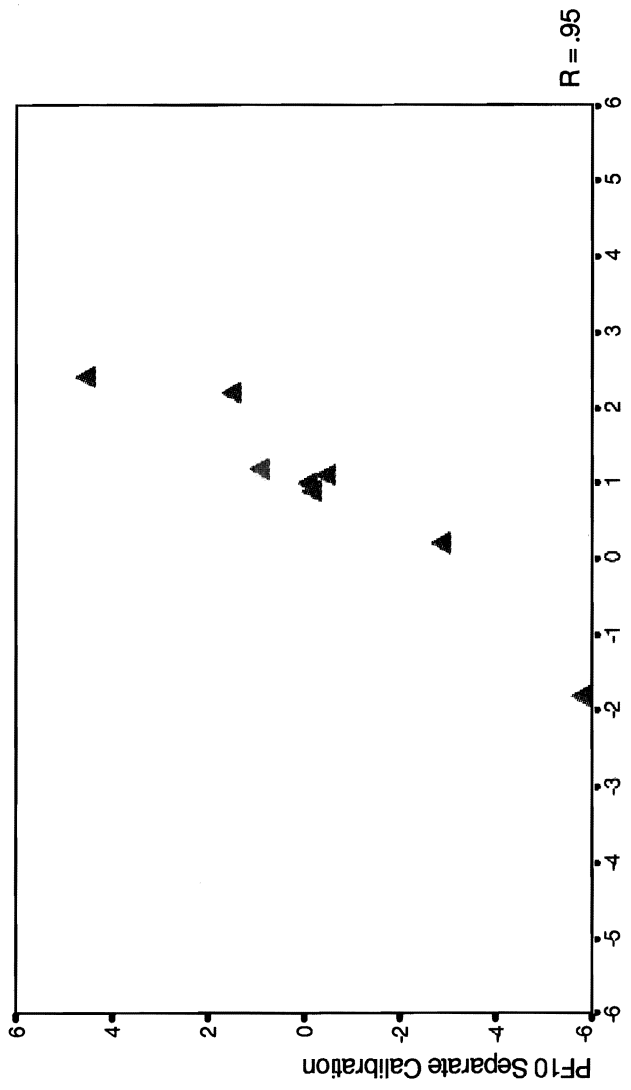


FIGURE 1 Eight pseudo-common PF10 and PFS items drawn from two separate calibrations of the full instruments compared.

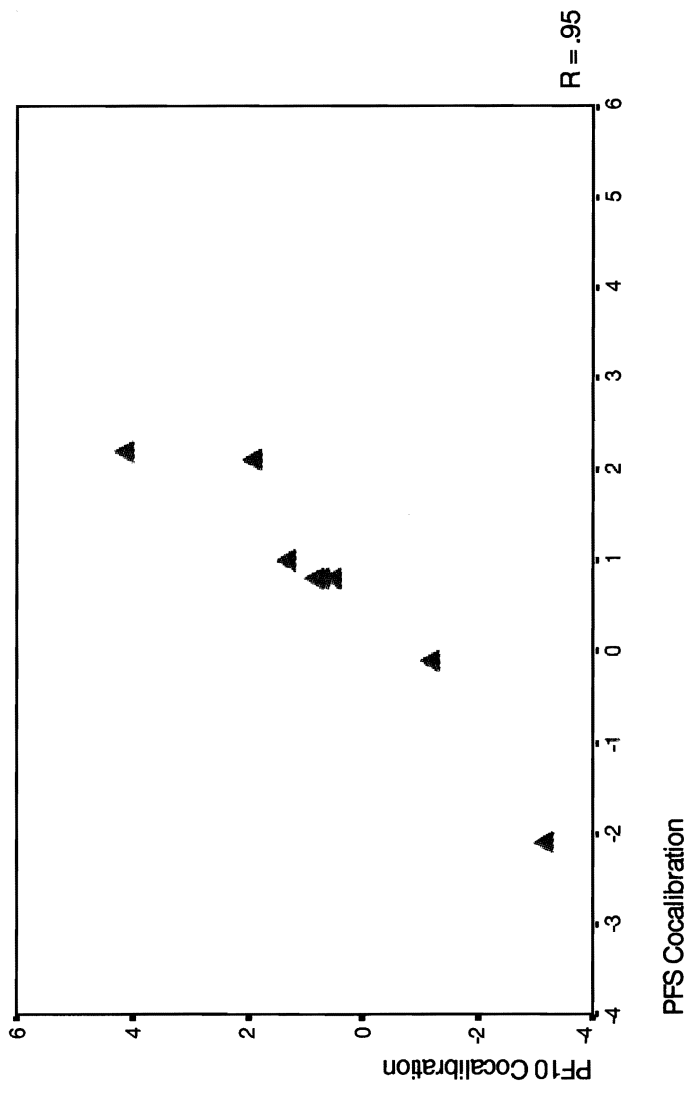


FIGURE 2 Eight pseudo-common PF10 and PFS items drawn from a co-calibration of the full instruments compared.

