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

Alan Tennant, Massimo Penta,† Luigi Tesio,‡ Gunnar Grimby,§ Jean-Louis Thonnard,† Anita Slade,* Gemma Lawton,* Anna Simone,‡ Jane Carter,|| Åsa Lundgren-Nilsson,§ Maria Tripolski,¶ Haim Ring,¶ Fin Biering-Sørensen,** Črt Marincek,†† Helena Burger,†† and Suzanne Phillips||*

Introduction: In Europe it is common for outcome measures to be translated for use in other languages. This adaptation may be complicated by culturally specific approaches to certain tasks; for example, bathing. In this context the issue of cross-cultural validity becomes paramount.

Objective: To facilitate the pooling of data in international studies, a project set out to evaluate the cross-cultural validity of impairment and activity limitation measures used in rehabilitation from the perspective of the Rasch measurement model.

Methods: Cross-cultural validity is assessed through an analysis of Differential Item Functioning (DIF) within the context of additive conjoint measurement expressed through the Rasch model. Data from patients undergoing rehabilitation for stroke was provided from 62 centers across Europe. Two commonly used outcome measures, the Mini-Mental State Examination (MMSE) and the Functional Independence Measure (FIM) motor scale are used to illustrate the approach.

Results: Pooled data from 3 countries for the MMSE were shown to fit the Rasch model with only 1 item displaying DIF by country. In contrast, many items from the FIM expressed DIF and misfit to the model. Consequently they were allowed to be unique across countries, so resolving the lack of fit to the model.

Conclusions: Where data are to be pooled for international studies, analysis of DIF by culture is essential. Where DIF is observed, adjustments can be made to allow for cultural differences in outcome measurement.

Key Words: Rasch, DIF, outcomes, cross-cultural, rehabilitation

(*Med Care* 2004;42: I-37–I-48)

Throughout Europe, a wide range of therapeutic interventions is used within rehabilitation, often country-specific and reflecting the different traditions of health care. To match this diversity of treatment modalities, there is also a broad range of outcome measures used to evaluate the efficacy of treatment.[1] These measures focus mostly on aspects of impairment and activity limitation (disability) as defined by the International Classification of Functioning, Disability and Health (ICF).[2] Their use, which attempts to quantify the results of rehabilitation, is seen as an increasingly important part of good clinical practice. However, if we wish to identify the most effective and efficient treatment modalities at the European level, it follows that we require outcome measures that serve this purpose.

Some attempt at standardization of outcome measurement in rehabilitation has been made in North America. The Functional Independence Measure (FIM),[3] a measure of disability, is used across a wide range of conditions and in a wide range of situations in rehabilitation. This scale has also

From the *Academic Unit of Musculoskeletal and Rehabilitation Medicine, University of Leeds, 36 Clarendon Road, Leeds, LS2 9NZ, United Kingdom; †Unité de Réadaption et de Médecine Physique, Université Catholique de Louvain, B-1200 Bruxelles, Belgium; ‡Divisione di Recupero e Rieducazione Funzionale, Salvatore Maugeri Foundation, IRCSS, Pavia, Italy; §Department of Rehabilitation Medicine, Sahlgrenska Academy at Göteborg University, Sweden; ||The Bath Head Injury/Neuro-Rehabilitation Unit, Royal National Hospital for Rheumatic Diseases, National Health Services Trust, Bath, United Kingdom; and ¶Loewenstein Hospital, Rehabilitation Centre, Raanana, Tel Aviv University, School of Medicine, Israel.

Reprints: Alan Tennant BA, PhD, Professor of Rehabilitation Studies, Academic Unit of Musculoskeletal & Rehabilitation Medicine, The University of Leeds, 36 Clarendon Road, Leeds LS2 9NZ, UK. E-mail: a.tennant@leeds.ac.uk.

ISSN: 0025-7079/04/4200-0037

DOI: 10.1097/01.mlr.0000103529.63132.77

been widely adopted throughout Europe.[4–6] Despite this, across Europe, each diagnostic group tends to have its own set of commonly used outcome measures.[1] For example, for the diagnosis of stroke, as well as the FIM, measures such as the Barthel Index[7] and the Mini-Mental State Examination (MMSE)[8] are also widely used.

Once outcome measures are introduced, it is common for them to be translated for use in other languages.[9–12] In this context, the issue of cross-cultural validity becomes paramount. Although guidelines, both old and new, exist for adapting instruments into different languages,[13–14] these tend to focus on self-complete instruments, whereas those measures in use in routine clinical practice tend to involve assessment of impairment and disability by professionals. Although in principle the adaptation procedures should be similar, it is often unclear as to exactly how (and by whom) these clinical measures were first translated.[5,15–18] For purposes of comparing the rehabilitation services across Europe, it thus becomes essential to determine whether or not different language versions of such measures do work in the same way.

The objective of this paper is to demonstrate the use of Differential Item Functioning (DIF) as a mechanism to evaluate the cross-cultural validity of outcome measures used in rehabilitation and so facilitate the pooling of data in international studies. The approach is set within the framework of the Rasch unidimensional measurement model.[19,20] Within Item Response Theory [IRT] this is known as the 1-parameter model. IRT is a general statistical theory about item (question or task) and scale performance and how that performance relates to the abilities measured by the items in the scale.[21] Although early work in IRT was contiguous with that of Rasch (and it was known as Latent Trait Theory in the 1950s) it was Lord and Novick's 1968 publication[22] of their seminal work on Statistical Theories of Mental Test Scores, and particularly the contribution of Birnbaum,[23] that formally established IRT. Today, IRT encompasses many different types of parametric and nonparametric models, each with various extensions of the original dichotomous model.[24] However, the Rasch model has unique properties which are crucial to attaining additive conjoint measurement,[25] a prerequisite for the calculation of change scores, effect sizes and drawing correct inferences from various statistical tests which are common in health outcome studies.[26–30]

The Rasch model is a unidimensional model which asserts 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. The dichotomous Rasch model assumes that the probability of a given respondent to give a "correct" answer to a particular item is a logistic function of the relative distance between the item location and the respondent location on a linear scale. In other words the probability that a person will affirm an item is a logistic function of the difference between the person's ability [θ] and the difficulty of the item [b] (ie, the ability required to affirm item i), and only a function of that difference.

$$P_{ni} = \frac{e^{(\theta n - b_i)}}{1 + e^{(\theta n - b_i)}} \quad (1)$$

where p_{ni} is the probability that person n will answer item i correctly (or be able to do the task specified by that item), θ is person ability, and b is the item difficulty parameter.

