J Rehabil Med 2007; 39: 163–169

ORIGINAL REPORT

CROSS-DIAGNOSTIC VALIDITY OF THE SF-36 PHYSICAL FUNCTIONING SCALE IN PATIENTS WITH STROKE, MULTIPLE SCLEROSIS AND AMYOTROPHIC LATERAL SCLEROSIS: A STUDY USING RASCH ANALYSIS

Annet J. Dallmeijer[1,2], Vincent de Groot[1,2], Leo D. Roorda[3], Vera P. M. Schepers[4], Eline Lindeman[4], Leonard H. van den Berg[5], Anita Beelen[6], Joost Dekker[1,2] on behalf of the FuPro study group

From the [1]Department of Rehabilitation Medicine, [2]Institute for Research in Extramural Medicine (EMGO Institute), VU University Medical Center, [3]Jan van Bremen Institute, Center for Rheumatology and Rehabilitation, Amsterdam, [4]Rehabilitation Center De Hoogstraat and Rudolf Magnus Institute of Neuroscience, [5]Department of Neurology, University Medical Center, Utrecht and [6]Department of Rehabilitation, Academic Medical Center, Amsterdam, The Netherlands

Objective: **The aim of this study was to investigate unidimensionality and differential item functioning of the SF-36 physical functioning scale (PF10) in patients with various neurological disorders.**

Patients: **Patients post-stroke (n = 198), with multiple sclerosis (*n* = 151) and amyotrophic lateral sclerosis (*n* = 193) participated.**

Methods: **Unidimensionality of the PF10 within the patient groups was investigated by performing a separate Rasch analysis for each group. Differential item functioning was investigated in a pooled Rasch analysis of the 3 groups.**

Results: **Within each group, all items fitted the Rasch model, except the "bathing/dressing" item in the amyotrophic lateral sclerosis group. The pooled analysis showed inadequate fit to the Rasch model for one item ("walking several hundred metres"). Of the other 9 fitting items, 5 showed differential item functioning for stroke vs multiple sclerosis and amyotrophic lateral sclerosis, while no differential item functioning was found between multiple sclerosis and amyotrophic lateral sclerosis.**

Conclusion: **All items of the PF10, except one for the amyotrophic lateral sclerosis group, form a unidimensional scale, supporting the use of a sum score as a measure of physical functioning within these diagnostic groups. When comparing the data of patients after stroke, with that of patients with multiple sclerosis and/or amyotrophic lateral sclerosis patients, adjustments for differential item functioning are required.**

Key words: amyotrophic lateral sclerosis, cross-diagnostic validity, differential item functioning, multiple sclerosis, physical functioning, Rasch model, stroke

J Rehabil Med 2007; 39: 163–169

Correspondence address: Annet J. Dallmeijer, Department of Rehabilitation Medicine, VU University Medical Center, PO Box 7057, NL-1007 MB Amsterdam, The Netherlands.
E-mail: a.dallmeijer@vumc.nl

Submitted March 21, 2006; accepted October 18, 2006.

INTRODUCTION

Describing and examining the outcomes of rehabilitation treatment is becoming increasingly important for evidence-based practice and for policy-makers. Adequate measurement instruments to assess disability are therefore essential. The 36-item Short Form health survey of the Medical Outcome Study (SF-36) is a generic measurement instrument that was developed to measure health-related quality of life in patients and healthy persons. It consists of 8 sub-scales that are often used separately as outcome measures of various aspects of health-related quality of life (1–4). The 10-item Physical Functioning sub-scale (PF10) is of specific interest for application in rehabilitation because of its focus on physical disability (5), measured at the level of activity according to the International Classification of Functioning (6). It is important that instruments that are used to measure (changes in) physical disability in different patient groups fulfil several psychometric requirements (7). Among those, it is important that all individual items of multi-item scales measure the same underlying construct on a hierarchical scale (show unidimensionality). This is particularly important when ordinal item ratings are summed up to yield a total score (8, 9). If the items form a hierarchical measurement scale, less disabled subjects are more likely to pass difficult items than more disabled subjects, and vice versa. When a measurement instrument fulfils this requirement, unidimensionality of the scale is supported, and the calculation of total sum scores is allowed. These characteristics should be tested in each patient group before the instrument is used.