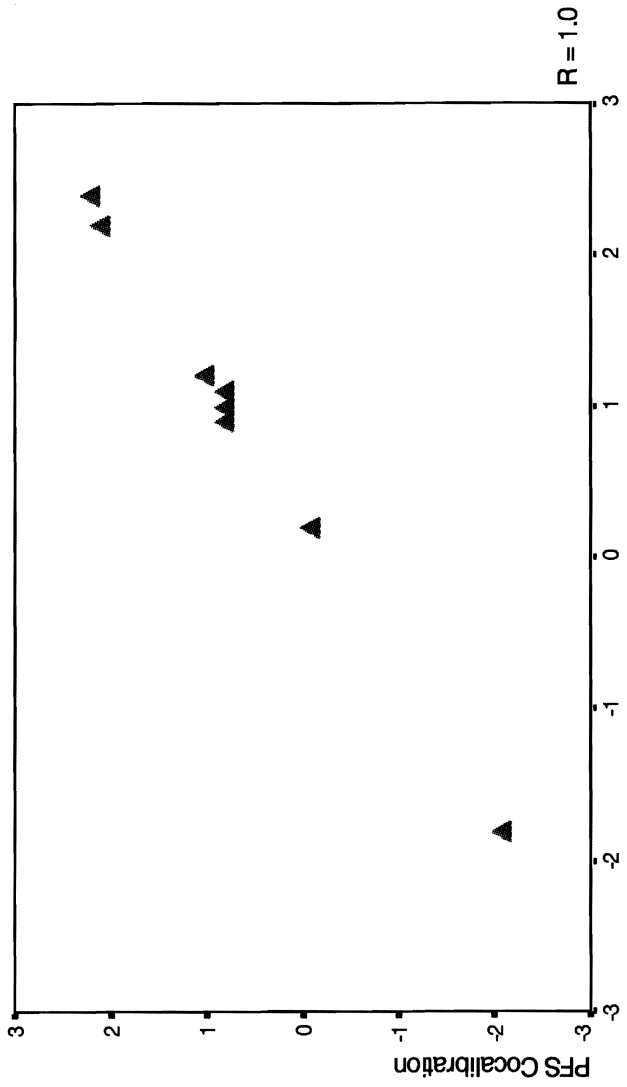


FIGURE 3 Eight PFS items drawn form two calibrations of the full instrument compared.

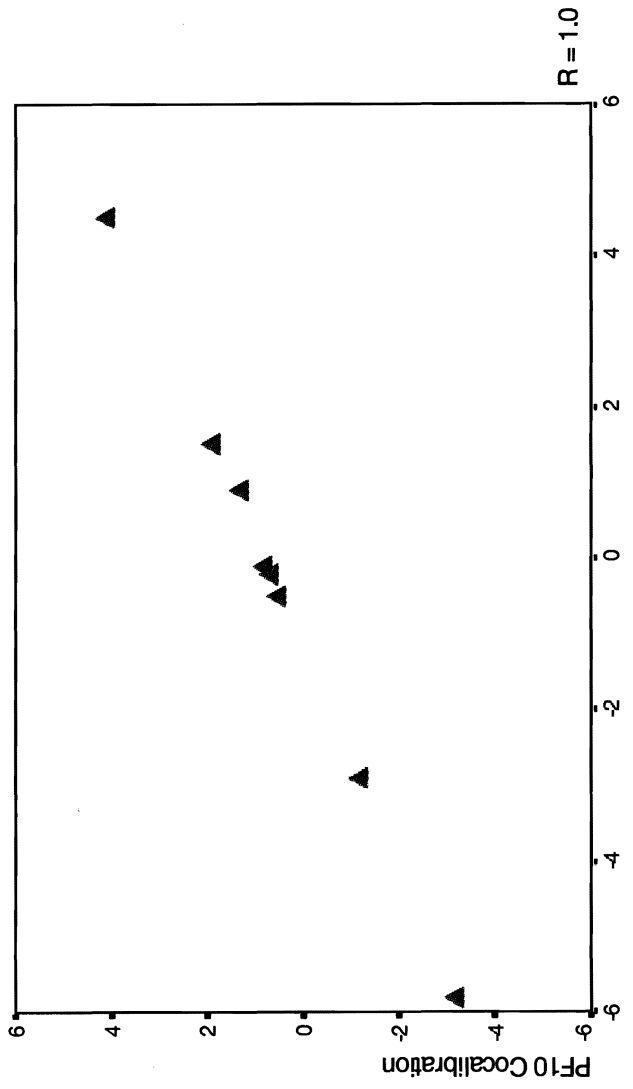


FIGURE 4 Eight PF10items drawn from two calibrations of the full instrument compared.

Table 3
PFS and PF10 Calibration Results
For a Subset of Similar Items

PF10 Separate Calibration N=153	PF10 Combined Calibration* N=113	PF10 Items Names	PFS Separate Calibration N=198	PFS Combined Calibration N=113	PFS Item Names
4.5 (.3)	4.1 (.4)	Vigorous activities	2.4 (.1)	2.2 (.2)	Exercise hard for half an hour
1.5 (.2)	1.9 (.2)	Walk more than a mile	2.2 (.1)	2.1 (.2)	Walk 2 miles
.9 (2)	1.3 (.2)	Moderate activities	1.2 (.1)	1.0 (.2)	Do a day's office work
-.1(.2)	.8(.2)	Lift/Carry Groceries	1.0 (.1)	.8 (.1)	Carry a bag of groceries
-.2 (.2)	.7 (.2)	Walk several blocks	.9 (.1)	.8 (.1)	Walk half a mile
-.5 (.2)	.5 (.2)	Climb one flight of stairs	1.1 (.1)	.8 (.1)	Walk up a flight of stairs
-2.9 (.2)	-1.2 (.2)	Walk one block	.2 (.1)	-.1 (.1)	Walk a block
-5.8 (.3)	-3.2 (.2)	Bathe/Dress	-1.8 (.1)	-2.1 (.2)	Dress yourself

*Difficulty estimates are shown with standard errors in parentheses.

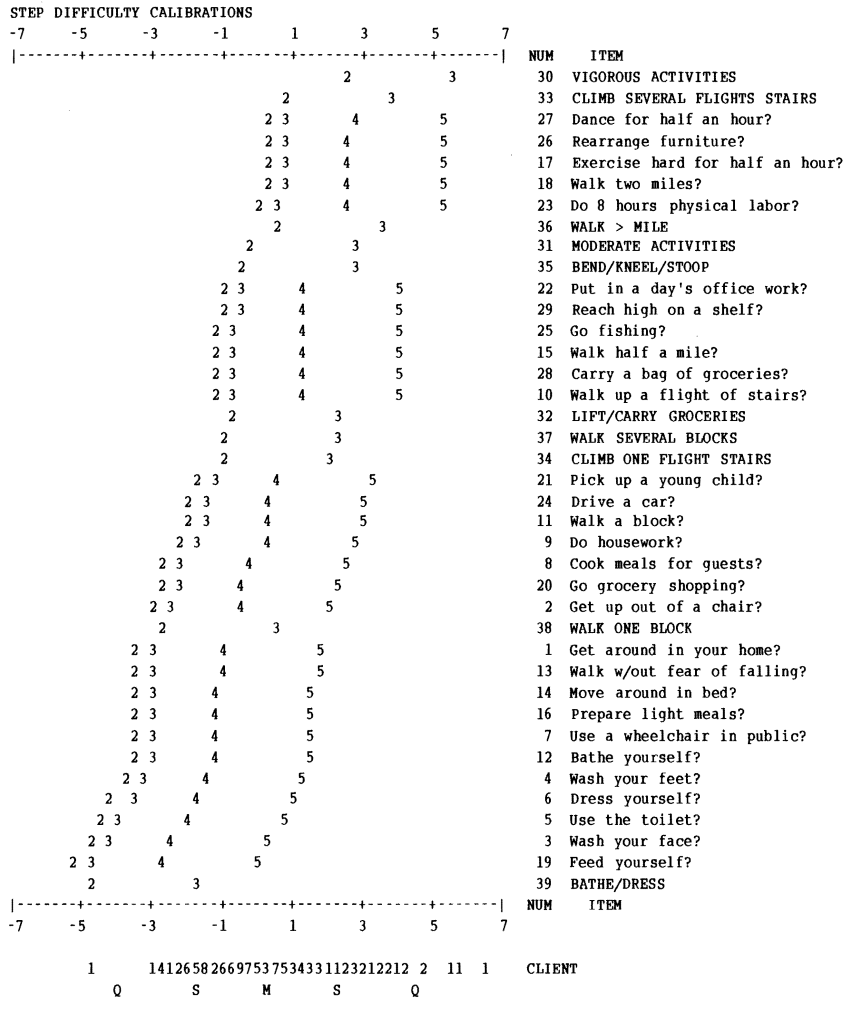
relation is high (.95), but the scatter plot shows the items crossing the identity line with a steep slope.