From this, the expected pattern of responses to an item set is determined given the estimated θ and b. When the observed response pattern coincides with or does not deviate too much from the expected response pattern, then the items constitute a true Rasch scale.[31] Taken with confirmation of local independence of items, that is, no residual associations in the data after the Rasch trait has been removed, this supports unidimensionality.[32,33]

The model can be extended to cope with items with more than 2 categories,[34] and this involves an explicit "threshold" parameter (τ), where the threshold represents the equal probability point between any 2 adjacent categories within an item. In the logit form this is:

$$\ln\left(\frac{P_{nik}}{1-P_{nik}}\right) = \theta_n - b_i - \tau_k \quad (2)$$

where P is the probability of person n affirming category k in item i; θ is person ability, b is the item difficulty parameter, and τ_k is the difficulty of the k threshold. The model used in the present analysis is a further derivation, the Partial Credit Model[35]

$$\ln\left(\frac{P_{nik}}{1-P_{nik}}\right) = \theta_n - b_{ik} \quad (3)$$

where no assumptions are made about the equality of threshold locations relative to each item.

It is easy to see how the approach was readily adopted in rehabilitation in the late 1980s.[36–38] Patients undergoing rehabilitation have a given ability level, and they are presented with a range of tasks with differing degrees of difficulty. The language of ability and difficulty easily transferred from education to rehabilitation. From this early work the approach quickly moved into the mainstream of health status measurement.[39–40] The early published work on Rasch analysis in rehabilitation explored issues of unidimensionality and scaling properties[39] and this has remained a central theme to date.[41–43]

However, the approach can also contribute to the issue of cross-cultural validity. Essentially the scale should work in the same way, irrespective of which group is assessed. Thus, the location of items along the measurement construct should

remain the same across cultures. Originally, tests for such differences came under the rubric "item bias." Here, individuals with the same score on a test are expected to have the same probability of affirming an item, irrespective of any group membership.[44] This type of analysis has latterly been given the name DIF.[45] The basis of the DIF approach lies in the item response logistic function, the proportion of individuals *at the same ability level* who 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 proportion is found irrespective of the nature of the group for whom a function is plotted.[46] Items that do not yield the same item response function for 2 or more groups display DIF and are violating the requirement of unidimensionality.[47] 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. It was thus the potential of the IRT approach, and specifically the application of the Rasch model with its attendant measurement properties, which led to the project to examine cross-cultural validity, funded under the European Commission's BIOMED2 program. This project was the European Standardization of Outcome Measurement in Physical Medicine and Rehabilitation which, as with all European funded projects, is given a mnemonic, in this case PROESOR.

MATERIALS AND METHODS

Rasch Analysis

Thus the project sought to explore the potential of the IRT Rasch model for evaluating the cross-cultural validity of outcome scales used in Rehabilitation in Europe. A survey of the use of outcome measures across Europe (1) identified a set of measures for evaluation. Two scales identified in this survey have been chosen to illustrate the approach, both commonly used with patients who have had a stroke. These are the MMSE[8] and the FIM motor subscale.[3] The former is a measure of cognitive impairment and the latter, a measure of activity limitation, looking at the independence in activities of daily living.

The approach, utilizing the Rasch model described above, follows a sequence of analyses. All analysis is undertaken on pooled data with age, gender, and country entered as "person factors" for subsequent DIF analysis. In the first instance, where data are polytomous, an analysis is undertaken of the ordering of each category. An "ordered category" implies that the response categories (eg, 1, 2, 3...) reflect an increasing amount of the latent variable under investigation. The boundaries between categories, as we have seen above, are called "thresholds" and it is possible to observe "disordered thresholds" in the data, where, for example, with an increasing category number representing higher disability, the threshold between categories 2 and 3 is found to represent a lower level of disability than the threshold between categories 1 and 2.[20] Where this occurs, it will be necessary to collapse adjacent categories which can be undertaken as part of the ongoing Rasch analysis.

Following this, the data are then (re)fitted to the Rasch model to determine overall fit, and how well each item fits the model. These statistics indicate how far the observed data match that expected by the model. Three overall fit statistics are considered. Two are item-person interaction statistics distributed as a Z statistic with mean of zero and SD of 1 (which indicates perfect fit to the model). A third is an item-trait interaction statistic reported as a χ^2, reflecting the property of invariance across the trait. This means that the hierarchical ordering of the items remains the same at different levels of the underlying trait, indicated by a nonsignificant χ^2. In addition, individual item-fit statistics are presented, both as residuals (a summation of individual person and item deviations—usually acceptable within the range $\pm$ 2) and as a χ^2 statistic (deviation from the model by groups of people defined by their ability level—requiring a nonsignificant χ^2; ie, a P value of 0.05 and above, with appropriate adjustment for repeated tests). Misfit of items indicates a lack of the expected probabilistic relationship between the item and other items in the scale. This may derive from DIF (see below) or may indicate that the item does not contribute to the trait under consideration.

Where data fit the Rasch model, an estimate of both person ability and item difficulty is provided on the same linear scale. The logit scale is an interval scale, where each unit represents an increase in the odds of a person succeeding on an activity by 2.716 times. It is customary to set the mean difficulty of items as 0.0 logits to solve the indeterminancy of the scale. The items are situated along the interval scale according to their difficulty. The average person ability and spread (ie, the SD) will indicate how well the scale is targeted at the sample, with less well-targeted scales having an average person ability well away from the central zero logit. A person separation index can be used to compute the number of discernible strata in the distribution of patient measures.[48] The index gives a practical indication of how precisely patients have been spread out along the measurement construct defined by the items.

As we have seen, within the framework of Rasch measurement, the scale should work in the same way, irrespective of which group is being assessed. Thus, in the case of disability, the probability of a person affirming an item (or category) at a given level of disability should be the same for younger or older people, men and women, and so on. Items that do not yield the same item response function for 2 or more groups display DIF and are violating the requirement of unidimensionality.[47] Consequently, every item is checked for

DIF by age and gender and, for the current analysis, by country.

Briefly, for the response of each person to each item, the standardized residual Z_{ni} of the observed score X_{ni} from that predicted by the model, $E[X_{ni}]$ is calculated according to

$$Z_{ni} = \frac{X_{ni} - E[X_{ni}]}{\sqrt{V[X_{ni}]}} \quad (4)$$

Then each person is classified according to one of G class intervals and for example, in the case of gender, giving a set of residuals suitable for a 2-way analysis of variance (ANOVA) design.[49] Thus, the statistical test used for detecting DIF is an ANOVA of the person-item deviation residuals with person factors (eg, gender) and class intervals (eg, Group along the trait) as factors. Two types of DIF can be identified: uniform and nonuniform DIF. With the former, there is a constant difference between groups in the probability of affirming an item (or category) across the trait (ANOVA main effect), and with the latter the difference varies across the trait (ANOVA interaction effect).