In rehabilitation, generic measures are frequently used for pooling or comparing the data of different patient groups (10). In this case, the items should function the same across all the patient groups. When items do not function similarly across groups, this is called differential item functioning (DIF). Although DIF is being increasingly investigated for the cross-cultural validation of (translations of) questionnaires used in rehabilitation (11–13), very little attention has been paid to disease-specific characteristics as a cause of DIF (cross-diagnostic validity). The Rasch measurement model is a method that can be used to investigate unidimensionality and DIF. In Rasch analysis, ordinal scores are converted into interval measures of person ability and item difficulty along a common measurement continuum (14). This makes it possible to make a detailed investigation of the unidimensionality of a scale, as well as DIF across groups (13, 15).

 ISSN 1650-1977
DOI: 10.2340/16501977-0024

A few studies have investigated the unidimensionality of the PF10 (5, 16–19) and DIF across groups of patients (5, 17, 19) or countries (18). In some of these studies it was found that the items represent a unidimensional construct in most patient groups and in the general populations of several European countries (18), and that the item hierarchy was reproducible across different patient groups (5) and countries (18). In contrast, other studies concluded that the items of the PF10 do not form a unidimensional scale in some patient groups (16, 17), and showed that not all PF10 items function similarly across diagnostic groups (17, 19). These contradictory results may be due to differences in the type of patient groups or countries. It is therefore not clear in which patient groups the PF10 can be used and whether the results of the PF10 can be pooled or compared between different neurological patient groups.

The aims of this study were: (*i*) to investigate the unidimensionality of the PF10 in patients with stroke, multiple sclerosis (MS) and amyotrophic lateral sclerosis (ALS), within the groups; and (*ii*) to investigate DIF across the groups.

METHODS

Subjects and design

This investigation was performed as part of a 3-year follow-up study on the functional prognosis of patients with neurological disorders. The SF-36 scores at 6 months after inclusion were used for the analyses of this study. Three patient groups with different neurological disorders were investigated: (*i*) patients with a first-ever supratentorial stroke, who had been admitted for inpatient rehabilitation, but had left the rehabilitation centre at the time of measurement (6 months post-stroke), (*ii*) patients with recently diagnosed MS, 6–12 months after diagnosis, and (*iii*) patients with probable or definite ALS, according to the revised El Escorial (20) criteria, who were attending a outpatient clinic of an university hospital (department of neurology or rehabilitation medicine), with a disease duration from onset of symptoms ranging from 6 months to 5 years (average 1.5 years).

Physical Functioning scale of the SF-36

The PF10 is 1 of the 8 sub-scales of the SF-36, each of which measures a different construct of health-related quality of life. This self-reported health status measurement instrument was developed in the USA with data from the Medical Outcome Study (1). The PF10 consists of 10 items that assess the extent of health-related limitations in physical functioning. Its reliability and validity have been supported in previous studies (3, 4). The items are scored on a 3-point Likert scale (1 = limited a lot, 2 = limited a little, 3 = not limited at all). The scale was designed to be applicable to general populations as well as patients with acute and chronic diseases.

Fit with the Rasch model

Rasch analysis was applied to investigate the unidimensionality of the PF10 scale within the 3 patient groups, and to investigate DIF between the groups. If the data fit the Rasch model, item scores can be used to determine item difficulty and person ability on a common interval scale. Person ability and item difficulty are expressed in log-odd units (logits). The Rasch model assumes that easier items are more likely to be passed, and that less disabled persons are more likely to pass an item than more disabled persons. For this analysis we used the Rasch partial credit model, which allows the intervals between the thresholds of the answer categories (point of equal probability between 2 adjacent categories) to vary between items. Fit of the items with the model is investigated by examining the fit residuals and χ^2 statistics (21). The standardized fit residuals of the responses of all persons to the item (distributed as a Z statistic with a mean of 0 and a standard deviation (SD) of 1 for perfect fit with the model) indicate the deviation of the observed item score from the model-expected scores. High fit residuals (> 2.5) indicate that the observed scores are higher or lower than the model expected values. Low fit residuals (< –2.5) indicate that items are redundant (21). The χ^2 item-trait interaction statistic was applied to investigate whether the fit with the model was satisfactory along the whole scale (invariance of the scale at different levels of ability). To investigate this, the group was sub-divided, based on level of ability, into different class intervals of approximately the same number of patients along the trait, with around 50–60 patients in each class interval. Non-significant χ^2 statistics indicate that items fit the model at all levels of ability (21). The level of significance was adjusted by the Bonferroni procedure for multiple testing ($p = 0.005$). The χ^2 statistics were calculated for each item individually, and for the overall scale (sum of the item χ^2). If all items fit the Rasch model, an additional principal component analysis of the residuals is required to support unidimensionality. The first residual component should account for less than 40% of the variance to support unidimensionality of the scale (21). Analyses were performed with the RUMM2020 software package (21).

