CROSS-CULTURAL VALIDITY OF FUNCTIONAL INDEPENDENCE MEASURE ITEMS IN STROKE: A STUDY USING RASCH ANALYSIS

Åsa Lundgren-Nilsson,[1] Gunnar Grimby,[1] Haim Ring,[2] Luigi Tesio,[3] Gemma Lawton,[4] Anita Slade,[4] Massimo Penta,[5] Maria Tripolski,[2] Fin Biering-Sørensen,[6] Jane Carter,[7] Crt Marincek,[8] Suzanne Phillips,[7] Anna Simone[3] and Alan Tennant[4]

From the [1]Rehabilitation Medicine, Sahlgrenska Academy at Göteborg University, Göteborg, Sweden, [2]Loewenstein Hospital, Rehabilitation Centre, Raanana, Tel Aviv University School of Medicine, Israel, [3]Unit of Rehabilitation Research, Istituto Auxologico Italiano, Milan, Italy, [4]Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds, UK, [5]Unité de Réadaption et de Médecine Physique, Université Catholique de Louvain, Bruxelles, Belgium, [6]Clinic for Para- and Tetraplegia, The Neuroscience Centre, Rigshospitalet, Copenhagen University Hospital, Denmark, [7]The Bath Head Injury/Neuro-Rehabilitation Unit, Royal National Hospital for Rheumatic Diseases, National Health Services Trust, Bath, UK and [8]University Institute for Rehabilitation Ljubljana, Ljubljana, Slovenia

Objective: **To analyse cross-cultural validity of the Functional Independence Measure (FIM㊊) in patients with stroke using the Rasch model.**

Settings: **Thirty-one rehabilitation facilities within 6 different countries in Europe.**

Participants: **A total of 2546 in-patients at admission, median age 63 years.**

Methods: **Data from the FIM㊊ were evaluated with the Rasch model, using the Rasch analysis package RUMM2020. A detailed analysis of scoring functions of the 7 categories of the FIM items was undertaken prior to testing fit to the model. Categories were re-scored where necessary. Analysis of Differential Item Functioning was undertaken in pooled data for each of the FIM motor and social-cognitive scales, respectively.**

Results: **Disordered thresholds were found on most items when using 7 categories. Fit to the Rasch model varied between countries. Differential Item Functioning was found by country for most items. Adequate fit to the Rasch model was achieved when items were treated as unique for each country and after a few country-specific items were removed.**

Conclusion: **Clinical collected data from FIM for patients with stroke cannot be pooled in its raw form, or compared across countries. Comparisons can be made after adjusting for country-specific Differential Item Functioning, though the adjustments for Differential Item Functioning and rating scales may not generalize to other samples.**

Key words: measurement, cross-cultural validity, Rasch analysis, rehabilitation, stroke, Functional Independence Measure.

J Rehabil Med 2005; 37: 23–31

Correspondence address: Åsa Lundgren Nilsson, Department of Rehabilitation Medicine, Sahlgrenska University Hospital, SE-413 45 Göteborg, Sweden.
E-mail: asa.lundgren-nilsson@rehab.gu.se

Submitted December 12, 2002; accepted March 24, 2004

INTRODUCTION

In order to assess the results of rehabilitation, which is seen as an increasingly important part of good clinical practice, the use of standardized, reliable and valid measures is becoming essential. However, if we wish to identify the most effective and efficient treatment modalities within diagnostic groups at the regional, national and international level, it follows that we require measures that serve this purpose. Ideally, such measures should have both a solid conceptual basis, e.g. being based upon the International Classification of Functioning, Disability and Health (ICF) (1), and adequate psychometric properties.

Some attempt at standardization of measurement in rehabilitation has been made in North America. The Functional Independence Measure (FIM^{TM}) (2), a measure of disability, is used across a wide range of conditions and in a wide range of situations in rehabilitation. The items are mainly within the dimension of Activity limitation according to ICF. There is an extensive body of literature, in general supporting reliability and validity of FIM^{TM}, as also sensitivity, which, however, may vary due to the population being assessed (3).

A methodological approach to measurement has emerged based on Rasch analysis (4), recently reviewed as a tool for rehabilitation research (5). This advance, in the understanding of the scientific basis of measurement, has led to a psychometric re-appraisal of existing measures (6, 7). In particular, fitting data to the Rasch model allows for a detailed examination of the internal construct validity of the measure, including the ordering of categories, unidimensionality, and whether or not items work in the same way across groups, including country (Differential Item Functioning, DIF).

The requirement for outcome measures to work in a consistent manner at the European level would contribute to standardization of measurement at the European level. To facilitate these

The Editor-in-Chief has not had the responsibility for this article (which originated from his own department). It has been handled fully by a co-editor, who has made the decision to accept the article.

ISSN 1650–1977
DOI 10.1080/16501970410032696

objectives a project called “European Standardisation of Outcome Measurement in Rehabilitation” (Pro-ESOR), was established under the Framework IV programme of the European Commission (EC). In this project, commonly used outcome measures were identified in 9 diagnostic groups from 416 facilities offering rehabilitation in European countries (8). The most commonly used instruments in 6 of the diagnostic groups (3 had no common measures) were chosen for further analyses of their internal construct validity and cross-cultural validity using Rasch analysis. In patients with stroke, the FIM^{TM}, and the Barthel index (different modifications) were the most widely used instruments.