Where some but not all items display DIF, adjustments can be made to allow items with DIF to vary by group. The approach adopted has been described as an iterative "top-down purification" approach in that a requirement for identifying DIF is a baseline set of "pure" items.[50] 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.[51] The rejected items are then reintroduced to test the results. When an item displays DIF, it is rendered unique to the group(s) which display DIF. For example, suppose that Italy displayed DIF for a bathing item, compared with England and Sweden (post hoc tests for the ANOVA identify where the difference(s) lie). In this case, the item would be split into 2, with 1 item for Italy (and the responses for patients from England and Sweden entered as structural missing values) and 1 item for England and Sweden together (with the responses for Italian patients entered as structural missing values). Consequently, there would be now 2 bathing items, an Italian item and an Anglo-Swedish item. The unsplit (pure) items act as links in the calibration. In this way, item difficulty for some items is allowed to vary across countries. Fit is again reassessed, and if items still display misfit to the model, then they are removed from the scale and do not contribute to the person estimate.

Finally, 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 and indicate that the attributes of conjoint additive measurement are retained in the data.

Sample Size and Statistical Software

Sixty-two rehabilitation facilities contributed anonymous data from consecutively admitted patients. It was known that a sample size of 150 patients would estimate the item difficulty, with α of 0.01, to within $\pm$ 0.5 logits.[52] This sample size is also sufficient to test for DIF where (a) at α of 0.01 a difference of 0.5 SD within the residuals can be detected for any 2 groups with β of 0.20; or (b) with identical α and β, a difference in SDs within the residuals can be detected where this difference varies by 0.1 logits across each of 6 groups.[53] Bonferroni corrections are applied to both fit and DIF statistics due to the number of tests undertaken for any given scale.[54]

In any event, 3 countries, Italy (n = 108), Slovenia (n = 78) and Sweden (n = 149) contributed 335 cases for the analysis of the MMSE, and data were received from 6 countries (Belgium 143, France 157, Italy 1046, Israel 319, Sweden 642, and UK 239) for the FIM motor scale. Due to the unequal sample size for the latter scale, a random sample of cases was taken from Italy, Sweden, and Israel to provide approximately 150 cases from each country for the analysis. This was necessary to conform to equal group size requirements for the ANOVA of residuals.[55]

Rasch analysis was undertaken using the RUMM2010 package.[56]

The Outcome Measures

Briefly, the FIM motor scale comprises 13 items with a 1–7 category response. Items are tasks, ranging from bladder management, to climbing stairs. A category response of 1 reflects total dependence, and 7, total independence. Thus a total score can range from 13 to 91. The MMSE covers 6 domains, including orientation and recall. The domains have a variable number of items, and the items have a different number of categories (ranging from dichotomous to a 6-category response). In practice, 11 items are summated with a maximum score of 30. A low score reflects cognitive dysfunction, with a score of 23 considered nonnormal; 24 and above, normal.

RESULTS

The MMSE

Overall, items were found to fit the model (Table 1; Item Fit Mean −0.038, SD 1.300) and the Person Separation Index was satisfactory (0.899). However, Item-Trait Interaction showed marginal significance, indicating a variable hierarchical ordering of items across the trait (χ^2 (df =33) 58.203, P = 0.004357). Also, 5 items displayed disordered thresholds (Fig. 1). For example, while the item "disorientation in time" displayed ordered thresholds, in that each category demonstrates an increasing level of the trait of

TABLE 1. Fit of MMSE to Rasch model

Item	Location	SE	Residual	χ^2	Probability
Orientation time	0.61	0.07	0.28	0.24	0.6221
Orientation place	−0.09	0.08	−1.76	1.15	0.2843
Registration	−1.39	0.18	−0.40	0.81	0.3670
Attention and calculation	1.14	0.11	−0.41	0.81	0.3686
Recall	0.44	0.16	0.99	0.33	0.5669
Language naming	−3.09	0.44	−1.46	1.22	0.2690
Language repeating	0.40	0.16	1.08	1.29	0.2567
Language command—verbal	−0.48	0.14	0.56	0.14	0.7042
Language command—written	−0.24	0.18	−1.27	3.04	0.0811
Language write	0.58	0.16	0.41	0.04	0.8406
Copying	2.13	0.14	1.61	2.72	0.0988

mental state, the item "registration" failed to display an ordering over the trait and could thus not be presented in a graphical fashion as the categories are out of order.

These items were rescored to see if fit to the model was improved. During rescoring, 2 items had to be reduced to dichotomous responses. These were "Language—naming" and "Recall". This produced an improvement in overall fit to model (Item Fit Mean −0.034, SD 1.113; Person Fit Mean −0.144, SD 0.464) and all individual item fit statistics were satisfactory, both for residuals and χ^2 (Table 2). The Item-Trait Interaction statistic became nonsignificant (χ^2 (11) 11.799, P = 0.378962), demonstrating invariance across the trait. The Person Separation Index remained satisfactory (0.897), indicating the scale was able to differentiate almost 4 separate groups of patients across the scale.[48]

Although the overall fit of the data to the model was good, 2 items were found to display uniform DIF by age, and 1 item displayed nonuniform DIF by country. The "language—naming" and "language—copying" displayed uniform DIF by age. The "language command—written" item displayed DIF by country, and the comparative response functions for this item are shown in Figure 2. A Bonferroni corrected probability of 0.0045 would indicate a significant difference in these circumstances.

In Figure 2, the model expectation curve for the item is given, and also for each country. In practice, for each country,

