Comparison of 4 Functional Indexes in Psoriatic Arthritis with Axial or Peripheral Disease Subgroups Using Rasch Analyses

YING-YING LEUNG, LAI-SHAN TAM, EMILY WAI-LIN KUN, KWOK-WAH HO, and EDMUND KWOK-MING LI

ABSTRACT. ***Objective.*** Rasch item response theory analysis is essential in evaluating measurement tools in specific disease cohorts. We compared the performance of 4 functional indexes in patients with psoriatic arthritis (PsA) in axial or peripheral disease subgroups.

Methods. A cross-sectional study was performed in a single center. Functional outcomes assessed by the Health Assessment Questionnaire (HAQ), Bath Ankylosing Spondylitis Functional Index (BASFI), Dougados Functional Index (FI), and the physical functioning scale of the Medical Outcome Study Short-form 36 (SF-36-PF) were analyzed by the Rasch model for item fit, item separation, measurement span, and distribution properties. Patient subgroups with axial or peripheral disease were analyzed for differential item functioning (DIF).

Results. One hundred eight patients with PsA were assessed. The 4 functional indexes were highly correlated with each other and moderately correlated with patients' perception of health and pain scores. Floor effects were less marked in SF-36-PF. The 4 indexes satisfied the unidimensionality assumption of the Rasch model. HAQ and SF-36-PF had better information-weighted fit statistics (INFIT) and outlier-sensitive (OUTFIT) statistics. HAQ had the poorest item separation. SF-36-PF had the highest item separation (6.99), reliability (0.85), and the longest span of item threshold (9.03 logits). Only 1 and 2 items in BASFI and Dougados-FI had DIF in patients with sacroiliitis.

Conclusion. HAQ, BASFI, Dougados-FI, and SF-36-PF provide unidimensional measures of functional disability in PsA. SF-36-PF was the best in terms of less floor effect, highest item separation, longest span of item threshold, and better distributional properties. BASFI and Dougados-FI behaved similarly in patients with and without sacroiliitis and conferred no superiority in patients with axial disease. (First Release July 1 2008; J Rheumatol 2008;35:1613–21)

Key Indexing Terms:
PSORIATIC ARTHRITIS FUNCTIONAL ASSESSMENT RASCH ANALYSIS

Psoriatic arthritis (PsA) is an inflammatory arthritis associated with psoriasis. It affects young adults in their working ages, causing deformities, impaired quality of life, and poor physical function[1-3]. Functional disability is one of the major outcome domains used in randomized controlled trials, observational studies, and daily practice.

The Health Assessment Questionnaire (HAQ) was developed for rheumatoid arthritis (RA)[4]. In PsA, HAQ has been used for measuring physical function in observational studies and clinical trials[5,6]. HAQ was shown to be strongly correlated to grip strength, American College of Rheumatology functional class, and fibromyalgia tender points, but was correlated only moderately to other measures of disease activity and poorly correlated with measures of spinal mobility. HAQ scores have been shown to be adequately sensitive to change after effective biologic therapies for peripheral joints in PsA[7,8]. A modification of the HAQ for spondyloarthropathies (HAQ-S)[9] was found to behave similarly to the original HAQ, even in people with the axial pattern of disease. The HAQ has also been criticized for its floor effects, nonlinearity, and confusing items[10].

The Bath Ankylosing Spondylitis Functional Index (BASFI) has been validated and adopted by the Assessment in AS (ASAS) Working Group for measuring functional outcome in ankylosing spondylitis (AS). It has been used in observational trials in spondyloarthropathies, which included patients with PsA with some evidence of validity[11]. It may be useful in PsA and may represent functional impairment as a result of spinal involvement in the PsA cohort. The Medical Outcome Study Short-Form Health Survey (SF-36) is a generic measurement of quality of life. It has been validated and used in epidemiology and treatment trials in PsA[1,7,8,12]. The 10 items on physical functioning (SF-36-PF) were extracted for analysis. The Dougados Functional

From the Department of Medicine and Geriatrics, Tai Po Hospital; Department of Medicine and Therapeutics, Princes of Wales Hospital, Chinese University of Hong Kong; and Department of Statistics, Chinese University of Hong Kong, Hong Kong, China.