The FIM^{TM} in stroke patients with data from 31 clinical facilities within 6 European countries is the subject of the present paper. It will be demonstrated that pooling of raw score data across countries is not valid but, after necessary adjustments, comparison between different countries can be made. However, within the recorded DIF between countries, non-country specific factors may exist but are not analysed further in the present study.

METHODS

Functional Independence Measure

The FIM^{TM} consists of 13 motor and 5 social-cognitive items, assessing self-care, sphincter management, transfer, locomotion, communication, social interaction and cognition (2). It uses a 7-level scale anchored by extreme rating of total dependence as 1 and complete independence as 7; the intermediate levels are: 6 modified independence, 5 supervision or set-up, 4 minimal contact assistance or the subject expends >75% of the effort, 3 moderate assistance or the subjects expends 50–74% of the effort, and 2 maximal assistance or the subject expends 25–49% of the effort. Although developed originally as an 18-item scale, it has been shown that there are 2 scales, a 13-item motor and a 5-item social-cognitive scale (9). Thus, in the present study these will be referred to as FIM motor and FIM social-cognitive items or scales respectively.

Rasch analysis

The Rasch model is used as a methodological basis for examining the internal construct validity of the scale; its scaling properties and, where appropriate, through analysis of DIF, its cross-cultural validity (10). It is a unidimensional measurement model, which assumes that the easier the item the more likely it will be passed, and the more able the person, the more likely they will pass an item compared with a less able person (6). The original dichotomous model has been extended to accommodate polytomous responses, Partial Credit Model (PCM), which was used in this paper (11). When the observed response pattern coincides with or does not deviate too much from the expected response pattern then the items constitute a valid measure. If there are no associations in the residuals derived from the difference between observed values and model expectations (local independence) this further supports the claim of unidimensionality (12).

The PCM involves a threshold *k*, which represents the equal probability point between any 2 adjacent categories within an item. Threshold estimates should be correctly ordered if the categories are being assigned in the intended way. Consequently, this can be empirically verified against the model expectation and deviations identified where the categories fail to express an increasing level of the trait (disordered thresholds) (7). Where categories are disordered with respect to the underlying trait (i.e. where a 2 represents less of the trait than a 1 when it should represent more) it is necessary to collapse adjacent categories as part of the ongoing Rasch analysis (7). Once disordered thresholds are removed, fit of data to the Rasch model is assessed by examining deviations from model expectations, including DIF.

Differential Item Functioning

Early published work on Rasch analysis in rehabilitation explored issues of unidimensionality and scaling properties (13) and this has remained a central theme to date (14–16). However, Rasch analysis allows for much more than an empirical test for unidimensionality. Within the framework of Rasch measurement, cross-cultural validity can be examined from a comparison of item performance *after* the requirements for Rasch measurement have been met (17). The basis of the approach to the analysis of cross-cultural validity lies in the item response function, the S-shaped trace of the proportion of individuals at the same ability level who, in the dichotomous case, answer a given item correctly (or can do a particular task). Under the requirement that the ability under consideration is unidimensional, if the item measures the same ability across groups then, except for random variations, the same curve is found irrespective of the nature of the group for whom a function is plotted (18). Items that do not yield the same item response function for 2 or more groups display DIF and are violating the requirement of unidimensionality (18). Consequently, it is possible to examine whether or not a scale works in the same way by contrasting the response function for each item across cultures.

Analytical strategy

Under the assumption that the distances between thresholds vary across items, the PCM was used (6). This assumption was formally tested in the present study by graphical representation of thresholds and by a log-likelihood test.

Data from each country were initially analysed separately and then pooled to assess cross-cultural differences. Where the categories were found to be not working as intended for an item (i.e. disordered thresholds) they were collapsed. This was done uniquely for each country first, then again for the pooled data. Although disordered thresholds are identified explicitly in the RUMM2020 programme, decisions still need to be taken as how best to collapse categories. Initially a visual examination of the way in which categories were working suggested possible ways to collapse categories (Fig. 1). For example, Fig. 1 shows the “eating item” and that categories 1 and 2 (the categories always start from 0, e.g. equivalent to 1 in the FIM scale) do not appear to be functioning in the correct manner. At no time are categories 1 and 2 more probable than category 0, and thus these would be collapsed together. Where alternative collapsing strategies seem possible, that pattern which produces the best fit for the item is chosen.

Following this, the data are refitted to the Rasch model to determine overall fit and how well each item fits the model. Three overall fit statistics are considered initially. Fit of the items is given by a standardized fit statistic for persons and items (mean 0, SD of 1 where the data fit the model perfectly) and a chi-square (χ^2) Item-Trait interaction statistic to determine scale invariance and which should indicate non-significant deviation from the model. Perfect fit indicates that the hierarchical ordering of the items remains exactly the same at different levels of the underlying trait. This is calculated by summing all the chi-square values for each of the individual items and calculating the significance value using the summated degrees of freedom. In addition to these overall fit statistics a Person Separation Index, similar to Cronbach’s α, and indicates the degree to which the scale can separate patients into discrete groups. A value of 0.8 is the minimum required to differentiate 2 groups (19).

Individual item-fit statistics are considered, both as residuals (a summation of individual person and item deviations, and usually acceptable within the range ±3.0 (6)), and/or as a chi-square statistic, reflecting the deviation from the model by groups of people defined by their ability level (called class intervals in RUMM2020) and requiring a non-significant chi-square i.e. >0.05, with appropriate adjustment for repeated tests (20).