Rasch analysis within each patient group: unidimensionality

We first investigated whether the scale fulfilled the criteria for unidimensionality in each separate group (research question one). We investigated whether individual items fitted the Rasch model, by examining the fit residuals and χ^2 statistics, as described above. Non-fitting items were removed until an overall fitting model (indicated by a non-significant overall χ^2) was achieved for each group. This was followed by a principal component analysis for each separate group.

Rasch analysis for the total group: differential item functioning

Subsequently, to investigate whether items functioned similarly across patient groups, DIF was investigated by performing a Rasch analysis on the total group (3 patient groups combined). Because items have to fit the Rasch model before DIF can be investigated, non-fitting items were first removed. DIF was investigated in all remaining fitting items by performing a two-way analysis of variance (ANOVA) of the residuals, using patient group (stroke, MS, ALS) and class interval (ability groups, as described above) as factors (21). DIF was identified by a significant ANOVA main effect of patient group (i.e. a constant difference between groups: uniform DIF) or an ANOVA interaction effect of group and class interval (i.e. the difference between groups varied across the trait: non-uniform DIF) (21).

To make group comparisons possible, we adjusted for DIF by sub-dividing items that showed DIF, into 2 or 3 group-specific items, as described by Tennant et al. (13). The final overall analysis was performed with the items that showed no DIF, combined with the group-specific items for the items that showed DIF. Again, non-fitting items were removed until a fitting model was achieved.

The PF10 Rasch estimates for person ability were transformed to a score from 0 to 100. Mean patient group values were calculated using the adjusted PF10 estimates (with the split DIF items) and using the PF10 estimates without adjustments (no split items). To determine the effect of the adjustment procedure, an ANOVA for repeated measures was performed, with adjusted vs non-adjusted as within-subject factor, and patient group as between subject-factor.

RESULTS

Subjects

The subjects comprised: 198 patients with stroke, 151 patients with MS and 194 patients with ALS. The mean age (SD) was 57.3 (11.8), 38.4 (9.8) and 57.9 (10.8) years, and the percentage of females was 39%, 63% and 32% for patients with stroke, MS and ALS, respectively.

Table I. *Descriptive 10-item Physical Functioning (PF) sub-scale*

	Stroke *n* = 198	MS *n* = 151	ALS *n* = 193
Floor (%)	0.5	0.7	11.9
Ceiling (%)	3.5	15.2	2.6
PF10 Raw score (mean (SD))	58.4 (25.4)	69.0 (26.8)	40.5 (31.4)
PF10 Rasch estimates (no adjustment for DIF) (mean (SD))	50.4 (19.8)	60.8 (24.4)	38.0 (25.5)
PF10 Rasch estimates with adjustment for DIF (split items) (mean (SD))	49.7 (19.6)	55.1 (22.2)	33.9 (23.5)

MS: multiple sclerosis; ALS: amyotrophic lateral sclerosis; SD: standard deviation; DIF: differential item functioning.

In the stroke group, 116 (59%) patients had left hemisphere lesions, 80 patients (40%) had right hemisphere lesions and 2 (1%) had bilateral lesions. A total of 147 (74%) of the patients had a cerebral infarction, and 50 (25%) patients had a hemorrhagic stroke, 17 of whom had a subarachnoid haemorrhage. For one patient the type of lesion was unknown. For 28 patients who had signs of aphasia, the SF-36 was completed by interviewing proxies (relatives). To investigate the effect of including the 28 proxies, we performed the analysis with and without proxies, but the results remained the same. It was therefore decided to include these patients in the study.