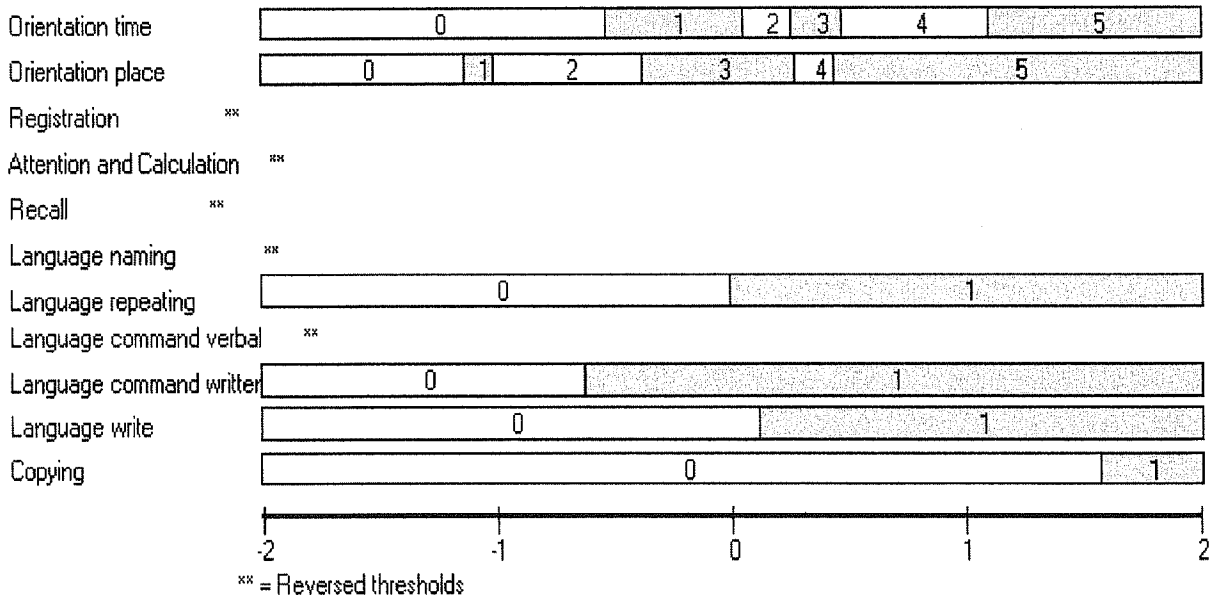

FIGURE 1. Threshold ordering of MMSE.

TABLE 2. Fit of MMSE (adjusted for DIF) to Rasch model

Item	Location	SE	Residual	χ^2	Probability
Orientation time	0.58	0.07	0.31	3.49	0.4793
Orientation place	−0.13	0.08	−1.76	4.60	0.3313
Registration	−1.42	0.18	−0.41	2.87	0.5798
Attention and calculation	1.11	0.11	−0.36	3.27	0.5143
Recall	0.42	0.16	0.98	3.50	0.4778
Language naming	−3.12	0.44	−1.44	3.99	0.4069
Language repeating	0.37	0.16	1.09	7.03	0.1344
Language command—verbal	−0.52	0.14	0.56	2.96	0.5640
Language command—written	−0.56	0.22	−1.34	3.83	0.4294
Language command—written, Slovenia	0.63	0.34	−0.86	2.41	0.6610
Language write	0.55	0.16	0.46	1.80	0.7726
Copying	2.10	0.14	1.63	2.21	0.6973

the expected score is given for 2 groups of people at different levels of the trait (the number of groups is determined by sample size). It becomes obvious from Figure 2 (as well as post hoc tests) that the difference in response function can be attributed to the Slovenia data. Consequently, the item was split in 2; 1 for Italy and Sweden together and 1 for Slovenia. Following this, fit of the resulting 12-item scale to the model was found to be good (Item Fit Mean −0.095, SD 1.099; Person Fit Mean −0.146, SD 0.462; Item-Trait Interaction χ^2 (48) 41.957, P = 0.717653; Table 3), and DIF for country was absent. PCA analysis of residuals showed no discernable pattern, with the first factor taking 13% of the variation among the residuals, so supporting unidimensionality.

The FIM Motor Subscale

Overall fit to the model of the FIM motor scale was poor (Item Fit Mean −0.360, SD 4.642). The Item-Trait Interaction statistic was significant (χ^2 (117) 504.309, $P <$ 0.0001) showing a lack of invariance across the trait. Individual item fit is shown in Table 3.

Only 5 of the 13 items were found to have ordered thresholds. Consequently, 8 disordered items had to be rescored, 3 of which had to be dichotomized (Fig. 3).

Following rescoring overall fit to the model remained poor (Item Fit Mean −0.444, SD 2.450) and the Item-Trait Interaction Statistic remained significant (χ^2 (117) 229.288, $P <$ 0.0001) although the Person Separation Index was satisfactory (0.969). DIF was tested with respect to age, gender, and country. No items were found to have DIF by gender or age, but 8 items were found to have DIF by Country (Table 4). Consequently, the items that displayed DIF were split and made unique across Countries. A new analysis was run on the resulting 53 items (5 original "pure" items and 8 split across 6 countries). Items with disordered thresholds were once again

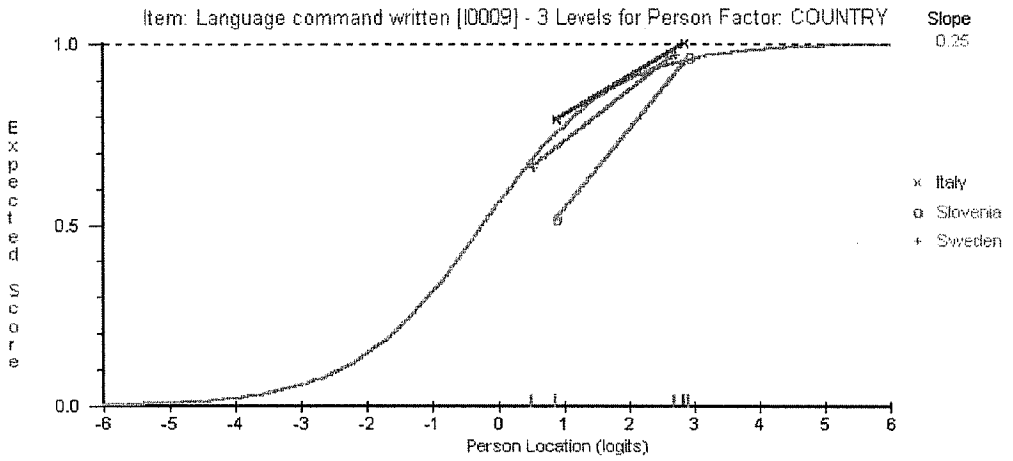

FIGURE 2. Comparative response functions for the item "language command—written" in the MMSE.