Figure 2 shows the effect of the PFS's greater number of items and rating categories on the respective calibrations of these similar items from the two instruments. Notice that the PF10's range is shrunk by almost 3.0 logits in the combined calibration, and that the PFS's range is about the same as it was in the first calibration. The items are much closer to the identity line. Only the most difficult pair of items are more than an error from the identity line, and this may result more from an actual difference in the items' difficulties than from problems related to the instruments' numbers of rating categories.

The same phenomenon can be viewed from another angle in Figures 3 and 4. Figure 3 shows the eight PFS items from the two calibrations on the identity line. Figure 4 shows that the two calibrations of the PF10 correlate 1.0, but that the variation in the estimates from the co-calibration is compressed. The compression is especially evident in the lower, easier end of the distribution, where the PF10 has many fewer items than the PFS.

Figure 5 shows the step difficulty calibrations for the items from the two instruments, with the items listed on the right in difficulty order. Notice that the difficulty of the step from category 2 to category 3 on the PFS items is about a fourth or a fifth of the difficulty associated with taking the next step, and an even smaller proportion of the difficulty of taking the final step. The transition from category 2 to category 3 on the PF10 is slightly larger than the largest step on the PFS. These variations in the distance between categories are what make the raw scores indicative only of order and not quantity. The natural logarithm of the odds that a step is taken produces the equal-interval -7.0 to 7.0 logit continuum across the top and bottom of Figure 5. The spacing of the rating categories on the logit number line shows that the amount of physical functioning represented by one additional raw score unit varies depending on which category is changed.

The constancy of the physical functioning construct across instruments evident in Figures 1-4 and in Table 3 also shows itself in Figure 5. The PFS items involving bathing and dressing calibrate at the bottom of the scale with the PF10 Bathe/Dress item, and the PFS items involving dancing, rearranging furniture, and exercise calibrate at the top of the scale with the PF10 Vigorous Activities item. The PF10 category 2 (Yes, Limited a Little) typically falls near the PFS categories 2 (Difficult) or 3 (Manageable), with the PF10 category 3 (No, Not Limited at All) falling be-



PF10 item names are in capital letters.
Category 1 is not shown for either instrument as it does not represent a step from one category to another.

FIGURE 5 Item and category order and spacing initial co-calibration.

tween the PFS categories 4 (Easy) and 5 (Very Easy).

The correspondence in the difficulty estimates for similar items and categories from the two instruments supports the hypothesis that the two instruments measure the same variable. The next test is to compare the measurement estimates for the respondents across the two separate calibrations.

Comparison of Person Measures from the Initial Analyses

Figure 6 is a plot of the 113 common measures produced in the separate analyses of the two instruments. The correlation of .80 ($p < .01$) indicates a fairly strong association between the two sets of measures. Correcting for attenuation, the correlation is .86. The correlation is disattenuated for error by dividing it by the square root of the product of the instruments' reliabilities (.90 for the PF10, and .97 for the PFS). Disattenuation does nothing to improve the measures or their predictive power, and it does not take the place of precise measurement, but it does indicate the extent to which a correlation is affected by measurement error, as opposed to expressing the extent to which two sets of measures are actually correlated (Muchinsky, 1996; Schumacker, 1996). In this instance, the PF10's lower person separation reliability is making the correlation appear lower than it would be if that instrument's measurement error were lower, given the same amount of variation in its measures. The .86 disattenuated correlation is strong enough to not falsify the hypothesis that the two instruments measure the same variable.

The differences in the two sets of measures' means produce a statistically significant paired-sample t of 7.61. The large difference in the variances for each set of measures indicates that equating the measures will entail more than simply adding a constant to one or the other sets of measures. An equating function can be calculated from the two sets of measures' standard deviations and their average difference. This equation for the PF10, for instance, would divide the PFS's standard deviation (2.3) by the PF10's (3.4), and multiply that result (.6765) by each of the PF10 measures. Then, 1.2, the average difference in the measures minus a small adjustment, would be added to each transformed measure. The results of this process are plotted in Figure 7, which, compared with Figure 6, shows that the equated PF10 measures have been shifted vertically, up the page, so that the measures from the two instruments now intersect along the identity line. A parallel procedure applied to the PFS measures produced equivalent results.