Misfit of items indicates a lack of the expected probabilistic relationship between the item and other items in the scale. One potential source of this lack of fit is DIF. Thus analysis of DIF was undertaken first on individual country data where age and gender were entered as “person factors” for DIF analysis. Subsequently, data were pooled for analysis

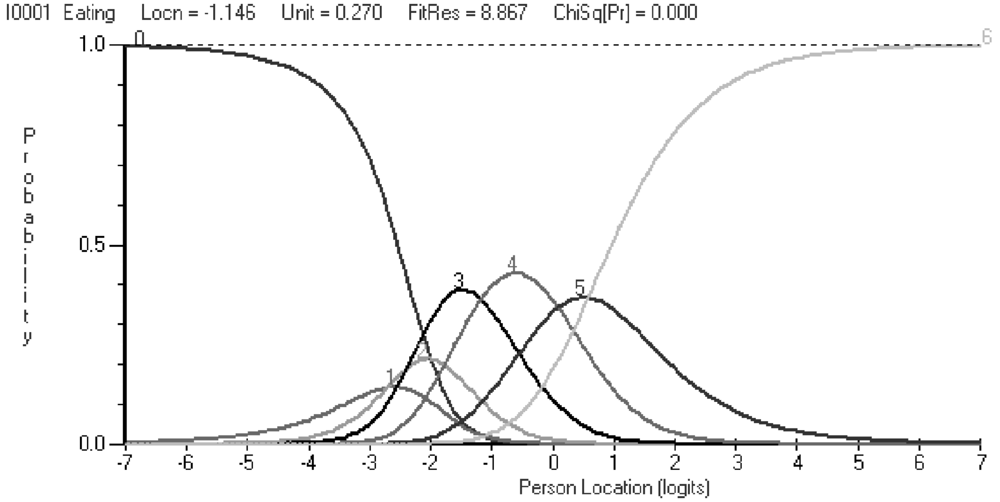

Fig. 1. Category probability curves of the FIM Motor "eating" item.

of DIF by country. Where data were pooled across countries and some but not all items were found to display DIF, adjustments were made to allow items with DIF to vary by group. The approach is an iterative "top-down purification" approach in that a requirement for identifying DIF is a baseline set of "pure" items (21, 22). Consequently a biased item with the poorest fit to the Rasch model is removed first and the procedure repeated until an unbiased and scalable subset of items is identified (23). The rejected items are then re-introduced to test the results. To do this, the items that display DIF are rendered unique to the group that display DIF.

The statistical test used for detecting DIF is an ANOVA of the person-item deviation residuals with person factors (e.g. country) and class intervals (e.g. Group along the trait) as factors. Two types of DIF can be identified – uniform and non-uniform DIF. With the former, there is a (constant) difference between groups (ANOVA main effect) and with the latter the difference varies across the trait (ANOVA interaction effect).

For example, suppose that France displayed DIF for an item, compared with Belgium and Sweden (*post hoc* tests for the ANOVA identify where the difference(s) lay). In this case, 2 nation-specific items would be formed, with 1 item for France (and the responses for patients from Belgium and Sweden entered as structural missing values) and 1 item for Belgium and Sweden together (with the responses for French patients entered as structural missing values). Those items without DIF for country acted as links in the calibration. Thus the item difficulty is allowed to vary across countries (23). Fit is again reassessed and items still displaying misfit to the model are removed and do not contribute to the person estimate (the Rasch model can easily estimate item and person parameter with missing values).

Finally, on this expanded data set (i.e. original plus split items), person-item deviation residuals are examined by Principal Components Analysis (PCA) for associations which may be indicative of the breach of the assumptions of local independence. The absence of such associations, taken with adequate fit to the Rasch model, support unidimensionality of the construct.

The Rasch analysis was undertaken with the RUMM2020 package (24). Due to the number of tests of fit undertaken (e.g. 13 for each item in the motor FIM, and much more when items were split for DIF) Bonferroni corrections (20) were applied giving a significant value of 0.008 for motor FIM and 0.002 for social-cognitive FIM.

These analyses allow for a series of measurement quality parameters to be identified, which are described briefly below:

- The *number of disordered thresholds* is the proportion of items over the total number in the scale that has ordered response categories (range 0–1).
- *Unidimensionality* is the proportion of items over the total number in the scale that fit the underlying latent trait after re-scoring (range 0–1).
- *Range of Measurement* is the ratio between the number of patients with extreme scores, maximum or minimum, and the total number of patients analysed (range 0–1). This gives the combined floor and ceiling effect.
- *Person Separation Reliability (PSR)* is the ratio between the patient true measure variance (expressed by the standard deviation of patient measures corrected for measurement error) and the observed (true + error) measure variance (6).
- *Invariance of scale* is the number of items displaying the absence of DIF (after re-scoring) by age and gender and, for the pooled data, by country.

Patients and settings

Initially a postal questionnaire was distributed to health professionals working in units providing rehabilitation during November 1998.