TABLE 3. FIM individual item fit*

Item	Location	SE	Residual	χ^2	Probability
Eating	−1.11	0.04	6.39	80.46	0.0000
Grooming	−0.41	0.03	0.34	8.95	0.4423
Bathing	0.39	0.03	−3.37	27.06	0.0014
Dressing—upper body	0.03	0.03	−0.87	23.67	0.0049
Dressing—lower body	0.47	0.03	−6.00	41.53	0.0000
Toileting	0.11	0.03	−4.14	31.97	0.0002
Bladder management	−0.64	0.03	5.74	49.70	0.0000
Bowel management	−0.88	0.03	4.70	42.88	0.0000
Transfer bed	−0.15	0.03	−5.79	39.26	0.0000
Transfer toilet	−0.04	0.03	−7.39	57.19	0.0000
Transfer tub	0.80	0.03	1.88	24.29	0.0039
Walk/wheelchair	0.24	0.03	1.57	14.17	0.1164
Stairs	1.19	0.03	2.27	63.20	0.0000

*Bonferroni corrected significance level $P = 0.0008$.

rescored, and misfitting items were deleted, reducing the scale to 50 items (all 5 original items—Eating, Dressing-Upper Body, Bladder Management, Walk/Wheelchair and Stairs—to serve as "link" items, and 45 split items). The unique "transfer-toilet" items for Sweden and Israel were deleted, along with the "grooming" item for Italy. The final fit of this scale was good at the individual level, and at the Item Fit level (mean −0.462, SD 1.550), although the Item-Trait Interaction statistic remained borderline nonsignificant (χ^2 (50) 89.889, $P = 0.000462$). Final fit of the individual 50 items is shown in Table 5. Once again, PCA of residuals showed no discernable pattern among the data, with a first factor accounting for only 19% of the total variation.

DISCUSSION

A DIF approach within the framework of the Rasch measurement model offers a sophisticated way of identifying and, under certain circumstances, of adjusting for problems associated with cross-cultural differences in outcome scales.

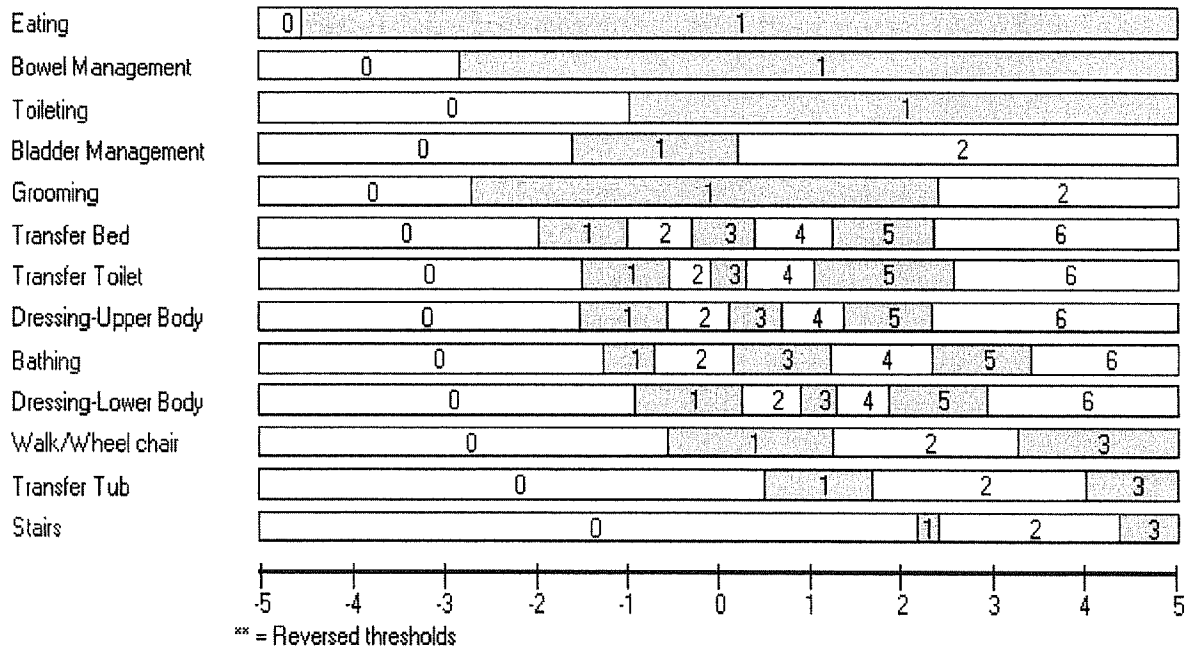

FIGURE 3. Threshold Map of FIM motor scale following rescoring.

TABLE 4. FIM motor items displaying DIF by country*

	Uniform		Nonuniform	
Item	**F**	***P***	**F**	***P***
Grooming	4.759	0.00025		
Bathing	10.340	<0.0001		
Dressing—lower body			1.991	<.0001
Toileting	8.548	<0.0001		
Bladder management	4.735	<0.0001		
Bowel management				
Transfer bed	20.56	<0.0001		
Transfer toilet	5.359	<0.0001	2.086	<0.0001
Transfer tub	39.11	<0.0001	3.232	<0.0001

*Bonferroni corrected significance level $P = 0.0008$.

The 2 scales chosen for illustration in this paper, despite their widespread use as international standards, both demonstrated initial problems with basic measurement requirements. At a given level of the trait, the probability of response to various categories was shown to be different across Countries. Through adjusting for DIF, these problems were resolved for the MMSE, and largely for the FIM. Given this adjustment, unidimensionality was supported, although the solution for the FIM required the omission of 3 county-specific items.

Fit of the data to the Rasch model, including a PCA of the residuals to test the local independence assumption, supports the unidimensionality of the scale. Although factor analysis is a traditional approach to determine unidimensionality, there are considerable problems associated with using ordinal data for this type of analysis and, for example, spurious factors can be caused by DIF.[57]

MMSE scores have previously been found to be influenced by both culture and educational level[58] and Teresi and colleagues[59] have identified DIF for culture and education. The latter has not been considered in the present analysis but, particularly with an expanding Europe, may introduce another important source of DIF which will need to be considered. It has also been suggested that for cognitive screening tests, shorter tests composed of easier items which rely more on memory recall, rather than performance related to language and literacy, would perform better.[60] The interaction between culture and educational level, and its effect upon DIF, remains unknown.

The FIM is widely used in rehabilitation throughout Europe and North America and poses particular challenges for cross-cultural studies. This is explicitly acknowledged by proponents of the FIM who have established formal procedures for translation and training across countries. It became clear during the analysis that the original scoring function of 7 categories is not working well. This is true both across and within countries, such as Italy, where formal training programs help therapists and others assign grades according to strict criteria. The rescoring undertaken in the analysis above suggests that a common scoring function across items may be difficult to achieve,in that some items had to be dichotomized. This indicates that some tasks are perceived by therapists as an "all or nothing" situation for patients, compared with other tasks which can be graded. However, it is possible that such a situation may vary by diagnostic group.

Many items displayed DIF by Country, but through making those items unique to each Country, a solution was found that included most items for most Countries. However, problems remain. The "grooming" item in Italy for example, would not fit the model as a Country-specific item. Analysis of the Italian data only (not shown) showed this item to lack fit to the model. Is this because the adaptation of the scale to the Italian culture brought about a fundamental change in its relationship with other items? The FIM motor scale works in other Countries, for example the UK, which suggests that the adaptation may have changed the item. We must also stress here that it is not the wording of the item that necessarily changes, but rather the way in which the task is evaluated in the clinical setting. Once again, different traditions across Europe may influence this, for example, in the way in which patients are bathed in a hospital setting.[61] This may result in subtle differences in the degree of difficulty of the tasks.