Y-Y. Leung, MD, Resident Specialist; L-S. Tam, MD, Department of Medicine and Therapeutics, Princes of Wales Hospital; E.W-L. Kun, MD; K-W. Ho, PhD, Department of Statistics, Chinese University of Hong Kong; E.K-M. Li, MD, Department of Medicine and Therapeutics, Princes of Wales Hospital.

Address reprint requests to Dr. Y-Y. Leung, Department of Medicine and Geriatrics, Tai Po Hospital, HKSAR 9 Chuen On Road, Tai Po, NT, Hong Kong SAR. E-mail: katyccc@hotmail.com

Accepted for publication March 23, 2008.

Index (Dougados-FI) has been used in spondyloarthropathy with some validity[11,13].

Despite the application of these indexes in PsA, their validity in measuring physical disability in PsA may not be completely secured. Construct or criterion validity is not well studied; neither do we know if different patterns of PsA, particularly axial disease, would affect patients' responses to these instruments.

The Rasch model is one model of the item response theory, which is a statistical theory about item (i.e., a question on a disability scale) and scale performance[14]. The Rasch model assumes the probability of a specified response is modelled as a function of person and item parameters. This relationship is expressed through a formula:

$$\Pr(X_{ij} = k) = \frac{\exp(-\sum_{l=0}^{k}\tau_l + k(\theta_j - b_i))}{\sum_{l=0}^{m}\exp(-\sum_{l=0}^{k}\tau_l + k(\theta_j - b_i))}$$

where Pr (X_{ij} = k) is the probability that the jth subject falls in the kth category for item i. The number of categories is m + 1, θ_j is the jth person's ability, and b_i is the ith item's difficulty. τ_i are step thresholds between category l and l – 1 with $\tau_0 = 0$.

The Rasch model calibrates ability and item difficulty onto a single common metric scale, and deconstructs each item into a series of thresholds. The results are reported in logits. A logit is the natural log of an odds ratio. An increment on the logit scales increases the odds of affirming an event by 2.718 (the base of natural logarithm)[15,16].

The Rasch analysis is a method to obtain objective, fundamental, and linear measures for stochastic observations of ordered category responses. In general, an ideal questionnaire instrument should be statistically reliable, long enough to identify the full range of functional activities, and measure only one dimension (unidimensional). It should be linear such that one unit change in the scale has the same meaning anywhere in the scale. The scale should also work similarly in different patient subgroups, and remain unaffected by subgroups such as sex or age. Over the last 2 decades, Rasch analysis has been adopted as pivotal in judging the quality of existing psychometric outcome instruments and in developing new instruments. It provides disease-specific and comparative quality of life (QOL) measures by "item banking" items (or questions) onto the same underlying metric in other rheumatic diseases[17-19]. It has also been employed in development of the specific QOL measure for PsA[20].

In our study, the 4 functional indexes were compared using Rasch analysis. We evaluated the statistic fit of the item response theory by the Rasch model and thus the sufficiency of the indexes as an estimate of a person's ability. We also tried to evaluate whether items of the BASFI and Dougados-FI display differential item functioning (DIF) or item bias with respect to the axial or peripheral disease subgroups. This may suggest these indexes work differently in axial or peripheral disease subgroups.

MATERIALS AND METHODS

Patients. Consecutive patients with PsA followed in a single rheumatology tertiary referral center were invited for a cross-sectional study using a standardized protocol. All patients were > 18 years old and fulfilled the Classification of Psoriatic Arithitis (CASPAS) criteria for PsA[21].