Table I. *Number of hospitals and age and gender characteristics of participants recruited into the study per country*

Country	Number of patients	% Females	Number of hospitals	Range of patients within hospital	Mean age	Median age	Number below median age for total sample	Number above median age for total sample	Range
Belgium	143	48	3	33–60	64	67	52	91	32–87
France	157	47	2	21–136	62	66	64	93	20–86
Israel	319	31	4	50–166	59	61	187	132	19–94
Italy	1046	41	11	13–248	68	69	299	747	15–95
Sweden	642	37	7	15–190	57	57	454	188	22–90
UK	239	45	4	24–78	53	55	178	61	16–99
Total	2546	40	31	13–248	62	63	1234	1312	15–99

Table II. *Summed raw score for motor and social-cognitive FIM items*

	FIM motor			FIM social-cognitive		
Country	Mean	Median	Range	Mean	Median	Range
Belgium	43	39	13–88	25	29	5–35
France	42	37	13–91	22	24	5–35
Israel	55	60	13–91	25	28	5–35
Italy	42	38	13–91	26	30	5–35
Sweden	65	73	13–91	26	29	5–35
UK	54	56	13–91	24	25	5–35
Total	51	49	13–91	24	25	5–35

In total, 418 surveys were returned, identifying a range of measures used across diagnosis. After selection of measures, 31 rehabilitation facilities within 6 countries agreed to contribute data from the FIM^{TM} in patients with stroke. The only data required from the facilities were ratings on the FIM^{TM}, age and gender. Admission data were collected from 2546 patients entering rehabilitation. For the study of DIF for age, the common median age of 63 years for the whole sample was used. The proportion of patients below and above that median value is also shown in Table I, and varied between the countries. The number of contributing hospitals per country is also shown together with number of observations. For the analysis of DIF by country, due to the disproportionate number of cases from some countries, a random sample of 150 cases was taken from each of Italy, Sweden, UK and Israel, and added to those cases from Belgium and France. This gave 895 cases for the cross-cultural analysis.

RESULTS

The ratings for FIM^{TM} motor and social-cognitive items respectively (Table II) indicate that the whole scale is used. There is variation in the mean and median values between the countries for the FIM^{TM} motor scores.

Scaling properties and fit within countries before re-scoring

The FIM^{TM} motor scale frequently displayed a lack of ordered responses (disordered thresholds), especially in the items “toileting”, “bladder” and “bowel management”, “transfer tub/shower”, “walk/wheelchair” and “stairs” (Table III). In the social-cognitive scale, however, this problem was largely absent, not being more than 1 disordered threshold in most countries and Belgium and Sweden displayed scoring categories as expected.

The quality parameters (Table IV) for the motor scale indicated that the proportion of ordered response categories varied across countries, with France, Sweden and UK exhibiting a low proportion of ordered categories in the motor scale, and Israel and Italy in the social-cognitive scale. The range of measurement was consistently good for the FIM^{TM} motor items, indicating only minor floor or ceiling effect, but not so for the social-cognitive items. All countries showed high person separation reliability, indicating that patients were well spread along the measurement construct defined by the FIM motor and social-cognitive items.

Scaling properties and fit within countries after re-scoring

For the FIM motor scale the “eating” item was the easiest in most countries (that is, independence would be achieved first), except France, where “bowel management” was the easiest (Table V). “Transfer tub/shower” and “stairs” were the most

Table III. *Items with disordered thresholds by country*

	Number of disordered thresholds					
Item	Belgium	France	Israel	Italy	Sweden	UK
Motor items						
Eating	2	2	2	2	3	2
Grooming		3			2	1
Bathing		3			2	
Dressing upper body		2			1	
Dressing lower body		3	1			
Toileting	2	4	3	4	2	3
Bladder management	4	4	5	5	5	5
Bowel management	5	4	5	4	4	5
Transfer bed	2					2
Transfer toilet	2	3				3
Transfer tub/shower	3	4	1	2	3	2
Walk/wheelchair		3	3	3	3	4
Stairs	4	4	3	3	4	3
Total number of items with disordered thresholds	8	12	8	7	10	10
Social-cognitive items						
Comprehension			1			
Expression				3		3
Social interaction			1			2
Problem solving		3	2	2		
Memory		4	1	3		
Total number of items with disordered thresholds	0	2	4	3	0	2

Table IV. *Summary of fit characteristics at individual country level*

Measure	Belgium	France	Israel	Italy	Sweden	UK
Motor items						
Number of patients	135	154	294	1046	642	239
Order in Response Categories	0.38	0.08	0.38	0.46	0.23	0.23
Unidimensionality	0.92	0.92	0.76	0.69	0.62	1.00
Range of Measurement	0.04	0.02	0.07	0.06	0.15	0.07
Person Separation Reliability	0.966	0.953	0.975	0.963	0.968	0.968
Social-cognitive items						
Number of patients	135	154	294	1046	642	239
Order in Response Categories	1.00	0.60	0.20	0.40	1.00	0.60
Unidimensionality	1.00	1.00	0.20	0.80	1.00	1.00
Range of Measurement	0.27	0.16	0.28	0.32	0.19	0.17
Person Separation Reliability	0.967	0.949	0.964	0.959	0.930	0.948

difficult items. For the FIM social-cognitive scale the "problem solving" item was always the most difficult item, but the easiest item varied from country to country (Table V).

After re-scoring, only data from the UK showed fit to the Rasch model for both individual and overall fit (Item-Trait Interaction) of the items, in both motor and social-cognitive scales (Table VI). Data from France also showed acceptable Item Trait Interaction, but inadequate fit of the "eating" item. Belgium was similar, but with a significant overall chi-square interaction. The "grooming" and "toileting" items showed misfit in Italy, as did "bowel management" in Sweden, both of which showed significant chi-squared values. For the social-cognitive items there was no significant misfit at either the individual item, or total scale level.

In summary, at the individual country level, fit to the Rasch model varied for the FIM motor scale, with all items meeting model expectations in the UK, marginally less so in France and Belgium, and less so in Italy, Israel and Sweden. The FIM social-cognitive scale fit the Rasch model in all countries except Israel, where only 1 in 5 items, after re-scoring, showed fit to the Rasch model (not shown). Both scales were free of DIF by gender in all countries, and DIF by age was found only in Sweden with a DIF age proportion of 0.77 in motor items and 0.80 in social-cognitive items.