In the MS group, 115 (76%) patients had relapsing-remitting MS, 22 (15%) had primary-progressive MS, 8 (5%) had secondary-progressive MS. For 6 (4%) patients the type of MS could not be determined at the time of diagnosis.

In the ALS group 42 (22%) patients had a bulbar onset of the disease, and 150 (77%) had a spinal onset. For one (1%) patient the onset was unknown.

Missing values were treated as described in the SF-36 manual (22) (the mean value of the non-missing items was used to calculate sub-scale scores if 50% or more of the sub-scale items were not missing). Following this procedure, one subject in the ALS group was excluded because of missing items. The mean raw PF10 scores, as well as the floor and ceiling effects for each group are shown in Table I. Except for the ceiling effect of 15% for the MS group, all the other floor/ceiling effects were negligible.

Rasch analysis within each group: unidimensionality

Item difficulties and fit residuals for each patient group are shown in Table II. The item difficulties are expressed on an interval scale in logits. Because the PF10 items have 3 answer categories, the item difficulty is the mean of the 2 thresholds (the point of equal probability between 2 adjacent answer categories). The answer categories showed no disordered

Table II. *Difficulty and fit of the 10-item Physical Functioning (PF) sub-scale items for each patient group analysed separately*

	Stroke					MS					ALS				
Item	Location	SE	Fit residual	χ^2	*p*-value*	Location	SE	Fit residual	χ^2	*p*-value*	Location	SE	Fit residual	χ^2	*p*-value*
PF1: Vigorous activities	4.49	0.22	–0.58	6.59	0.086	4.15	0.22	2.45	0.83	0.661	3.21	0.22	–1.47	3.38	0.184
PF2: Moderate activities	1.14	0.15	0.36	0.92	0.820	0.32	0.21	–0.53	0.59	0.744	0.25	0.16	–1.44	0.94	0.626
PF3: Lifting/ carrying groceries	1.14	0.14	–0.59	4.05	0.256	–0.16	0.21	–1.94	4.66	0.097	0.25	0.16	–0.12	0.18	0.912
PF4: Climbing several flights of stairs	–0.65	0.15	–1.29	2.48	0.479	0.81	0.20	–0.61	0.91	0.636	0.04	0.16	–1.47	3.12	0.210
PF5: Climbing 1 flight of stairs	–1.68	0.16	–0.56	1.61	0.657	–1.65	0.23	–1.00	2.93	0.231	–1.13	0.17	–1.71	1.64	0.441
PF6: Bending/ kneeling/stooping	0.06	0.14	2.41	**13.70**	**0.003**	–0.22	0.20	2.51	9.98	0.007	–0.30	0.16	0.30	1.66	0.437
PF7: Walking more than 1 km	0.78	0.13	–0.66	1.10	0.776	1.30	0.19	–0.99	4.20	0.123	0.86	0.16	–1.47	4.62	0.099
PF8: Walking several 100 m	–0.68	0.14	–2.43	10.63	0.014	–0.82	0.21	–2.15	1.76	0.416	–0.49	0.16	–2.44	5.05	0.080
PF9: Walking 100 m	–1.93	0.17	–1.44	5.52	0.138	–1.54	0.21	–1.60	0.86	0.652	–1.57	0.16	–1.13	1.66	0.435
PF10: Bathing/ dressing	–2.67	0.18	0.01	2.38	0.498	–2.19	0.23	–0.30	0.96	0.619	–1.13	0.16	**4.37**	**23.57**	**0.000**
Total	–	–	–	48.98	0.016	–	–	–	27.66	0.118	–	–	–	45.83	0.001

*Adjusted *p*-value of 0.005 was applied; misfit in bold type.
MS: multiple sclerosis; ALS: amyotrophic lateral sclerosis; SE: standard error.

thresholds for any of the items, indicating that the difficulty of the answer categories was as expected.