Assessment. Demographic data were recorded. Clinical features assessed included swollen, tender, and damaged joint counts in 66/68/68 diarthrodial joints, respectively. Severity of joint pain was self-reported on a 0–10 visual analog scale (VAS). Patients' perceptions of health were reported on a 0–5 numerical scale. Functional and quality of life scores recorded included the HAQ, BASFI, Dougados-FI, and the Chinese (Hong Kong) version of the SF-36[22]. For each item of the BASFI, a numerical rating scale from 0–10, ranging from "easy" to "impossible," was incorporated instead of the original VAS[23,24]. Radiographs of hands and wrists and sacroiliac joints were taken and the presence or absence of hand and wrist erosions or sacroiliitis were determined by a rheumatologist and a radiologist blinded to patients' clinical features. The presence of sacroiliitis was defined as grade 2 or above bilaterally or grade 3–4 unilaterally, according to the New York grading system[25]. Patients with radiographic sacroiliitis were classified as the axial subgroup, the rest were classified as the peripheral subgroup.

The study protocol was reviewed and approved by the Joint Chinese University of Hong Kong – New Territories East cluster (CUHK-NTEC) clinical research ethics committees. Before entry to the study, participants were informed of the nature and purpose of study.

Statistical analysis. Demographic and clinical characteristics were noted using descriptive parametric or nonparametric statistics as appropriate. Analyses were performed using the Statistic Package for Social Science (SPSS for Windows, version 10.0 2000; SPSS, Chicago, IL, USA). All hypothesis tests were 2-tailed and p values < 0.05 were considered significant.

Winstep was used for Rasch analysis[26]. Well-fitted statistics indicate that the subscale items contribute to a single underlying construct (unidimensionality). The overall fit to the model was assessed using the item-trait chi-square interaction statistic. Nonsignificant chi-square values indicate model fit[27]. Other tests for unidimensionality included fit statistics, principal component analysis (PCA), and an individual t-test approach.

Fit of the observed data to the Rasch model was assessed by a chi-square statistic, the mean-square (MNSQ) of information-weighted fit statistics (INFIT) and outlier-sensitive statistics (OUTFIT). The acceptable range of fit statistics is in the range of 0.7–1.3[28]. A fit value of 1 + x indicates 100% times x more variation between the observed and modeled predicted data. For example, an INFIT value of 1.3 indicates 30% more variation between the observed and model-predicated value. INFIT takes particular note of the difference between observed and expected response for those items that have a difficulty level near the person's ability level, and thus is a weighted fit statistic that gives greater weight to responses to items close to the person's ability level. OUTFIT includes the differences for all items, irrespective of how far the item difficulty is from the person's ability. Taken together, INFIT and OUTFIT allow one to construct a detailed picture of the working of items within a scale. A high fit statistic, > 1.3, denotes noise in the data and generally implies the item does not belong to the unidimensional construct. A low fit statistic, < 0.7, indicates that the item is "muted" or often has interdependence with another item[29].

PCA was performed to confirm unidimensionality. PCA examines the residuals (i.e., the remaining data variance after extracting the data component modeled by Rasch analysis). A desirable instrument should not produce a second factor structure with PCA, as the Rasch solution should represent the single factor in the data.

Unidimensionality was further confirmed with an individual t-test

approach[29]. This is a comparison of person estimates based on subsets on items and is a robust test of unidimensionality used to avoid significantly different person estimates driven by multidimensionality. The fundamental idea is to conduct a series of independent t-tests with the pairs of person measures fitted from 2 subsets of items. A significant t-test indicates the level of trait differs depending on which items are included in calibration and hence indicating multidimensionality. The items were separated using the item loadings on the first factor of the PCA of the residuals. Person estimates derived from the positive set of items are contrasted against those derived from the negative set. A series of individual t-tests was undertaken to compare the estimates for each person[27]. The percentage of tests that are significant at the 5% level is computed. A binomial proportion confidence interval is then calculated for this percentage. The 95% binomial proportions confidence intervals should cover 5% for nonsignificant violation of unidimensionality.