Pooled data and cross-cultural validity

When data were pooled only 5 of the 13 motor items had ordered thresholds (categories) ("bathing", "dressing upper body", "dressing lower body", "transfer bed" and "transfer toilet"). After re-scoring, the number of categories used varied for the motor items between 2 and 7 (Table VII). Only 1 of the social-

Table V. *Relative location of FIM motor and social-cognitive items within each countries following re-scoring*

	Belgium	France	Israel	Italy	Sweden	UK	Pooled
Motor items							
Eating	1	3	1	1	1	1	1
Grooming	6	6	3	5	3	3	5
Bathing	8	11	8	9	11	8	9
Dressing upper body	9	10	7	7	5	5	8
Dressing lower body	11	7	10	10	9	9	10
Toileting	10	8	9	3	8	12	3
Bladder management	3	2	4	4	4	4	4
Bowel management	2	1	2	2	2	2	2
Transfer bed	4	4	5	6	6	6	6
Transfer toilet	5	5	6	8	7	7	7
Transfer tub/shower	13	13	11	12	10	11	12
Walk/wheelchair	7	9	12	11	12	10	11
Stairs	12	12	13	13	13	13	13
Overall range of measure values	−1.44 to 2.43	−2.44 to 2.78	−2.5 to 2.84	−5.20 to 2.74	−1.50 to 3.21	−1.51 to 2.48	−4.59 to 2.99
Social-cognitive items							
Comprehension	1	3	1	1	2	2	1
Expression	2	2	3	4	3	4	3
Social interaction	3	1	4	2	1	1	2
Problem solving	5	5	5	5	5	5	5
Memory	4	4	2	3	4	3	4
Overall range of measure values	−0.63 to 0.837	−0.834 to 1.649	−1.316 to 0.543	−0.544 to 0.632	−0.43 to 1.469	−0.437 to 0.649	−0.52 to 0.71

A low number indicates that the item is easier to achieve independence.

Table VI. *Significant misfit of individual items and overall fit following re-scoring of disordered thresholds by country*

Motor items												
	Individual Item Fit (following re-scoring)											
	Belgium		France		Israel		Italy		Sweden		UK	
Item	Residual	Chi Prob.	Residual	Chi Prob.	Residual	Chi Prob.	Residual	Chi Prob.	Residual	Chi Prob.	Residual	Chi Prob.
Eating	**7.17**	**0.0000**	**3.49**	**0.0003**	2.72	**0.0000**	0.10	0.0008	1.25	**0.0000**	−0.22	0.7781
Grooming	0.94	0.8223	−0.46	0.3544	0.71	0.5050	**4.54**	**0.0000**	0.27	0.3986	0.902	0.8443
Toileting	−1.51	0.1141	−2.76	0.0175	−1.42	0.0176	−**4.34**	**0.0000**	−1.86	0.1424	−1.231	0.1677
Bowel management	−0.95	0.1199	1.77	0.4078	0.97	0.3050	2.91	0.3561	**6.26**	**0.0000**	1.567	0.1816
	Overall Fit											
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item fit	−0.289	2.738	−0.476	1.454	−0.452	2.328	−0.648	2.535	−0.844	3.567	−0.186	1.554
Person fit	−0.518	0.987	−0.306	0.825	−0.328	1.049	−0.257	0.608	−0.384	1.076	−0.338	1.107
	Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.	
Item-trait interaction	**0.0000**		0.0198		**0.0003**		**0.0000**		**0.0000**		0.0016	
Social-cognitive items												
	Overall Fit											
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Item fit	0.003	0.696	0.306	0.944	−2.390	1.861	0.369	2.439	0.103	2.073	0.325	1.168
Person fit	−0.288	0.862	−0.274	0.933	−0.724	1.003	−0.0319	1.049	−0.266	0.934	−0.329	1.095
	Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.		Chi Sq. Prob.	
Item-trait interaction	0.5729		0.8754		0.0039		0.0121		0.1000		0.8566	

Misfit is in bold.

cognitive items ("expression") showed disordered thresholds (categories). Thus the proportion of ordered response categories was still rather low (0.38) for FIM motor items, but higher (0.80) for social-cognitive items.

The range of measurements for FIM motor items at 0.08 indicated only minor floor or ceiling effects, but was somewhat higher (0.24) for the social-cognitive items (Table VIII). Person separation reliability was high 1.0, indicating that patients were well spread along both measurement constructs. After re-scoring, 4 of the motor items still showed misfit to the Rasch model, giving a proportion for unidimensionality of 0.77.

Table VII. *Location, individual item fit and number of used categories across all countries (pooled data, after re-scoring)*

	Location	SE	Residual	Chi Sq.	Probability	No. of categories
Motor items						
Eating	−4.59	0.19	0.03	16.32	0.0604	2
Grooming	−0.15	0.08	1.19	8.99	0.4384	3
Bathing	0.88	0.04	0.44	4.86	0.8465	7
Dressing upper body	0.41	0.04	3.01	31.81	**0.0002**	7
Dressing lower body	1.07	0.04	−2.74	11.26	0.2585	7
Toileting	−0.97	0.10	−2.77	56.84	**0.0000**	2
Bladder management	−0.69	0.07	**3.23**	24.62	0.0034	3
Bowel management	−2.81	0.12	−0.63	3.46	0.9432	2
Transfer bed	0.14	0.04	−2.39	20.41	0.0156	7
Transfer toilet	0.31	0.04	−**5.65**	26.42	0.0017	7
Transfer tub	2.08	0.06	0.32	5.99	0.7410	4
Walk/wheelchair	1.33	0.06	0.29	5.32	0.8053	4
Stairs	2.99	0.07	−0.11	12.99	0.1632	4
Social-cognitive items						
Comprehension	−0.52	0.04	−0.23	12.42	0.1906	7
Expression	−0.11	0.04	1.60	6.78	0.6604	5
Social interaction	−0.17	0.03	0.63	6.61	0.6779	7
Problem solving	0.71	0.03	−2.46	9.09	0.4291	7
Memory	0.10	0.03	1.07	8.52	0.4823	7

Misfit is in bold.