In the stroke and the MS groups, all items fitted the Rasch model. In the stroke group, the item "Bending, kneeling, stooping" (item 6) showed a borderline χ^2 value (13.7, $p = 0.003$). However, because the fit residual of this item was within the accepted range (2.4), and the overall fit of the scale was acceptable ($\chi^2 = 48.98$, $p = 0.016$), we did not remove this item from the scale. In the MS group, all 10 items fitted the model (overall $\chi^2 = 27.66$, $p = 0.118$). In the ALS group, the item "bathing and dressing" (item 10) showed considerable misfit for both fit residual (4.37) and the χ^2 statistic (23.57, $p < 0.001$, Table II). After removing this item from the scale, the remaining 9 items in the ALS group all fitted the model, showing a good overall scale fit ($\chi^2 = 17.24$, $p = 0.507$). Principal component analysis of the residuals showed that the first residual component accounted for only 22%, 20% and 27% of the variance in the stroke, MS and ALS groups, respectively, supporting unidimensionality.

Table IV. *10-item Physical Functioning (PF) sub-scale adjusted for differential item functioning by diagnosis (with split items)*

Item	Location	SE	Fit residual	χ^2	*p*-value
PF10 MS/stroke	–2.70	0.14	–0.39	2.15	0.542
PF9	–2.15	0.10	–1.55	5.91	0.116
PF5	–1.99	0.10	–1.75	5.71	0.127
PF6 MS/ALS	–0.85	0.12	1.64	3.13	0.373
PF4 stroke	–0.85	0.15	–1.44	3.22	0.359
PF3 MS/ALS	–0.51	0.13	–1.13	1.99	0.574
PF2 MS/ALS	–0.32	0.13	–1.09	1.49	0.686
PF4 MS/ALS	–0.23	0.13	–1.59	4.20	0.241
PF6 stroke	–0.13	0.14	2.29	10.29	0.016
PF7	0.50	0.09	–0.97	1.34	0.719
PF3 stroke	0.95	0.14	–0.55	5.29	0.152
PF2 stroke	0.96	0.15	0.22	1.68	0.641
PF1 MS/ALS	2.98	0.15	0.68	0.72	0.869
PF1 stroke	4.33	0.22	–0.56	4.38	0.224

MS: multiple sclerosis; ALS: amyotrophic lateral sclerosis; SE: standard error.

Rasch analysis for the total group: differential item functioning

Initial analysis of the 3 groups combined showed that two items (item 6: "bending, kneeling, stooping" and item 10: "bathing and dressing") had fit residuals exceeding 2.5, and two other items (item 8: "walking several hundred metres" and item 9: "walking 100 metres") had fit residuals of less than –2.5. Items 8 and 10 also had significant χ^2 item-trait interaction statistics (Table III), indicating that the fit of the model was not the same for the different levels of ability (class intervals). All 10 items showed ordered thresholds between answer categories.

First, the 2 most misfitting items (items 8 and 10) were subdivided into group-specific items, and the analysis was repeated. All 3 group-specific items of item 8 showed large fit residuals and were therefore removed from the total sample for all groups. For ALS, item 10 showed a very large fit residual (4.35), while the fit residual and χ^2 statistic of item 10 for stroke and MS were within the acceptable range. Therefore, item 10 for ALS was removed from the sample. After repeating the analysis with the remaining items, all items fitted the model (both fit residuals and χ^2 statistics).

DIF was then investigated in the remaining items. The item showing the most serious DIF was first adjusted by defining 2 or 3 group-specific items, and then repeating the analysis. This procedure was followed until all items were free of DIF and fitted the model. Finally, 4 items (items 5, 7, 9, and item 10 for stroke and MS only) showed no DIF. Five items (items 1, 2, 3, 4 and 6) showed DIF between the stroke group and the MS and ALS groups. For these items, the stroke group-specific items were included in the final analysis. No DIF was found between the MS and the ALS group. In the final analysis, including split items (stroke vs ALS/MS) for items 1, 2, 3, 4 and 6, the fit residuals ranged from –1.75 to 2.29, and the χ^2 statistics were all non-significant, with a non-significant overall scale χ^2 (51.48, $p = 0.15$) (Table IV). Unidimensionality of the scale was supported by principal component analysis of the fit residuals, showing that the first residual component accounted for only 19% of the variance. The final item difficulties for each group are shown in Fig. 1.