Person separation reliability and item separation of the 4 indexes were also examined. Person separation reliability illustrates how well the index differentiates persons, while item separation provides a measure on the spread of item difficulties. An index is considered reliable if person separation reliability is > 0.8. Item thresholds for each item subscale of each index were plotted. A good instrument should have a good overall item separation index > 2.0. The greater the separations between item thresholds, the more distinct strata are identified. Individual items that are > 0.15 logits apart are considered as individual strata[30]. A good instrument should also have a long applicable measurement span.

Items of each index were examined for DIF or item bias by comparing item performance for different subgroups using t-tests. Subgroups subjected to analyses included sex and axial or peripheral subgroups. We analyzed the person-response residuals for each item, which mark the extent to which each person diverges from the expected response for their particular ability level. Item bias is suggested if divergence is common to a particular subgroup.

RESULTS

One hundred eight patients with PsA completed the functional indexes. Demographic data for 56 male and 52 female patients are shown in Table 1. As with previous studies, the 4 functional indexes were highly correlated with each other (Spearman's r ranged from 0.76 to 0.81, a negative value for SF-36-PF) and correlated moderately with patients' perception of health status and pain score (Spearman's r ranged from 0.44 to 0.56, negative values for SF-36-PF)[5,9]. The functional indexes were only poorly correlated to tender and swollen joint counts (only data for HAQ was shown in Table 1). All 4 functional indexes had floor effects with the profiles of scores skewed towards the less disabled end (Figure 1). It was found that 18.5%, 31.1%, and 24.5% of patients had zero scores for BASFI, Dougados-FI, and HAQ, respectively. Floor effects appeared less marked in SF-36-PF. Maximum score for SF-36-PF occurred in 7.4%. All instruments exhibited significant difference between 5 levels of patients' perception of health status (all Kruskal-Wallis test: BASFI, chi-squared 28.1, $p = 0.001$; Dougados-FI, chi-squared 24.8, $p < 0.001$; HAQ, chi-squared 27.1, $p < 0.001$; SF-36-PF, chi-squared 26.7, $p < 0.001$).

Under Rasch analysis, characteristics determining optimal psychometric properties for an instrument include unidimensionality; minimal floor or ceiling effect (i.e., having long applicable measurement span); adequate spread of items along the linear measurement scale (item separation); no item bias or DIF; and no category or step disordering.

Fit analysis to the Rasch model (unidimensionality). The basic assumption of the Rasch model is that the items of each index belong to a single construct (unidimensionality). Good fit of the data to the model denotes unidimensionality. Nonsignificant overall item-trait interaction chi-squares were obtained for the 4 functional indexes, suggesting good overall fit. The item-trait interaction chi-squares for BASFI, Dougados-FI, HAQ, and SF-36-PF were, respectively, 27.8 (df 20, $p = 0.11$), 35.7 (df 40, $p = 0.66$), 17.4 (df 16, $p = 0.36$), and 24.3 (df 20, $p = 0.23$).

The person separation reliability for BASFI, Dougados-FI, HAQ, and SF-36-PF were all above the desired range of

Table 1. Demographic data of the psoriatic arthritis cohort.

	Characteristics mean (± SD)	Spearman Correlation HAQ	Patients' Global	Pain Score
Male/female, no.	52/56			
Age, yrs	49.3 (± 12.6)			
Duration of arthritis, yrs	9.00 (± 6.8)	–0.08		
Sacroiliitis (%)	35 (32.4)			
Swollen joint (0–66), no.	4.11 (± 5.42)	0.18		
Tender joint (0–68), no.	1.81 (± 2.72)	0.43*		
Damaged joint (0–66), no.	3.41 (± 4.72)	0.41*		
Pain score (0–100)	46.94 (± 26.46)			
Patients' global (0–5)	45.83 (± 23.88)			
PASI	4.7 (± 6.3)	0.16		
BASFI	24.41 (± 22.93)	0.81*	0.49*	0.52*
Dougados-FI	6.18 (± 7.08)	0.76*	0.44*	0.50*
HAQ	0.69 (± 0.67)	—	0.53*	0.56*
SF-36-PF	63.33 (± 25.50)	–0.80*	–0.44*	–0.499*