Table VIII. *Summary table of the psychometric qualities of the Functional Independence Measure (FIM) items across all countries (pooled data)*

Measure	FIM Motor	FIM Soc.-Cogn.
Number of patients	895	895
Order in response categories	0.38	0.80
Unidimensionality	0.77	1.00
Range of measurement	0.08	0.24
Person Separation Reliability	1.00	1.00
Invariance of the Scale		
DIF by gender	1.00	1.00
DIF by age	1.00	1.00
DIF by country	0.20	0.38

DIF = Differential Item Functioning.

Seven of the motor items showed DIF by country. Five items (“grooming”, “bathing”, “transfer bed”, “transfer toilet” and “transfer tub”) displayed both Uniform and Non-Uniform DIF by country, 2 items (“toileting” and “bowel management”) displayed Uniform DIF by country and 1 item (“dressing – lower body”) displayed Non-Uniform DIF by country. Consequently the scale was adjusted for DIF by creating country specific items. This produced 53 motor FIM items – 8 items split across 6 countries and 5 original items (“eating”, “dressing upper body”, “bladder management”, “walk/wheelchair” and “stairs”). Following re-scoring of items with disordered thresholds and the removal of 3 country-specific items which continued to misfit the model (“grooming” in Italy, “transfer toilet” in Israel and “transfer toilet” in Sweden), a 50-item scale could be resolved.[1] The PCA of residuals showed no discernable pattern amongst the data, with a first factor accounting for only 19% of the total variation.

The social-cognitive item “expression” required re-scoring after which all items showed no deviations from the model expectations.[1] However, all but 1 of the items (“memory”) displayed uniform DIF by country. Therefore the scale was adjusted for DIF, resulting in a 25 items scale[1] (4 items split across 6 countries and 1 original item – “memory”). The social-cognitive items showed (following adjustments for DIF) no deviations from model expectations. PCA of the residuals showed the first principal component to account for 35% of the variation, which is deemed to be of no importance (25).

In summary, for the pooled data and after re-scoring and adjusting for DIF, the FIM motor scale after deleting 3 country-specific items met model expectations. The social-cognitive scale also met model expectations without removal of items. Both scales had good person separation reliability and there was no significant DIF by age or gender.

DISCUSSION

This paper evaluates the internal construct validity and cross-cultural validity of the FIM from the perspective of the Rasch measurement model. It has demonstrated that the number of categories presently used in FIM with data from routine clinical settings in the participating countries in Europe is not sustainable. Also, after adjusting for such problems, items have different levels of difficulty across countries. In practical terms this means that: (i) fewer categories might be appropriate; and that (ii) comparison of raw score data between countries is limited. However, utilizing the power of the Rasch model, for example by allowing some items to be unique for each country, limitations can be overcome to a great extent.

Of considerable concern is that there are many disordered thresholds in some items in the FIM motor scale. One explanation for the “bowel” and “bladder management” items may be that in the FIM manual there are 2 ways of assessing the need of assistance and frequency of failure respectively. These may give rise to psychometrically inconsistent values. Also, the higher levels of independence are seen rather infrequently. The “walk/wheelchair” and “stairs” items also display 3 or 4 disordered thresholds. In this case, the scoring can be based both on the walk or wheelchair items, and this may also lead to inconsistency in the scores. Overall, the results indicate that for the FIM motor items, the 7 category scoring does not work as intended and that 3–5 categories would be more appropriate to produce ordered thresholds. The optimum number of categories probably varies between items and also between countries. Using a slightly different approach, where thresholds were constrained to be equal across items, (rating scale model) Grimby et al., in patients with cerebral palsy and spina bifida (26) and in stroke (27), suggested a 5-step scale as giving the best person separation, and no disordered steps as in the 7-step scale. It has also recently been proposed to use 4 steps in multidisciplinary FIM ratings (28). In analysing the ordered category scale by a rank-invariant statistical method, it was suggested that a 5-category (29) or a 4-category scale version of the FIM (30) would be give optimal discrimination. A 7-category scale motivated by being sensitive for clinical purposes could be considered as a clinical working tool, which can be condensed in the data analyses. It is not the aim of the present study to give detailed guidelines how to reduce the number of categories for scoring of FIM items but to indicate the inherent problems and the potential for further analyses to identify the optimal number of categories for each item.

One surprising finding in the current study was that “bladder” and “bowel management” showed adequate fit to the model, which is usually not the case (9, 27). This is also true for the item “stairs” (9). The absence of misfit for these items in the current study may be due to the importance given to the identification of disordered thresholds prior to testing to the model. As disordered thresholds compromise the necessary probabilistic relationship between items, previous misfit may have been due, in part, to this factor. Thus, the lack of discrimination across the thresholds of adjacent categories may have been the problem rather than lack of unidimensionality.