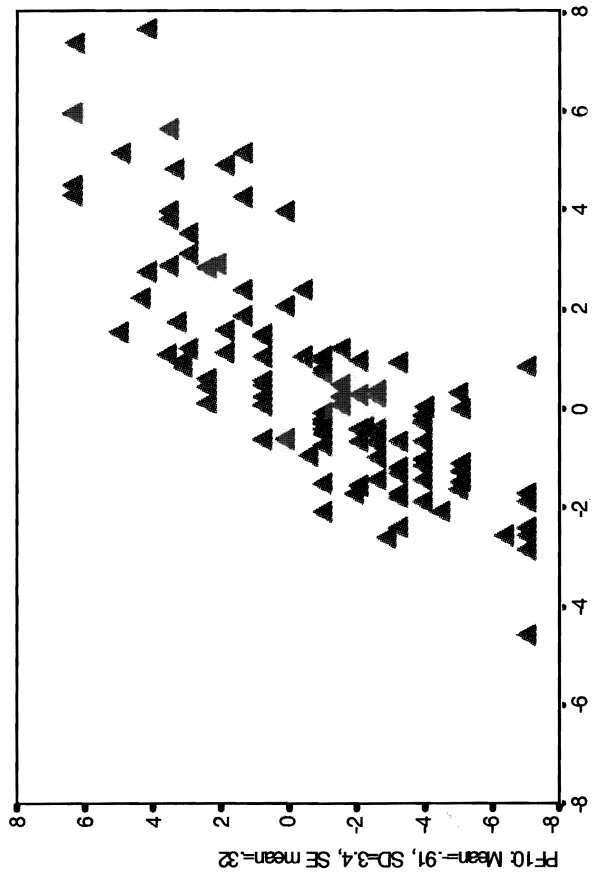


FIGURE 6 Measures from separate analyses of PF10 and PFS.

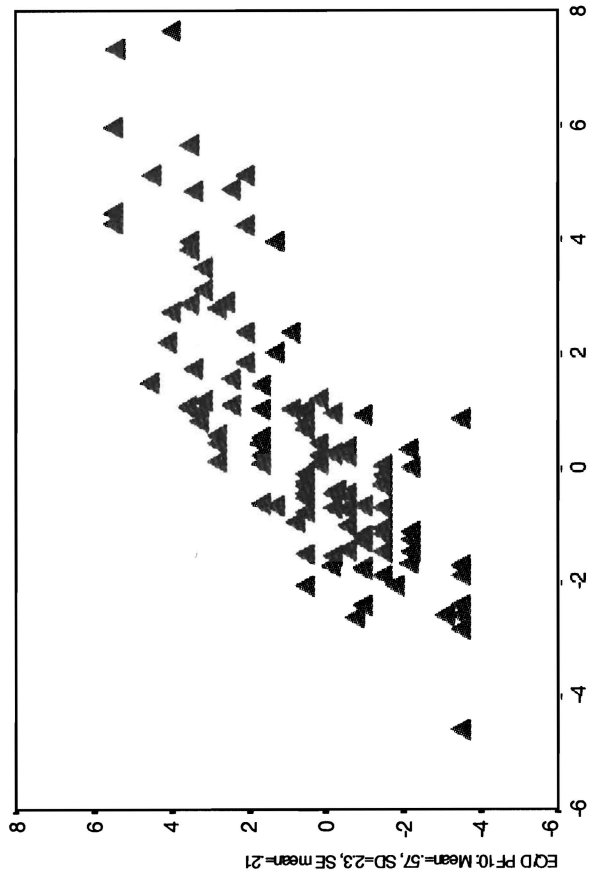


FIGURE 7 Measures from separate analysesequated via formula ((2.3/3.4)*PF10)+1.2.

If the only concern was to arrive at equivalent measures from each instrument, the results shown in Figure 7 might suffice. Insofar as they are to be quantitative, however, it is essential that the measures be interpretable as expressions of a unidimensional and invariant measurement continuum (Duncan, 1984; Fisher & Wright, 1994; Guttman, 1950; Krantz, et al., 1971; Luce & Tukey, 1964; Michell, 1990; Thurstone, 1959; van der Linden, 1994; Wright, 1980, 1984, 1985). The separate analyses of the two instruments have left their items and steps in unknown relation to one another. Figure 5, showing the co-calibrated step and item relations, presents a reasonable picture of what might be expected, but it is as yet unknown whether the measures from this co-calibration will be equivalent with one another, or with the measures from the separate analyses.

Second Separation of the Instruments: Comparison

Anchoring the item and the step difficulty estimates at their co-calibrated values in two additional analyses, one for each instrument, produced two new sets of 113 measures. The only significant changes in the measures from the original, unanchored analyses related to their means and standard deviations. These became more similar, with the PF10 statistics becoming more like the PFS's, as expected, with the PFS mean and standard deviation at .27 and 2.2, and the PF10 mean and standard deviation at .14 and 2.5. The paired-sample t is .95 (p=.34).

The measures' correlation remains at .80, and the disattenuated correlation remains at .86, as neither instrument's person separation reliability changed. A scatterplot of the measures from the anchored analyses takes the same shape as the plots in Figures 6 and 7, and so is not reproduced here. The measures from the anchored PF10 items correlate 1.0 with the the measures from the initial unanchored analysis; the measures from the anchored and unanchored PFS analyses also correlate 1.0.

The equating method used on the measures from the initial, separate, analyses was applied to the measures from the anchored analyses in order to reduce the average difference in the measures to zero, and to remove the small remaining difference in the standard deviations. The resulting paired-sample t was .00 (p=1.0). The average measure from each instrument was .27, with a standard deviation of 2.2 and standard error of the mean equal to .21. The average difference in the measures was .0000, with a standard deviation of 1.4, and a standard error of the mean equal to .13. The scatterplot of the measures from the two instruments remained identical in

its pattern with Figures 6 and 7, and the correlation remained .80. Correlations were again 1.0 for each instrument's respective pairs of measures.

Tables 4 and 5 relate raw scores to equated measures for each instrument. The score-measure relation shown in these tables is taken from the equating derived from the BIGSTEPS co-calibration analysis. Direct use of each table requires complete data from the instrument in question. A score summed from incomplete data can be related to the common metric by finding the score's proportion of the maximum possible score, and relating that proportion to the proportions implicit in the table. For instance, a score of 18 on the PF10, based on data from only 9 items, would represent a proportion of 18/27, or 2/3, given a maximum score of 3 on each of the 9 items. A score of 20 from all 10 items also represents a 2/3 proportion. Table 4 shows that a score of 20 is equivalent to a measure of .94, with an error of .55. The use of fewer items implies a higher error of measurement, so the errors shown in Tables 4 and 5 are under estimated for measures based on incomplete data.