Mean PF10 person abilities for values calculated with and without adjusting for DIF by splitting the DIF items are shown

Table III. *Difficulty and fit of the 10-item Physical Functioning (PF) sub-scale items for the total group*

Item	Location	SE	Fit residual	χ^2	*p*-value*
PF1: Vigorous activities	3.73	0.12	0.42	9.78	0.201
PF2: Moderate activities	0.61	0.10	–0.91	6.66	0.465
PF3: Lifting/carrying groceries	0.54	0.09	–0.67	20.66	0.004
PF4: Climbing several flights of stairs	–0.09	0.09	–0.91	11.10	0.134
PF5: Climbing one flight of stairs	–1.48	0.10	–1.75	10.77	0.149
PF6: Bending/kneeling/stooping	–0.13	0.09	**2.62**	6.06	0.533
PF7: Walking more than one km	0.90	0.09	–1.81	9.16	0.241
PF8: Walking several hundred metres	–0.64	0.09	**–4.61**	**30.23**	**0.000**
PF9: Walking 100 metres	–1.69	0.10	**–2.73**	20.03	0.005
PF10: Bathing/dressing	–1.77	0.10	**2.90**	**33.55**	**0.000**
Total	–	–	–	157.99	0.000

*Adjusted *p*-value of 0.005 was applied; misfit in bold type.
SE: standard error.

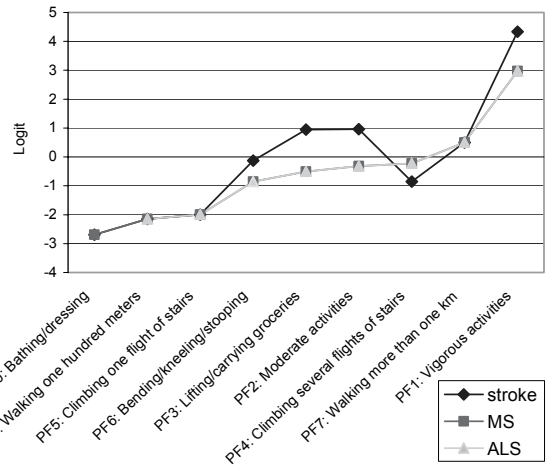

Fig. 1. Difficulties (expressed in logits) of 10-item Physical Functioning sub-scale (PF10) items for patients with stroke, multiple sclerosis (MS) and amyotrophic lateral sclerosis (ALS) after a combined Rasch analysis. Note that items 8 and 10 (for ALS only) have been removed.

in Table I. There was a significant ANOVA interaction effect of adjustment (yes or no) and group ($F = 388.7$, $p < 0.001$), indicating that the adjustment procedure affected the differences between groups.

DISCUSSION

The objective of this study was to investigate the unidimensionality of the PF10 scale in patients with stroke, MS and ALS, and to determine whether PF10 scores can be compared between groups or pooled for combined analysis. The results showed that, with the exception of item 10 in the ALS group, the PF10 items form a unidimensional scale within each patient group, which confirms that the items measure one underlying construct and that the summation of individual item scores to create a total score is justified in patients with stroke and MS, and, with exception of item 10, also in ALS patients. The results also demonstrated that some items showed DIF for the stroke group vs the MS and ALS groups, indicating that adjustments for DIF are necessary when pooling or comparing data between these groups.

Unidimensionality

Scaling assumptions of the PF10 of the SF-36 have been extensively investigated, using methods based on the classical test theory, such as Cronbach's alphas, item-scale correlations and factor analysis, all showing excellent internal consistency and high item-scale correlations across different patient groups and countries (3, 4, 23–26). Other studies that have investigated unidimensionality with the more stringent Rasch model showed that some items did not fit the Rasch model (5, 18, 19), but nevertheless concluded that unidimensionality was supported, because the proportion of subjects with unexpected responses (and thus causing misfit of these items) was sufficiently low (< 5%) (5). Others concluded that unidimensionality could not be supported because of some misfitting items (16, 17). In contrast, our results showed that, except for one item in the ALS group, all items fitted the Rasch model within the groups, indicating that the items form a unidimensional scale. The larger number of items showing misfit in earlier studies (5, 16–18) may be due to differences in group characteristics (different patient groups or a healthy population). Another possible explanation is that we used the more flexible partial credit model that allows for variations between items in the threshold levels of adjacent categories, while most previous studies used the Rasch rating scale model (5, 17, 18).