Despite this study being based on admission data the items of the motor scale were well distributed. The location of the

[1] This can be obtained from the first author.

items showed some variation between the countries with the "eating" item easiest in all countries except France, where "bowel management" was easiest. This is consistent with findings from previous studies of patients with stroke (9, 27, 31, 32). "Stairs" was the most difficult item except in Belgium and France, where "transfer to tub/shower" was the most difficult. This latter item showed variation in difficulty between countries, and may depend on the preferential use of bathtub or shower, being tasks with different components (33). A similar mixture of 2 activities in 1 item is seen in "walk/wheelchair", which may explain some cross-country differences, depending on the preference for mobility in the different case-mix. For the social-cognitive items there were a number of differences in difficulty of the item, which may be difficult to explain. Unfortunately, we have no information from the countries on the location of brain lesion or the occurrence of various impairments which may influence items as "expression" (aphasia and anartria) (27, 31, 32) or "problem solving". Nevertheless, "problem solving", as in several earlier studies (9, 32), was the most difficult item in all countries. Some differences in item difficulty have previous been demonstrated between USA and Japan (34) and Italy and USA (35), and cultural explanations for those differences were suggested. In general, it may be assumed that environmental factors, such as eating situation and type of food, dressing habits, physical arrangements in the ward, showers or tubs, cultural aspects in social interaction could vary between different countries as also case-mix and explain some of the differences in item difficulty.

Other factors may also contribute to cross-cultural variability. There may be some difference in the accuracy of the raters in using FIM^{TM} and in the translation of the FIM^{TM} manual. Different degree and form of training in different settings may have had an influence on the psychometric quality of the instrument. It has been shown that raters with formal FIM^{TM} training give more reliable ratings (36). Two of the participating countries (Italy and Sweden) have territorial licences with the central database in Buffalo, but the level of training in the other countries was unknown. Furthermore, despite training, differences in rater leniency will always exist and one solution may be the use of multi-faceted analysis that controls for rater leniency (37, 38). The influence of all the factors giving lack of invariance and the extent of their interaction are unknown and cannot be analysed in the present study.

The DIF approach was used in the present study to further objectively describe differences in item difficulty between countries. As many as 8 items in the motor scale displayed differential item functioning, which led to the conclusion that those ought to be split across countries when data should be pooled together. However, after removal of 3 country-specific misfitting items it was possible to achieve adequate fit to the Rasch model. A similar solution, without the deletion of items, was achieved for the social-cognitive scale. It is important to point out that the FIM scales may work well in clinical settings as a unidimensional scale within a single country, but that the scale works in a slightly different way across each or most countries, such that DIF compromises unidimensionality.

In such a large-scale study as the present, there are several limitations, as already mentioned. After having identified problems and established a potential solution to measurement in admission data, the next step would be to verify stability across time, a requirement for outcome measures. We have not been able, due to the format of data collection, to go into detailed studies of the impact of various stroke subgroups. In the current study, only comparison has been made between countries, but differences may also exist between centres within the same country. Differences in training across and within countries may contribute to variability, as may case mix. Consequently it is likely that the solution we have obtained, in order to facilitate the pooling of data from these countries, will be unique to the combination of centres involved. It is important to understand that a *requirement* of measurement is invariance across groups. At the present time, most of this variability can be accommodated within the framework of the Rasch measurement model, but this is a complex task. Resolving one or more of the issues highlighted in the present paper will make group comparisons more transparent and more widely available.

In conclusion, pooling of FIM data from stroke patients in Europe can be achieved, conditional upon careful and sophisticated adjustment of disordered thresholds and DIF such that adequate fit to the Rasch model is achieved, despite the potential variation in case mix, and rater training which may exist.

ACKNOWLEDGEMENTS

This study was funded by the European Commission within its BIOMED 2 programme under contract BMH4-CT98-3642. It has also been supported by grants from Swedish Research Council (VR K2002-27-VX-14318-01A) and The Swedish Society of Medicine. We would like to express our thanks to all our participating partners, who have supplied FIM^{TM} data; a list of the partners can be obtained from the first author.

REFERENCES

1. ICF International Classification of Functioning, Disability and Health. Geneva: World Health Organisation; 2001.
2. Keith RA, Granger CV, Hamilton BB, Sherwin FS. The functional independence measure: a new tool for rehabilitation. In: Eisenberg MG, Grzesiak RC, eds. Adv Clin Rehabil. New York: Springer; 1987; Vol 1, p. 6–18.
3. Cohen ME, Marino RJ. The tools of disability outcomes research functional status measures. Arch Phys Med Rehabil 2000; 81 (Suppl 2): S21–S29.
4. Rasch G. Probabilistic Models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1980.
5. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003; 35: 105–115.
6. Wright BD. Masters GN. Rating Scale Analysis. Chicago: MESA Press; 1982.
7. Andrich D. Rasch models for measurement. Newbury Park: Sage Publications; 1998.
8. Haigh R, Tennant A, Biering-Sørensen F, Grimby G, Marincek C, Phillips S, et al. The use of outcome measures in physical medicine and rehabilitation within Europe. J Rehabil Med 2001; 33: 273–278.
9. Linacre JM, Heinemann AW, Wright BD, Granger CV, Hamilton

BB. The structure and stability of the Functional Independence Measure. Arch Phys Med Rehabil 1994; 75: 127–132.