However, other issues may be relevant here; for example, the nature of the underlying lesion causing the stroke (DIF by case mix, based on lesion type). If the case mix is not homogeneous across countries, this may cause the discrepancy and should be investigated before definitive statements are made about the quality of the scale. Even on homogeneous case mix, with an apparent paradox, differential training may be a source of DIF, as long as it introduces systematic differences in interpretation of the items across countries.

There are several methodological issues which are raised by proposing this approach as a standard for assessing and adjusting for the cross-cultural validity of outcome measures across Europe. The detection of DIF by ANOVA techniques is vulnerable to sample size issues. With large samples, DIF can usually be found. Therefore this issue becomes one of substantive difference as opposed to statistical significance, yet there is little literature to guide what might constitute a substantive difference. Sample size also affects interpretation of χ^2-based fit statistics[62] and, again, with large enough sample sizes, misfit to the model can always be found. In the current study, we took samples of patients from some countries which returned large numbers of cases, to satisfy the equality of group requirements in the ANOVA analysis. We know, in the case of the Rasch model, that certain sample sizes can deliver given levels of precision for the parameter esti-

TABLE 5. FIM motor scale adjusted for DIF at the country level*

Item	Location	SE	Residual	χ^2	Probability
Eating	−5.11	0.19	0.09	5.53	0.0187
Grooming, UK	−0.41	0.12	3.28	4.07	0.0437
Grooming, Sweden	−1.16	0.14	−0.14	0.77	0.3810
Grooming, Israel	−0.40	0.11	0.15	2.10	0.1474
Grooming, France	−0.12	0.13	−0.37	0.15	0.7000
Grooming, Belgium	−1.16	0.09	1.66	1.00	0.3183
Bathing, UK	0.43	0.10	1.03	0.33	0.5664
Bathing, Sweden	0.80	0.11	0.57	0.20	0.6576
Bathing, Italy	0.39	0.10	−1.31	0.43	0.5141
Bathing, Israel	0.02	0.11	−1.36	0.02	0.8848
Bathing, France	0.74	0.18	−2.92	5.80	0.0160
Bathing, Belgium	0.20	0.11	−0.89	1.01	0.3144
Dressing—upper body	−0.10	0.04	2.80	2.62	0.1057
Dressing—lower body, UK	0.67	0.09	−1.23	0.56	0.4531
Dressing—lower body, Sweden	0.41	0.10	−1.58	0.26	0.6090
Dressing—lower body, Italy	1.00	0.10	−0.65	0.43	0.5102
Dressing—lower body, Israel	0.82	0.10	−2.17	0.32	0.5730
Dressing—lower body, France	0.16	0.14	−1.36	6.05	0.0139
Dressing—lower body, Belgium	0.86	0.10	−0.97	0.07	0.7971
Toileting, UK	0.10	0.19	−0.58	3.99	0.0458
Toileting, Sweden	−0.04	0.15	−1.69	1.38	0.2403
Toileting, Italy	1.00	0.19	−1.54	0.93	0.3349
Toileting, Israel	−0.02	0.17	−1.22	2.18	0.1399
Toileting, France	0.29	0.14	−3.17	1.82	0.1769
Toileting, Belgium	0.37	0.18	−1.27	2.49	0.1145
Bladder management	−1.25	0.07	2.49	5.06	0.0245
Bowel management, UK	−1.00	0.17	1.21	2.65	0.1034
Bowel management, Sweden	−1.53	0.21	2.88	2.68	0.1018
Bowel management, Italy	−1.42	0.17	0.30	0.54	0.4606
Bowel management, Israel	−1.64	0.18	0.45	0.82	0.3644
Bowel management, France	−2.49	0.17	1.41	1.50	0.2204
Bowel management, Belgium	−1.79	0.16	−0.56	1.22	0.2695
Transfer bed, UK	−0.08	0.14	−2.05	6.19	0.0129
Transfer bed, Sweden	−0.23	0.10	−2.22	0.16	0.6895
Transfer bed, Italy	−0.07	0.09	−0.52	0.57	0.4503
Transfer bed, Israel	−0.20	0.10	−2.63	0.15	0.6962
Transfer bed, France	−1.15	0.09	−0.29	0.47	0.4952
Transfer bed, Belgium	−0.42	0.10	−1.13	0.85	0.3558
Transfer toilet, UK	−0.20	0.12	−1.45	3.42	0.0644
Transfer toilet, Italy	0.91	0.16	−2.85	3.58	0.0584
Transfer toilet, France	−0.64	0.11	−2.47	0.22	0.6416
Transfer toilet, Belgium	−0.34	0.10	−2.13	0.63	0.4279
Transfer tub, UK	0.87	0.09	1.04	1.34	0.2468
Transfer tub, Sweden	1.03	0.14	−1.52	0.07	0.7928
Transfer tub, Italy	2.10	0.19	−0.08	0.39	0.5343
Transfer tub, Israel	0.96	0.11	0.56	1.99	0.1587
Transfer tub, France	2.95	0.27	0.39	8.16	0.0043
Transfer tub, Belgium	2.63	0.24	−0.24	0.21	0.6506
Walk/wheelchair	0.82	0.06	0.96	0.66	0.4165
Stairs	2.49	0.07	0.19	1.87	0.1716

*Bonferroni corrected significance level $P = 0.0002$.

mates.[52] However, where patient databases are much larger, is it appropriate to take a sample so that the analysis can be done with known (and perhaps commonly agreed) levels of power and significance?