* $p < 0.001$. Patients' global: Patients' global assessment of health; PASI: Psoriasis Area and Severity Index; BASFI: Bath Ankylosing Spondylitis Functional Index; Dougados-FI: Dougados Functional Index; HAQ: Health Assessment Questionnaire; SF-36-PF: Medical Outcome Study Short-Form 36, Physical Functioning.

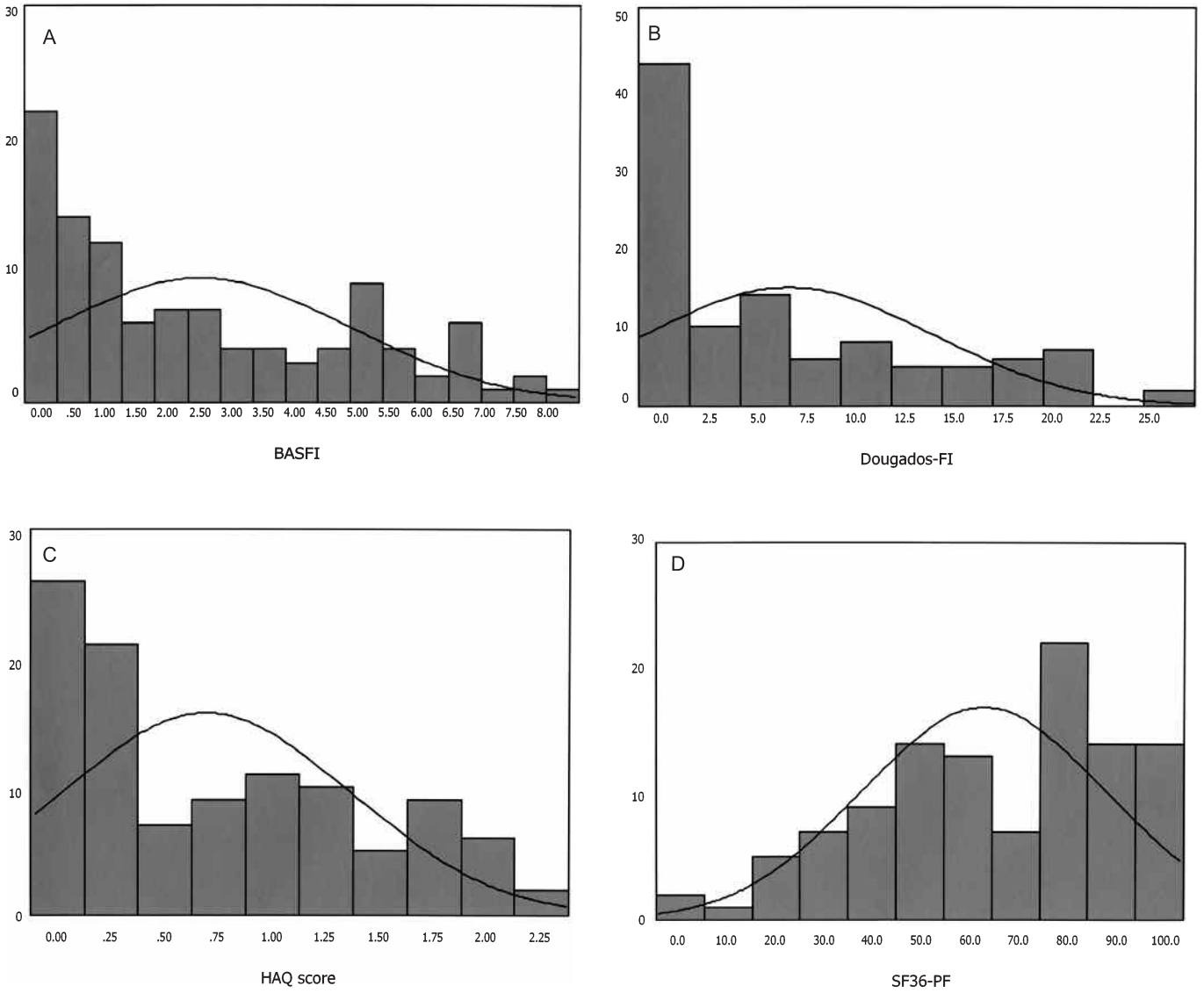

Figure 1. Raw scores and distribution: (A) Bath Ankylosing Spondylitis Functional Index, (B) Dougados Functional Index, (C) Health Assessment Questionnaire, and (D) Medical Outcomes Short-Form 36, physical functioning.

> 0.8. Item separation indexes were all above 2.0 logits (Table 2). For our sample size and cross-sectional study design, INFIT (MNSQ) and OUTFIT (MNSQ) within the range of 0.7–1.3 represent adequate fit to the Rasch model[26]. The INFIT/OUTFIT (MNSQ) statistics and item calibration (in logits) for each functional index are shown in Table 2. Higher item calibration indicates easier items and lower item calibration indicates more difficult items for the PsA cohort for BASFI, Dougados-FI, and HAQ. The reverse is true for the SF-36-PF. Greater separation in logits between item thresholds indicates more distinct strata. In this respect, we observe that the Dougados-FI performed poorly, as a few item difficulties were overlapping. The HAQ also had item calibrations close to each other (inadequate item separation).

Two items of the BASFI (full days activities and bend to floor) had high OUTFIT (MNSQ). A high OUTFIT (MNSQ) indicates unexpected response of the cohort towards these items, which is not appropriate with their given ability. It may indicate these items were not well understood. Although higher score for these items means higher disability, highly disabled patients do not find these particularly difficult. The person separation reliability and item separation of BASFI were 0.83 and 3.33, and were satisfactory. The span of item threshold or measurement span (2.97 logits) was the shortest among the 4 indexes.

The Dougados-FI had good person separation reliability (0.85) and reasonably long measurement span of 7.99 logits. However, many items displayed misfit to the Rasch model. Four items — cough or sneeze, get out of bed, sleep on your

Table 2. Summary of Rasch analysis with items in order of increasing difficulty. Fit statistics 0.7–1.3 denote adequate fit to the Rasch model.