10. Tennant A, Penta M, Tesio L, Grimby G, Thonnard J-L, Slade A, et al. Assessing and adjusting for cross cultural validity of impairment and activity limitation scales through Differential Item Functioning within the framework of the Rasch model: the Pro-ESOR project. Med Care 2004; 42(Suppl): 137–148.
11. Master GN. A Rasch model for partial credit scoring. Psychometrika 1982; 47: 149–174.
12. Smith RM. Fit analysis in latent trait measurement models. J Applied Measurement 2000; 2: 199–218.
13. Silverstein B, Kilore KM, Fisher WP, Harley JP, Harvey RF. Applying psychometric criteria to functional assessment in medical rehabilitation: I. Exploring unidimensionality. Arch Phys Med Rehab 1991; 72: 631–637.
14. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheum 1996; 35: 574–578.
15. Prieto L, Alonso J, Lamarca R, Wright BD. Rasch measurement for reducing the Items of the Nottingham Health Profile. J Outcome Meas 1998; 2: 285–301.
16. Tesio L, Valsecchi MR, Sala M, Guzzon P, Battaglia MA. Level of activity in profound/severe mental retardation (LAPMER): a Rasch-derived scale of disability. J Appl Measurement 2002; 3: 50–84.
17. Smith RM. A comparison of methods for determining dimensionality in Rasch measurement. Structural Equation Modelling 1996; 3: 25–40.
18. Angoff WH. Perspectives on Differential Item Functioning Methodology. In: Holland PW, Wainer H. Differential Item Functioning. Hillsdale, New Jersey: Lawrence Erlbaum; 1993.
19. Fisher WP. Reliability statistics. Rasch Measurement Transactions 1992; 6: 238.
20. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995; 310: 170.
21. Dorans NJ, Holland PW. DIF detection and description: Mantel-Haenszel and standardisation. In: Holland PW, Wainer H, eds. Differential Item Functioning. Hilldale, NJ: Lawrence Erlbaum Associates; 1993, p. 36–66.
22. Lange R, Irwin HJ, Houran J. Top-down purification of Tobacyk's revised paranormal beliefe scale. Personality and Individual Differences 2000; 29: 131–156.
23. Lange R, Thalbourne MA, Houran J, Lester D. Depressive response sets to gender and culture-based Differential Item Functioning. Personality and Individual Differences 2002; 33: 937–954.
24. Andrich D, Lyne A, Sheridon B, Luo G. RUMM 2020. Perth: RUMM Laboratory; 2003.
25. Hair JF, Andersen RE, Tatham RL, Black WC. Multivariate data analysis with readings. New Jersey: Prentice Hall; 1995.
26. Grimby G, Andren E, Holmgren E, Wright B, Linacre JM, Sundh V. Structure of a combination of Functional Independence Measure and Instrumental Activity Measure items in community-living persons: a study of individuals with cerebral palsy and spina bifida. Arch Phys Med Rehabil 1996; 77: 1109–1114.
27. Grimby G, Andren E, Daving Y, Wright B. Dependence and perceived difficulty in daily activities in community-living stroke survivors 2 years after stroke: a study of instrumental structures. Stroke 1998; 29: 1843–1849.
28. Heinemann AW, Semik P, Bode R. Reducing step disorder in multidisciplinary FIM ratings. Proceedings of the 1st World Congress of the International Society of Physical and Rehabilitation Medicine; 2001; July 7–13; Amsterdam, The Netherlands: Bolonga: Monduzzi Editore; 2001.
29. Claesson L, Svensson E. Measures of order consistency between paired ordinal data: application to the Functional Independence Measure and Sunnaas index of ADL. J Rehabil Med 2001; 33: 137–144.
30. Gosman-Hedström G, Svensson E. Parallel reliability of the Functional Independence Measure and the Barthel ADL index. Disabil Rehabil 2000; 22: 702–715.
31. Grimby G, Gudjonsson G, Rodhe M, Sunnerhagen KS, Sundh V, Ostensson ML. The functional independence measure in Sweden: experience for outcome measurement in rehabilitation medicine. Scand J Rehabil Med 1996; 28: 51–62.
32. Heinemann AW, Linacre JM, Wright BD, Hamilton BB, Granger C. Relationships between impairment and physical disability as measured by the Functional Independence Measure. Arch Phys Med Rehabil 1993; 74: 566–573.
33. Küçükdeveci AA, Günes Y, Tennant A, Süldür N, Sonel B, Arasil T. Adaptation of the modified Barthel index for use in physical medicine and rehabilitation in Turkey. Scand J Rehab Med 2000; 32: 87–92.
34. Tsuji T, Sonoda S, Domen K, Saitoh E, Liu M, Chino N. ADL structure for stroke patients in Japan based on the Functional Independence Measure. Am J Phys Med Rehabil 1995; 74: 432–438.
35. Tesio L, Granger CV, Perucca L, Franchignoni FB, Battaglia A, Deutsch A, Russell C. The FIM-Functional Independence Measure from USA to Italy: a comparison study. Am J Phys Med Rehabil 2002; 81: 168–176.
36. Fricke J, Unsworth C, Worell D. Reliability of the Functional Independence Measure with occupational therapist. Austral Occup Ther J 1993; 40: 2–15.
37. Fisher AG. Development of functional assessment that adjust for task simplicity and rater leniency. Wilson M, ed. Objective measurement: theory into practice. Norwood, NJ: Ablex; 1994; 2: p. 145–75.
38. Bernspång B. Rater calibration stability for the Assessment of Motor and Process Skills. Scand J Occup Ther 1999; 6: 101–109.