DISCUSSION

The best person separation reliability of .90 obtained for the PF10 in the initial analysis is identical with that obtained (Fisher, et al., 1995) for the same set of items in the reference data set published in the HSQ 2.0 manual (Radosevich, et al., 1994). Rasch generalizability theory (Linacre, 1993) predicts that a 10-item scale with three rating categories capable of producing a standard deviation of 1.8 logits in the person measures, as the PF10 does, will attain a reliability of about .90, not taking statistical inconsistencies in the data into account. The HSQ reference data achieve this reliability, probably as a result of being selected for their variability and consistency from a larger database, much as the data analyzed in this study were. Omitting model fit from the error in the present analysis boosts the PF10 person separation statistic from 3.0 to 3.4 and the person separation reliability from .90 to .92. Doing the same boosts the PFS person separation from 5.7 to 6.2, and person separation reliability remains unchanged at .97.

The differences in the scales' calibration errors can be explained by the fact that the PF10 has only three rating categories, where the PFS has six, causing the PF10 to have 86% greater calibration error, and 175% greater measurement error. Although overall model fit is satisfactory for both instruments, the PF10's data quality and reliability is generally poorer than the PFS's. In order to maximize person separation reliability, 82 cases

Table 4
PF10 Scores to Equated Measures for Complete Data

Score	Measure	S.E.
10	-4.56E	1.45
11	-3.64	1.14
12	-2.53	.86
13	-1.83	.72
14	-1.30	.64
15	-.86	.60
16	-.47	.57
17	-.10	.56
18	.25	.56
19	.59	.55
20	.94	.55
21	1.27	.55
22	1.61	.55
23	1.97	.56
24	2.32	.56
25	2.69	.58
26	3.10	.62
27	3.57	.68
28	4.18	.79
29	5.10	1.04
30	5.89E	1.38

Table 5
PFS Scores to Equated Measures for Complete Data

Score	Measure	S.E.	Score	Measure	S.E.	Score	Measure	S.E.
29	-7.39E	1.41	68	-1.74	.26	107	.96	.28
30	-6.70	1.00	69	-1.67	.26	108	1.04	.28
31	-6.00	.71	70	-1.60	.26	109	1.12	.28
32	-5.59	.59	71	-1.53	.26	110	1.20	.28
33	-5.28	.52	72	-1.46	.26	111	1.28	.29
34	-5.04	.47	73	-1.40	.26	112	1.36	.29
35	-4.84	.44	74	-1.33	.26	113	1.45	.29
36	-4.66	.41	75	-1.26	.26	114	1.54	.30
37	-4.50	.39	76	-1.19	.26	115	1.62	.30
38	-4.35	.37	77	-1.12	.26	116	1.71	.30
39	-4.22	.36	78	-1.06	.26	117	1.81	.31
40	-4.09	.35	79	-.99	.26	118	1.90	.31
41	-3.97	.34	80	-.92	.26	119	2.00	.31
42	-3.86	.33	81	-.86	.26	120	2.10	.32
43	-3.75	.32	82	-.79	.26	121	2.20	.32
44	-3.65	.32	83	-.72	.26	122	2.31	.33
45	-3.55	.31	84	-.66	.26	123	2.41	.33
46	-3.46	.31	85	-.59	.26	124	2.53	.34
47	-3.37	.30	86	-.52	.26	125	2.64	.34
48	-3.28	.30	87	-.45	.26	126	2.76	.35
49	-3.19	.29	88	-.39	.26	127	2.88	.35
50	-3.10	.29	89	-.32	.26	128	3.01	.36
51	-3.02	.29	90	-.25	.26	129	3.14	.37
52	-2.94	.29	91	-.18	.26	130	3.28	.38
53	-2.86	.28	92	-.12	.26	131	3.43	.38
54	-2.78	.28	93	-.05	.26	132	3.58	.39
55	-2.70	.28	94	.02	.26	133	3.74	.40
56	-2.62	.28	95	.09	.26	134	3.90	.42
57	-2.54	.28	96	.16	.26	135	4.08	.43
58	-2.47	.28	97	.23	.26	136	4.27	.44
59	-2.39	.27	98	.30	.27	137	4.47	.46
60	-2.32	.27	99	.37	.27	138	4.70	.48
61	-2.24	.27	100	.44	.27	139	4.94	.51
62	-2.17	.27	101	.51	.27	140	5.21	.54
63	-2.10	.27	102	.59	.27	141	5.53	.58
64	-2.03	.27	103	.66	.27	142	5.91	.65
65	-1.95	.27	104	.73	.27	143	6.40	.77
66	-1.88	.27	105	.81	.27	144	7.19	1.04
67	-1.81	.27	106	.88	.28	145	7.92E	1.45

were removed from the original sample size of 285 for the PF10, and 66 cases (20 percent fewer) were removed for the PFS. Even without removing any cases, the PFS's person separation (4.5, equivalent to a reliability of .95) was 50% greater than the PF10's best person separation (3.0, the reliability coefficient equivalent being .90).

The PFS's larger number of items make it possible to improve its targeting via adaptive administration of its items, so that patients need to respond only to the questions most relevant to their health care needs. Because the assessment and improvement of the PFS is based on Rasch's probabilistic measurement principles, missing data can be accounted for, meaning that, in addition to adaptive administration of its items, items can be added or deleted without compromising the unit of measurement, and without making old data incommensurable with new (Choppin, 1968; Choppin, 1976; Wright & Bell, 1984), opening the door to continuously improving the quality of the instrument (Holm & Kavanagh, 1985; Wright & Stone, 1979).

The LSU HSI PFS's items span a wider range overall and are more evenly distributed over the measurement continuum than the PF10's. One effect of the PFS's better person separation reliability relative to the PF10 is shown in Table 2 in the difference between the statistically distinct strata distinguished by the two instruments. The PF10's .90 reliability allows definition of 4.3 strata among the measures, providing enough statistical confidence to distinguish four levels of physical functioning.

The PFS, in contrast, with its .97 reliability and 7.9 strata, supports almost twice as many statistically significant distinctions. Whether the additional information provided by the PFS is useful will depend on the application. Marketing, and perhaps accreditation, needs might be met with less statistical power, but the research needed to determine the least meaningful unit of observation for treatment planning, program evaluation, and treatment effectiveness research has yet to be done. It may be that the PFS's .97 measurement reliability provides sufficient power to detect some small treatment effects within a particular range of physical functioning, but not within others.