The large misfit for item 10 in the ALS group was in agreement with Jenkinson et al. (17). Other studies also reported large fit residuals for item 10 in other patient groups (5, 16, 17), including patients with stroke (19), and in the general population across several countries (18). This misfit may be due to the fact that this item measures more than one activity (i.e. bathing or dressing). The good fit for all items of the MS group is in agreement with 2 former studies, using Rasch analysis (19) and a non-parametric item response theory model (Mokken analysis) (27) both reporting that the PF10 is a strong unidimensional, hierarchical scale. The negative misfit of item 8 ("walking several hundred metres") in the total group may be explained by a large interdependency of this item with other items.

Our results showed that, in accordance with the original scale in several diagnostic groups (5), and in the general Dutch population (18), item 1 was the most difficult, and items 9 and 10 were the easiest. The intermediate items showed more variation in item hierarchy, when compared with US patients with several chronic conditions (5), than when compared with the general (Dutch) population (18). Comparison with the latter study showed that item 7 ("walking more than one mile") was more difficult in our study (rank 2) than in Dutch healthy persons (rank 6) (18). It was also found that the interval between the most difficult item (item 1) and the adjacent item (item 2) is much larger than the item intervals in the mid-range. This finding is consistent with the results of previous studies (5, 18, 28, 29), and leads to an underestimation of change in physical functioning score (less numerical gain) at the high end of the score distribution when using the raw ordinal PF10 scores (29). It has been argued that, because of the unequal intervals between raw item scores, a Rasch-based score would improve discrimination and sensitivity to change in the PF10 scale (18, 28, 29). This should be further investigated in patients with stroke, MS and ALS.

Differential item functioning

The results show that there were differences in item difficulty between stroke patients, and patients with MS or ALS. More specifically, 4 out of the 9 fitting PF10 items were more difficult for stroke patients, and 1 item was easier than for MS and ALS patients, while for MS and ALS patients all items demonstrated similar characteristics. For example, item 3 ("lifting/carrying groceries") was more difficult for patients

with stroke than for patients with MS and ALS. This may be clinically explained by the unilateral impairment of the arms of stroke patients. These results suggest that caution should be taken when comparing the results of the PF10 scale between stroke patients and the other patient groups, or when pooling the data in a single analysis.

We used a method to adjust for DIF, described by Tennant et al. (13), that allows pooling of the data of different (patient) groups when DIF is present. However, in order to apply this method, some common items (that do not show DIF) are required. In our study, 3 items (5, 7 and 9) showed no DIF between groups, so these could be used to link the groups (13). In the pooled analysis we had to delete the ALS group-specific item 10, and also item 8 because the fit residuals were too large. The other items, including the 5 stroke-specific items, were used to describe the physical functioning of the 3 patient groups on one common measurement continuum, and can therefore used to compare physical functioning between the groups.

Our results differed from the results of an earlier study investigating the item response patterns in different patient groups by applying Rasch analysis (5). The authors concluded that the item structure was reproducible across patient groups, but they nonetheless also reported large differences in item difficulty between the groups. Another study investigating differences in item responses between MS patients and patients with other chronic diseases using Mokken analysis, reported no differences in item hierarchy (27). In contrast, Jenkinson et al. (17) and Bode et al. (19) reported results that were comparable to the results of the present study, showing differences in item difficulty between patients with ALS or MS and other diseases. These results confirm that differences in item difficulty do exist between patient groups, and should be investigated before comparing PF10 data. In addition, our results showed that group differences are influenced by the adjustments for DIF, indicating that adjustments for DIF are required when comparing stroke with ALS or MS.

It is concluded that the PF10 is a useful scale for measuring physical functioning in patients with neurological disorders during rehabilitation. The results also indicate that it is not appropriate to compare or combine the PF10 scale of patients after stroke with that of patients with MS and ALS without adjusting for DIF. However, the PF10 data of patients with MS and ALS can be compared without adjustments.

ACKNOWLEDGEMENTS

This investigation was performed as part of the "Functional prognostication and disability study on neurological disorders", supervised by the Department of Rehabilitation Medicine of the VU University Medical Center, Amsterdam, and was funded by the Netherlands Organization for Health Research and Development (grant: 1435.0001).