It has been argued that DIF is evaluated once the conditions for fit to the Rasch model have been satisfied.[63] However, it has also been argued that DIF is a breach of the assumptions of unidimensionality.[47] It is possible to have data which fit the model but which also display DIF. We have adopted the approach whereby we consider DIF as a violation to unidimensionality and thus one possible contribution to misfit, and thus deliberately refitted data to the model after adjustment for DIF. This adjustment involves a physical separation of the item for those countries displaying DIF and almost always results in an improvement of fit. An alternative approach would be to allow for that variation statistically by using the 2PL model (which adds a discrimination parameter).[64] Unfortunately, the latter has problems with convergence when dealing with relatively small item sets, missing values, and small sample sizes typically found in medical outcome studies.[49]

It is true that the approach used here, the 1-parameter Rasch model, places great demands upon the quality of measurement derived from the outcome measures. The Rasch measurement model is chosen because it conforms to the axioms of additive conjoint measurement and, given adequate fit to the model and local independence, can support the claim for an invariant unidimensional scale and provide a linear transformation of the ordinal data into interval measures. This class of measurement is a *requirement* for carrying out arithmetic operations such as the calculation of valid change scores. Otherwise, analysis should be restricted to appropriate nonparametric statistics.[30] Some analysis has used 2-parameter models as more appropriate for examining DIF in already developed measures.[58] However, in addition to the much larger sample size requirements for the 2PL and 3PL models (the latter which adds a guessing parameter),[65] it is known that almost 100% of the time their parameters violate interval scaling.[66] Thus, these models do not provide the quality of measurement which is required for most clinical outcome studies. Despite this, 2PL models have entered the mainstream of health status measurement.[67]

While this paper has addressed the issue of cross-cultural validity from a DIF perspective within the Rasch measurement model framework, there are many other non-IRT approaches, all well catalogued by Angoff.[46] A commonly used approach is the Mantel Haenzel procedure which estimates a common odds ratio across groups.[68] Newly developed packages such as SIBTEST facilitate analysis of DIF using this method.[69] However, it must be stressed that analysis of DIF within the Rasch framework is an analysis within a theoretical measurement framework which we believe to be crucial to fundamental measurement in health sciences.

For the future, developments based upon item banking, may bring the prospect of having different sets of items for different countries.[70,71] As long as these items are calibrated on the same underlying metric, any subset of items may be chosen, based on Rasch's notion of specific objectivity.[72] The strategy of allowing items to be unique by country, used in the analysis above, essentially establishes an item bank post hoc. In the future, new scales will use the power of Rasch analysis to allow for cultural variation by combining unique and common items in the development phase of the instrument.

In conclusion, where data are to be pooled for international clinical trials, it is essential to test for, and if necessary to adjust for DIF. Rasch's notion of specific objectivity means that, in practice, estimates of patient ability can be obtained when some (country-specific) items are omitted because of misfit to the model. We would argue that the requirements of outcome measurement in rehabilitation, specifically the calculation of change scores and other arithmetic operations on the data, demand that such analysis be undertaken within a framework that supports additive conjoint measurement. This can be operationalized, in a probabilistic framework, through the Rasch measurement model.

ACKNOWLEDGMENTS

This work was funded by the European Commission within its BIOMED 2 program under contract BMH4-CT98-3642.

We would like to thank the anonymous reviewers of this paper who made several valuable suggestions which contributed to our revision.

REFERENCES

1. Haigh R, Tennant A, Biering-Sørensen F, et al. The use of outcome measures in physical medicine and rehabilitation within Europe. *J Rehabil Med* 2001;33:273–278.
2. Word Health Organization. *The International Classification of Functioning Disability and Health.* WHO: Geneva; 2001.
3. Keith RA, Granger CV, Hamilton BB The functional independence measure: a new tool for rehabilitation. In: Eisenberg MG, Grzesiak RC, eds. *Advances in Clinical Rehabilitation.* Vol. 1. New York: Springer Publishing Co; 1987:6–18.
4. Tesio L, Franchignoni FP, Perucca L, et al. The influence of age of length of stay, functional independence and discharge destination of rehabilitation inpatients in Italy. *Disabil Rehabil.* 1996;18:502–508.
5. Grimby G, Thylander G. *Guide for Use of the Uniform Data Set for Medical Rehabilitation.* Buffalo: State University of New York at Buffalo; Swedish version, 1991.
6. Grimby G, Gudjonsson G, Rodhe M, et al. The functional independence measure in Sweden: experience for outcome measurement in rehabilitation medicine. *Scan J Rehabil Med.* 1996;28:51–62.
7. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Md State Med J.* 1965;14:61–65.
8. Folstein ME, Folstein SE, McHugh PR. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12:189–198.