Regarding the collapsing of the first and second PFS rating categories into a single category, perhaps it is impossible to make a quantitative distinction, or at least a consistent one, between Impossible and Very Difficult in the area of physical functioning. Another explanation for the lack of a distinct step is the lack of respondents with measures in this very low range. The PFS will remain a six-category instrument until the utility of

the distinction between Impossible and Very Difficult functioning is more conclusively tested.

CONCLUSIONS

This study shows that the SF36 PF10 and LSU HSI PFS measure the same physical functioning variable, and that they can do so in the same quantitative unit. As demands for accountability, outcome comparability, and a consumer-oriented focus continue to increase in health care, so will the need for sample-free and scale-free units of measurement. Insofar as data from different instruments all intended to measure the same variable meet the requirements for measurement stated in a Rasch model, and so fit that model, these instruments can, in principle, be equated to measure in a single quantitative unit. It may be that universally-accepted metrics for each of the health-related variables measured with rating scales are on the horizon. Such metrics will require dissemination of standard instrument quality evaluation and equating procedures. An effort aimed at developing these procedures has begun, under the auspices of the American Society for Testing and Materials (ASTM) Committee E-31 on the Content and Structure of the Electronic Health Record (Fisher, 1996, 1997b).[1]

This study is offered as an introductory tutorial in the exploration and comparison of rating scale instruments' measurement properties. It is hoped that the methods and the metric described here will be critically applied and extended so that, as new and improved measures of physical functioning are introduced, this particular Tower of Babel will begin to be dismantled. Perhaps, with effort and care, the cacophonous bedlam of incommensurable, scale-, and sample-dependent raw scores can be transformed into a harmonious chorus through the application of scale-free measurement principles.

FOOTNOTE

[1]All interested parties are invited to contribute to the development of these standards. Contact Teresa Cendroska (610/832-9500) at ASTM for more information.

ACKNOWLEDGMENTS

Portions of this paper were presented at the 1997 meetings of the American Medical Informatics Association in San Jose, CA; the American Educational Research Association in Chicago, IL; the International Objective Measurement Workshops, at the University of Chicago; and the American Public Health Association in Indianapolis. Thanks to Benjamin D. Wright, J. Michael Linacre, Richard M. Smith, Theo L. Dawson, and the reviewers for their comments on and contributions to this paper. Thanks also to James H. Diaz, MD, for his ongoing support of this work.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 357-374.

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.

Choppin, B. (1968). An item bank using sample-free calibration. *Nature, 219,* 870-872.

Choppin, B. (1976). Recent developments in item banking. In D. N. DeGruitjer, M & L. J. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement* (pp. 233-245). New York: Wiley.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical.* New York: Russell Sage Foundation.

Fisher, W. P., Jr. (1996, October). *Rating scale measurement standards relevant to ASTM 1384 on the content and structure of the electronic health record.* Unpublished paper presented at a semi-annual meeting of the ASTM Committee E31 on the Electronic Health Record, Washington, DC.

Fisher, W. P., Jr. (1997a). Physical disability construct convergence across instruments: Towards a universal metric. *Journal of Outcome Measurement, 1*(2), 87-113.

Fisher, W. P., Jr. (1997b, June). What scale-free measurement means to health outcomes research. *Physical Medicine & Rehabilitation State of the Art Reviews,* 11(2), 357-373.

Fisher, W. P., Jr., Harvey, R. F., & Kilgore, K. M. (1995). New developments in functional assessment: Probabilistic models for gold standards. *NeuroRehabilitation, 5*(1), 3-25.

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., 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)* (pp. 109-127). Gaithersburg, Maryland: Aspen Publications, Inc.

Fisher, W. P., Jr., Marier, R. L., & Hunter, S. (1995). *Comparing probabilistic calibrations of two health status instruments: The LSU HSI and the HSQ 2.0.* HCMS Measurement Research Reports, no. 1. New Orleans: LSUMC Department of Public Health & Preventive Medicine.

Fisher, W. P., Jr., & Wright, B. D. (1994). Introduction to probabilistic conjoint measurement theory and applications. *International Journal of Educational Research, 21*(6), 559-568.

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.

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., 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.

Holm, K. & Kavanagh, J. (1985). An approach to modifying self-report instruments. *Research in Nursing and Health, 8,* 13-18.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement. Volume 1: Additive and polynomial representations.* New York: Academic Press.

Linacre, J. M. (1993). Rasch generalizability theory. *Rasch Measurement Transactions, 7*(1), 283-284.

Linacre, J. M. & Wright, B. D. (1997). *A User's Guide to BIGSTEPS Rasch-Model Computer Program.* Chicago: MESA Press.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new kind of fundamental measurement. *Journal of Mathematical Psychology, 1*(1), 1-27.

Ludlow, L. H. & Haley, S. M. (1995, December). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement, 55*(6), 967-975.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Masters, G. N. (1985, March). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82.

McHorney, C. A., Haley, S. M., & Ware, J. E. (1997). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology, 50*(4), 451-461.

Michell, J. (1990). *An Introduction to the Logic of Psychological Measurement.*

Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*(1), 63-75.

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.

Schumacker, R. E. (1996, Spring). Disattenuating correlation coefficients. *Rasch Measurement Transactions, 10*(1), 479.

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.

Thurstone, L. L. (1959). *The measurement of values.* Chicago: University of Chicago Press, Midway Reprint Series.

van der Linden, W. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 3-24). Norwood, NJ: Ablex Publishing Corporation.

Wilson, M. (1994). Comparing attitude across different cultures: Two quantitative approaches to construct validity. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice, Volume 2* (pp. 271-294). Norwood, NJ: Ablex.

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. (1995, Summer). Which standard error? *Rasch Measurement Transactions, 9*(2), 436-437.

Wright, B. D. & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*(4), 331-345.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA Press.

Zhu, W. (1996). Test equating: What, why, how? *Research Quarterly for Exercise and Sport,* submitted for publication.