FuPro study group: G. J. Lankhorst, J. Dekker, A. J. Dallmeijer, M. J. IJzerman, H. Beckerman, V. de Groot: VU University Medical Center, Amsterdam (project co-ordination); A. J. H. Prevo, E. Lindeman, V. P. M. Schepers: University Medical Center, Utrecht; H. J. Stam, E. Odding, B. van Baalen: Erasmus Medical Center, Rotterdam; A. Beelen: Academic Medical Center, Amsterdam, the Netherlands.

REFERENCES

1. Ware JE, Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30: 473–483.
2. Ware JE, Jr. SF-36 health survey update. Spine 2000; 25: 3130–3139.
3. McHorney CA, Ware JE, Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993; 31: 247–263.
4. McHorney CA, Ware JE, Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994; 32: 40–66.
5. Haley SM, McHorney CA, Ware JE, Jr. 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.
6. WHO. International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.
7. Dekker J, Dallmeijer AJ, Lankhorst GJ. Clinimetrics in rehabilitation medicine: current issues in developing and applying measurement instruments. J Rehabil Med 2005; 37: 193–201.
8. Silverstein B, Kilgore KM, Fisher WP, Harley JP, Harvey RF. Applying psychometric criteria to functional assessment in medical rehabilitation: I. Exploring unidimensionality. Arch Phys Med Rehabil 1991; 72: 631–637.
9. Nunnally JC, Bernstein IA. Psychometric theory. 3rd edn. New York: McGraw-Hill; 1994.
10. Haigh R, Tennant A, Biering-Sorensen 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.
11. Lundgren-Nilsson A, Grimby G, Ring H, Tesio L, Lawton G, Slade A, et al. Cross-cultural validity of functional independence measure items in stroke: a study using Rasch analysis. J Rehabil Med 2005; 37: 23–31.
12. Roorda LD, Jones CA, Waltz M, Lankhorst GJ, Bouter LM, van der Eijken JW, et al. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann Rheum Dis 2004; 63: 36–42.
13. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, 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 1: I37–I48.
14. Wright BD, Linacre JM, Smith RM, Heinemann AW, Granger CV. FIM measurement properties and Rasch model details. Scand J Rehabil Med 1997; 29: 267–272.
15. Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. J Rehabil Med 2003; 35: 105–115.
16. Davidson M, Keating JL, Eyres S. A low back-specific version of the SF-36 Physical Functioning scale. Spine 2004; 29: 586–594.
17. Jenkinson C, Fitzpatrick R, Garratt A, Peto V, Stewart-Brown S. Can item response theory reduce patient burden when measuring health status in neurological disorders? Results from Rasch analysis of the SF-36 physical functioning scale (PF-10). J Neurol Neurosurg Psychiatry 2001; 71: 220–224.
18. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, et al. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998; 51: 1203–1214.
19. Bode RK, Lai JS, Cella D, Heinemann AW. Issues in the development of an item bank. Arch Phys Med Rehabil 2003; 84 Suppl 2: S52–S60.
20. Brooks BR, Miller RG, Swash M, Munsat TL. El Escorial revisited:

revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotroph Lateral Scler Other Motor Neuron Disord 2000; 1: 293–299.
21. Andrich D, Sheridan BS, Luo G. RUMM2020: Rasch Unidimensional Models for Measurement. Perth Western Australia: RUMM Laboratory; 2002.
22. Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 Health survey manual and interpretation guide. Boston MA: New England Medical Center, The Health Institute; 1993.
23. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R, et al. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol 1998; 51: 1055–1068.
24. Failde I, Ramos I. Validity and reliability of the SF-36 Health Survey Questionnaire in patients with coronary artery disease. J Clin Epidemiol 2000; 53: 359–365.
25. Gandek B, Ware JE, Jr, Aaronson NK, Alonso J, Apolone G, Bjorner J, et al. Tests of data quality, scaling assumptions, and reliability of the SF-36 in eleven countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998; 51: 1149–1158.
26. Ware JE, Jr, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, et al. The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998; 51: 1159–1165.
27. Moorer P, Suurmeije T, Foets M, Molenaar IW. Psychometric properties of the RAND-36 among three chronic diseases (multiple sclerosis, rheumatic diseases and COPD) in The Netherlands. Qual Life Res 2001; 10: 637–645.
28. McHorney CA, Haley SM, Ware JE, Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol 1997; 50: 451–461.
29. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol 1996; 49: 711–717.