9. Guillemin EG, Bombardier C, Beaton D. Cross-cultural adaptation of health related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol.* 1993;46:1417–1432.
10. Badia X, Alonso J. Rescaling the Spanish version of the Sickness Impact Profile: an opportunity for the assessment of cross-cultural equivalence. *J Clin Epidemiol.* 1995;48:949–957.
11. Wiesinger GF, Nuhr M, Quittan M, et al. Cross-cultural adaptation of the Roland-Morris Questionnaire for German-speaking patients with low back pain. *Spine.* 1999;24:1099–1103.
12. Salaffi F, Piva S, Barreca C, et al. Validation of the Italian version of the Arthritis Impact Measurement Scales 2 (ITALIAN-AIMS2) for patients with osteoarthritis of the knee. *Rheumatology.* 2000;39:720–727.
13. Brislin RW. Translation and content analysis of oral and written materials. In: HC Triandis, JW Berry, eds. *Handbook of Cross-Cultural Psychology- Methodology.* Vol 2. Boston: Allyn and Bacon Inc; 1980: 389–444.
14. Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine.* 2000;25: 3186–3191.
15. Mazzoni M, Ferroni L, Del Torto E, et al. Mini-Mental State Examination (MMSE): sensitivity in an Italian sample of patients with dementia. *Ital J Neurol Sci.* 1992;13:323–329.
16. Haglund Y, Edman G, Murelius O, et al. Does Swedish amateur boxing lead to chronic brain damage? I. a retrospective medical, neurological and personality trait study. *Acta Neurol Scand.* 1990;82:245–252.
17. Diard C, Ravaud JF, Held JP. French survey of postpolio sequelae: risk factors study and medical social outcome. *Am J Phys Med Rehabil.* 1994;73:264–267.
18. Eldar R, Ring H, Tshuwa M, et al. Quality of care for urinary incontinence in a rehabilitation setting for patients with stroke: simultaneous monitoring of process and outcome. *Int J Qual Health Care.* 2001;13: 57–61.
19. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: University of Chicago Press; 1980.
20. Andrich D. *Rasch Models for Measurement.* London: Sage; 1988.
21. Van der Linden WJ, Hambleton RK. Item response theory: brief history, common models, and extensions. In: Van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory.* New York: Springer; 1997.
22. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores.* Reading MA: Addison-Wesley; 1968.
23. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR, eds. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley; 1968:397–479.
24. Van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory.* New York: Springer; 1997:1–510.
25. Perline R, Wright BD, Wainer H. The Rasch model as additive conjoint measurement. *Appl Psychol Measure.* 1979;3:237–256.
26. Kaziz L, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care.* 1989;27:S178–S189.
27. Winship C, Mare RD. Regression models with ordinal data. *Am Soc Rev.* 1984;49:512–525.
28. Maxwell S, Delaney H. Measurement and statistics: an examination of construct validity. *Psychol Bull* 1985;97:85–93.
29. Embretson SE. Item response theory models and inferential bias in multiple group comparisons. *Appl Psychol Measure.* 1996;20:201–212.
30. Svensson E. Guidleines to statistical evaluation of data from rating scales and questionnaires. *J Rehabil Med.* 2001;33:47–48.
31. Van Alphen A, Halfens R, Hasman A, et al. Likert or Rasch? nothing is more applicable than a good theory. *J Adv Nurs.* 1994;20:196–201.
32. Banerji M, Smith RM, Dedrick RF. Dimensionality of an early childhood scale using Rasch analysis and confirmatory factor analysis. *J Outcome Measure.* 1997;1:56–85.
33. Smith RM. Fit analysis in latent trait measurement models. *J Appl Measure.* 2000;2:199–218.
34. Andrich D. Rating formulation for ordered response categories. *Psychometrica.* 1978;43:561–573.
35. Masters GN. A Rasch model for partial credit scoring. *Psychometrika.* 1982;47:149–174.
36. Silverstein B, Kilore KM, Fisher WP, et al. Applying psychometric criteria to functional assessment in medical rehabilitation: I. exploring unidimensionality. *Arch Phys Med Rehabil.* 1991;72:631–637.
37. Silverstein B, Fisher WP, Kilgore KM, et al. Applying psychometric criteria to functional assessment in medical rehabilitation: II. defining interval measures. *Arch Phys Med Rehabil.* 1992;73:507–518.
38. Fisher AG. The assessment of IADL motor skills: an application of many-faceted Rasch analysis. *Am J Occup Ther.* 1993;47:319–329.
39. Haley SM, McHorney CA, Ware JEJr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): I. unidimensionality and reproducibility of the Rasch item scale. *J Clin Epidemiol.* 1994;47:671–684.
40. Tennant A, Geddes JLM, Chamberlain MA. The Barthel Index: an ordinal score or interval level measure? *Clin Rehabil.* 1996;10:301–308.
41. Tennant A, Hillman M, Fear J, et al. Are we making the most of the Stanford Health Assessment Questionnaire? *Br J Rheumatol.* 1996;35: 574–578.
42. Prieto L, Alonso J, Lamarca R, et al. Rasch measurement for reducing the Items of the Nottingham Health Profile. *J Outcome Measure.* 1998;2:285–301.
43. Shulman JA, Wolfe EW. Development of a nutrition self-efficacy scale for prospective physicians. *J Appl Measure.* 2000;1:107–130.
44. Scheuneman JD. A method of assessing bias in test items. *J Educ Measure.* 1979;16:143–152.
45. Holland PW, Wainer H. *Differential Item Functioning.* Hillsdale. NJ: Lawrence Erlbaum; 1993.
46. Angoff WH. Perspectives on Differential Item Functioning methodology In: Holland PW, Wainer H, eds. *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum; 1993.
47. Dorans NJ, Holland PW. DIF detection and description: Mantel-Haenszel and standardisation. In: Holland PW, Wainer H, eds. *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1993: 36–66.
48. Fisher WP. Reliability statistics. *Rasch Measure Trans.* 1992;6:238.
49. Hagquist C, Andrich D. Is the Sense of Coherence instrument applicable on adolescents? a latent trait analysis using Rasch-modeling. *Pers Individ Dif.* 2003, in press.
50. Lange R, Irwin HJ, Houran J. Top-down purification of Tobacyk's revised paranormal belief scale. *Pers Individ Dif.* 2000;29:131–156.
51. Lange R, Thalbourne MA, Houran J, et al. Depressive response sets due to gender and culture-based Differential Item Functioning. *Pers Individ Dif.* 2002;33:937–954.
52. Linacre JM. Sample size and item calibration stability. *Rasch Measure Tran.* 1994;7:28.
53. Elashoff JD. *NQuery Advisor. Version 4.0. User's Guide.* Boston: Statistical Solutions; 2000.
54. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ.* 1995;310:170.
55. Tabachnick BG, Fidell LS, *Using Multivariate Statistics.* 4th ed. Boston: Allyn and Bacon; 2001.
56. Andrich D, Lyne A, Sheridon B, et al. *RUMM 2010.* Perth: RUMM Laboratory; 2000.
57. Stout WF. A nonparametric approach for assessing latent trait dimensionality. *Psycometrika.* 1987;55:293–326.
58. Hohl U, Grundman M, Salmon DP, et al. Mini-mental state examination and Mattis Dementia rating scale performance differ in Hispanic and non-Hispanic Alzheimer,s disease patients. *J Int Neuropsychol Soc.* 1999;5:301–307.
59. Teresi J, Golden R, Cross P, et al. Item bias in cognitive screening measures: comparisons of elderly white, Afro-American, Hispanic, and high and low education subgroups. *J Clin Epidemiol.* 1995;48:473–483.
60. Teresi JA, Holmes D, Ramirez M, et al. Performance of cognitive tests among different racial/ethnic and education groups: findings of differential item functioning and possible item bias. *J Ment Health Ageing.* 2001;7:79–89.
61. Küçükdeveci AA, Yavuzer G, Ehan AH, et al. Adaptation of the Functional Independence Measure for use in Turkey. *Clin Rehabil.* 2001;15:311–319.
62. Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. *J Outcome Measure.* 1998;2:66–78.
63. Smith RM. A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modelling.* 1996;3:25–40.

64. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. *J Rehabil Med.* 2001;33:47–48.
65. Streiner DL, Norman GR. *Health Measurement Scales.* Oxford: Oxford University Press; 1995.
66. Karabastos G. Axiomatic measurement theory as a basis for model selection in item response theory. 32nd Annual conference, Mathematical Psychology, 1999.
67. McHorney CA. Use of item response theory to link 3 modules of functional status items from the Asset and Health Dynamics Among the Oldest Old study. *Arch Phys Med Rehabil.* 2002;83:383–394.
68. Mantel N, Haenzel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Educ Measure.* 1959;22:719–748.
69. Stout WF, Roussos L. *SIBTEST Manual.* Urbana-Champaign, Ill: Statistical Laboratory for Educational and Psychological Measurement; 1996.
70. Dobby J, Duckworth D. *Objective assessment by means of item banking.* Schools Council Examination Bulletin 40. Evans/Methuen Educational, 1979.
71. Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. *Qual Life.* 1997;6:595–600.
72. Rasch G. On specific objectivity: an attempt at formalising the request for generality and validity of scientific statements. In: M Glegvad, ed. *The Danish Yearbook of Philosophy.* Copenhagen: Munksgaard; 1977: 58–94.