APPENDIX A

LSU HSI: PHYSICAL HEALTH STATUS

Form #: ☐☐☐☐

Instructions: Please tell us about your health by rating your ability to get around and do things. Use a pencil or pen to answer each question by putting a check mark or X in the relevant box. *Please keep your marks within the selected box.* **Thank you very much!**

How difficult is it for you to:	Very Impossible	Difficult	Difficult	Manageable	Easy	Very Easy
1. Get around in your home?	☐	☐	☐	☐	☐	☐
2. Get up out of a chair?	☐	☐	☐	☐	☐	☐
3. Wash your face?	☐	☐	☐	☐	☐	☐
4. Wash your feet?	☐	☐	☐	☐	☐	☐
5. Use the toilet?	☐	☐	☐	☐	☐	☐
6. Dress yourself?	☐	☐	☐	☐	☐	☐
7. Use a wheelchair in public places?	☐	☐	☐	☐	☐	☐
8. Cook meals for guests?	☐	☐	☐	☐	☐	☐
9. Do housework?	☐	☐	☐	☐	☐	☐
10. Walk up a flight of stairs?	☐	☐	☐	☐	☐	☐
11. Walk a block?	☐	☐	☐	☐	☐	☐
12. Bathe yourself?	☐	☐	☐	☐	☐	☐
13. Walk without being afraid of falling?	☐	☐	☐	☐	☐	☐
14. Move around in bed?	☐	☐	☐	☐	☐	☐
15. Walk half a mile?	☐	☐	☐	☐	☐	☐
16. Prepare light meals?	☐	☐	☐	☐	☐	☐
17. Exercise hard for half an hour?	☐	☐	☐	☐	☐	☐
18. Walk two miles?	☐	☐	☐	☐	☐	☐
19. Feed yourself?	☐	☐	☐	☐	☐	☐
20. Go grocery shopping?	☐	☐	☐	☐	☐	☐
21. Pick up a young child?	☐	☐	☐	☐	☐	☐
22. Put in a day's office work?	☐	☐	☐	☐	☐	☐
23. Do 8 hours physical labor?	☐	☐	☐	☐	☐	☐
24. Drive a car?	☐	☐	☐	☐	☐	☐
25. Go fishing?	☐	☐	☐	☐	☐	☐
26. Rearrange furniture in your home?	☐	☐	☐	☐	☐	☐
27. Dance for half an hour?	☐	☐	☐	☐	☐	☐
28. Carry a bag of groceries?	☐	☐	☐	☐	☐	☐
29. Reach high on a shelf for something?	☐	☐	☐	☐	☐	☐

APPENDIX B

THE MOS 36-ITEM SHORT-FORM HEALTH SURVEY (SF-36)

Form #: □□□□

Instructions: This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. Use a pencil or pen to answer each question by putting a check mark or X in the relevant box. *Please keep your marks within the selected box.* **Thank you very much!**

	Excellent	Very Good	Good	Fair	Poor
1. In general, would you say your health is:	☐	☐	☐	☐	☐

	Much better than 1 year ago	Somewhat better than 1 year ago	About the same	Somewhat worse than 1 year ago	Much worse than 1 year ago
2. **Compared to one year ago,** how would you rate your health in general **now?**	☐	☐	☐	☐	☐

3. The following items are about activities you might do during a typical day. Does **your health now limit you** in these activities?

If so, how much? (Check one box on each line.)	Yes, Limited a Lot	Yes, Limited a Little	No, Not Limited at All
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	☐	☐	☐
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	☐	☐	☐
c. Lifting or carrying groceries	☐	☐	☐
d. Climbing several flights of stairs	☐	☐	☐
e. Climbing one flight of stairs	☐	☐	☐
f. Bending, kneeling, or stooping	☐	☐	☐
g. Walking more than a mile	☐	☐	☐
h. Walking several blocks	☐	☐	☐
i. Walking one block	☐	☐	☐
j. Bathing or dressing yourself	☐	☐	☐

4. During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of your physical health?** (Check one box on each line.)

	Yes	No
a. Cut down on the amount of time you spent on work or other activities	☐	☐
b. Accomplished less than you would like	☐	☐
c. Were limited in the kind of work or other activities	☐	☐
d. Had difficulty performing work or other activities (for example, it took extra effort)	☐	☐

5. During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)? (Check one box on each line.)

	Yes	No
a. Cut down the amount of time you spent on work or other activities	☐	☐
b. Accomplished less that you would like	☐	☐
c. Didn't do work or other activities as carefully as usual	☐	☐

APPENDIX C

SPSS SYNTAX FOR READING BIGSTEPS PERSON MEASUREMENT FILES INTO SEPARATE DATA FILES

```
DATA LIST FILE 'BIGSTEPS.OUT1' FIXED
                / RECNO 2-6 MEASURE 7-14 (1) MEASSTAT 15-17
                  TEST 18-23 SCORE 24-30
                  ERROR 31-37 (1) MSINF 38-44 (2) STDINF 45-51 (2)
                  MSOUT 52-58 (2) STDOUT 59-65 (2) DISPLAC 66-72 (2)
                  PTBISER 73-79 (2).
EXECUTE.
*Note: the SPSS DATA LIST syntax will produce multiple meaningless warnings concerning
*character data in numeric fields if the four-line headers at the top of the BIGSTEPS
*files are not deleted before executing the statement. Manually delete these lines, or
*use the BIGSTEPS HLINES=N control variable to omit them from the output file.
VARIABLE LABELS RECNO 'RECORD NUMBER'
          MEASSTAT 'MEASURE TYPE'
          TEST '# ITEMS MEASURING' SCORE 'RAW RATINGS SUM'
          MEASURE 'LSU HSI MEASURES'
           ERROR 'ERROR (RANGE OF MEASURE)'
          MSINF 'MNSQ INFIT'
          STDINF 'STNDRDZD INFIT'
          MSOUT 'MNSQ OUTFIT'
          STDOUT 'STNDRDZD OUTFIT'.
VALUE LABELS MEASSTAT 1 'Estimated value' 2 'Anchored value' 0 'Extreme minimum'
               -1 'Extreme maximum' -2 'No available responses'
               -3 'Deleted by user' -4 'Combined w/ another item'.
MISSING VALUES MEASURE ERROR STDINF STDOUT SCORE TEST (0).
SET  BLANKS=SYSMIS BLANKS=SYSMIS UNDEFINED=WARN.
SORT CASES BY  recno (A) .
SAVE OUTFILE='BIGSTEPS1.SAV'  /COMPRESSED.
DATA LIST FILE 'BIGSTEPS.OUT2' FIXED
                / RECNO 2-6 MEASURE 7-14 (1) MEASSTAT 15-17
                 TEST 18-23 SCORE 24-30
                  ERROR 31-37 (1) MSINF 38-44 (2) STDINF 45-51 (2)
                  MSOUT 52-58 (2) STDOUT 59-65 (2) DISPLAC 66-72 (2)
                  PTBISER 73-79 (2).
EXECUTE.
*Add labels.
VARIABLE LABELS RECNO 'RECORD NUMBER'
          MEASSTAT 'MEASURE TYPE'
          TEST '# ITEMS MEASURING' SCORE 'RAW RATINGS SUM'
          MEASURE 'SF36 MEASURES'
          ERROR 'ERROR (RANGE OF MEASURE)'
          MSINF 'MNSQ INFIT'
          STDINF 'STNDRDZD INFIT'
          MSOUT 'MNSQ OUTFIT'
          STDOUT 'STNDRDZD OUTFIT'.
VALUE LABELS MEASSTAT 1 'Estimated value' 2 'Anchored value' 0 'Extreme minimum'
               -1 'Extreme maximum' -2 'No available responses'
               -3 'Deleted by user' -4 'Combined w/ another item'.
MISSING VALUES MEASURE ERROR STDINF STDOUT SCORE TEST (0).
SET BLANKS=SYSMIS BLANKS=SYSMIS  UNDEFINED=WARN.
SORT CASES BY  recno (A) .
SAVE OUTFILE='BIGSTEPS2.SAV' /COMPRESSED.
```

APPENDIX D

SPSS SYNTAX FOR COMBINING DATA FILES AND COMPARING MEASURES

```
GET
  FILE='BIGSTEPS1.SAV'.
EXECUTE.

MATCH FILES /TABLE=*
 /RENAME  measure=1STmeas measstat=1STmsta test=1STtest score=1STscor
  error=1STerro msinf=1STmsin stdinf=1STstin msout=1STmout
  stdout=1STsout displac=1STdisp ptbiser=1STptbi
 /FILE='BIGSTEPS2.SAV'
 /RENAME measure=2NDmeas measstat=2NDmsta test=2NDtest error=2NDerro
  msinf=2NDmsin stdinf=2NDstin msout=2NDmout stdout=2NDsout
  displac=2NDdisp ptbiser=2NDptbi score=2NDscor
 /BY recno.
EXECUTE.

SAVE OUTFILE='BIGSTEPS1&2.SAV'  /COMPRESSED.

USE ALL.
COMPUTE filter_$=(1STmsta = 1 & 2NDmsta = 1 & 1STtest >= [AT LEAST HALF THE ITEMS]
'+
'& 2NDtest >= [AT LEAST HALF THE ITEMS])'.
VARIABLE LABEL filter_$ '1STmsta = 1 & 2NDmsta = 1 & 1STtest >= [AT LEAST HALF THE
ITEMS] & 2NDtest >= [AT LEAST HALF THE ITEMS] (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

T-TEST
  PAIRS= 1STmeas  WITH 2ndmeas (PAIRED)
  /CRITERIA=CIN(.95)
  /FORMAT=LABELS
  /MISSING=ANALYSIS.
```

Journal of Outcome Measurement®
Volume 1
Author and Title Index

CONTRIBUTOR INFORMATION

Content: *Journal of Outcome Measurement* publishes refereed scholarly work from all academic disciplines relative to outcome measurement. Outcome measurement being defined as the measurement of the result of any intervention designed to alter the physical or mental state of an individual. The *Journal of Outcome Measurement* will consider both theoretical and applied articles that relate to measurement models, scale development, applications, and demonstrations. Given the multi-disciplinary nature of the journal, two broad-based editorial boards have been developed to consider articles falling into the general fields of Health Sciences and Social Sciences.

Book and Software Reviews: The *Journal of Outcome Measurement* publishes only solicited reviews of current books and software. These reviews permit objective assessment of current books and software. Suggestions for reviews are accepted. Original authors will be given the opportunity to respond to all reviews.

Peer Review of Manuscripts: Manuscripts are anonymously peer-reviewed by two experts appropriate for the topic and content. The editor is responsible for guaranteeing anonymity of the author(s) and reviewers during the review process. The review normally takes three (3) months.

Manuscript Preparation: Manuscripts should be prepared according to the *Publication Manual of the American Psychological Association* (4th ed., 1994). Limit manuscripts to 25 pages of text, exclusive of tables and figures. Manuscripts must be double spaced including the title page, abstract, text, quotes, acknowledgments, references, and appendices. On the cover page list author name(s), affiliation(s), address(es), telephone number(s), and electronic mail address(es). On the second page include a 100 to 150 word abstract. Place tables on separate pages. Include photocopies of all figures. Number all pages consecutively.

Authors are responsible for all statements made in their work and for obtaining permission from copyright owners to reprint or adapt a table or figure or to reprint a quotation of 500 words or more. Copies of all permissions and credit lines must be submitted.

Manuscript Submission: Submit four (4) manuscript copies to Richard M. Smith, Editor, *Journal of Outcome Measurement*, Rehabilitation Foundation Inc., P.O. Box 675, Wheaton, IL 60189 (e-mail:JOMEA@rfi.org). Prepare three copies of the manuscript for peer review by removing references to author(s) and institution(s). In a cover letter, authors should indicate that the manuscript includes only original material that has not been previously published and is not under review elsewhere. After manuscripts are accepted authors are asked to submit a final copy of the manuscript, original graphic files and camera-ready figures, a copy of the final manuscript in WordPerfect format on a 3 ½ in. disk for IBM-compatible personal computers, and sign and return a copyright-transfer agreement.

Production Notes: manuscripts are copy-edited and composed into page proofs. Authors review proofs before publication.

SUBSCRIBER INFORMATION

Journal of Outcome Measurement is published four times a year and is available on a calendar basis. Individual volume rates are $35.00 per year. Institutional subscriptions are available for $100 per year. There is an additional $24.00 charge for postage outside of the United States and Canada. Funds are payable in U.S. currency. Send subscription orders, information requests, and address changes to the Subscription Services, Rehabilitation Foundation, Inc. P.O. Box 675, Wheaton, IL 60189. Claims for missing issues cannot be honored beyond 6 months after mailing date. Duplicate copies cannot be sent to replace issues not delivered due to failure to notify publisher of change of address.

 ISSN 1090-655X.