Item	INFIT (MNSQ)	OUTFIT (MNSQ)	Item Calibration in logits (SE)	DIF Sex	DIF SAC
BASFI					
1. Put on socks	1.00	0.94	0.49 (0.07)	NS	NS
6. Stand unsupported	1.30	1.05	0.18 (0.06)	NS	NS
2. Bend to floor	1.01	1.37	0.13 (0.06)	NS	NS
3. Reach to shelf	0.85	0.87	–0.01 (0.06)	NS	NS
4. Up from chair	0.83	0.76	–0.02 (0.06)	NS	NS
8. Look over shoulder	1.03	0.95	–0.09 (0.05)	NS	NS
7. Climb steps	1.03	0.87	–0.13 (0.05)	NS	0.015
10. Full day's activities	1.17	1.31	–0.13 (0.05)	NS	NS
9. Demanding activities	1.09	1.28	–0.17 (0.05)	NS	NS
5. Up from floor	1.02	1.12	–0.25 (0.05)	NS	NS
Dougados-FI					
20. Breathe deep	1.08	0.84	2.19 (0.39)	NS	NS
19. Cough or sneeze	1.07	1.88	1.64 (0.35)	NS	0.045
8. Sit down	0.66	0.42	1.10 (0.31)	NS	NS
15. Get out of bed	0.80	1.92	1.10 (0.31)	NS	NS
13. Lie down	0.92	0.56	0.73 (0.30)	NS	NS
16. Sleep on your back	1.12	0.87	0.65 (0.29)	NS	NS
14. Turn in bed	0.76	0.82	0.56 (0.29)	NS	NS
2. Pull on trousers	0.76	0.59	0.48 (0.29)	NS	NS
4. Get into a bathtub	0.83	0.71	0.40 (0.28)	NS	NS
5. Remain standing 10 min	0.86	0.77	0.37 (0.28)	NS	NS
1. Put on your shoes	0.83	0.90	0.32 (0.28)	NS	NS
3. Pull on pullover	0.98	0.94	0.02 (0.27)	NS	NS
10. Get into a car	0.78	0.65	–0.47 (0.26)	NS	NS
9. Get up from a chair	0.88	0.77	–0.54 (0.26)	NS	NS
17. Sleep on your stomach	1.47	1.17	–0.73 (0.25)	NS	NS
18. Do your job or housework	0.80	0.84	–0.79 (0.25)	NS	NS
6. Climb 1 flight of stairs	1.49	1.64	–1.15 (0.24)	Ns	NS
11. bend over to pick up an object	1.26	1.25	–1.33 (0.24)	NS	NS
12. Crouch	1.08	1.05	–2.20 (0.23)	NS	NS
7. Run	1.19	1.20	–2.36 (0.23)	0.001	NS
HAQ					
5. Hygiene	1.05	0.90	0.56 (0.20)	NS	NS
3. Eating	0.95	0.82	0.32 (0.20)	NS	NS
1. Dressing & grooming	1.09	1.16	0.29 (0.19)	NS	NS
4. Walking	0.87	0.73	0.25 (0.19)	NS	NS
8. Activities	0.80	0.88	0.14 (0.19)	NS	NS
2. Arising	0.89	0.91	0.03 (0.19)	NS	NS
7. Grip	1.41	1.40	–0.73 (0.18)	0.001	NS
6. Reach	1.02	1.04	–0.86 (0.18)	NS	NS
SF-36-PF					
3j. Bathing or dressing yourself	1.16	0.85	–2.10 (0.26)	NS	NS
3i. Walking one block	1.07	0.86	–1.84 (0.25)	NS	NS
3e. Climbing one flight of stair	0.79	0.57	–1.55 (0.24)	NS	NS
3c. Lifting or carrying groceries	1.11	1.09	–0.61 (0.22)	NS	NS
3h. Walking several blocks	0.77	0.62	–0.52 (0.22)	NS	NS
3d. Climbing several flight of stairs	0.88	0.86	–0.02 (0.21)	NS	NS
3f. Bed making, kneeling, or stooping	1.08	1.08	0.50 (0.21)	NS	NS
3g. Walking more than a mile	0.80	0.79	1.25 (0.20)	NS	NS
3b. Moderate activities	1.37	1.30	1.29 (0.20)	NS	NS
3a. Vigorous activities	1.02	1.27	3.59 (0.24)	NS	NS

BASFI: Item separation index 3.33; person separation reliability 0.83; span of item threshold 2.97.
Dougados-FI: Item separation index 3.83; person separation reliability 0.85; span of item threshold 7.99.
HAQ: Item separation index 2.22; person separation reliability 0.84; span of item threshold 5.63.
SF-36-PF: Item separation index 6.99; person separation reliability 0.85; span of item threshold 9.03.

INFIT: information-weighted fit statistics; OUTFIT: outlier-sensitive statistics; MNSQ: mean square; DIF: differential item functioning; SAC: sacroiliitis.

stomach, and climb one flight of stairs — had high fit statistics. Another 4 items — sit down, lie down, pull on trousers, and get into a car — had low fit statistics, indicating redundancy and in-built dependency of these items with each other.

As for the HAQ, one item, grip, had high fit statistics. The HAQ was also limited by a short item separation of 2.22, which indicated relatively poor spread of item difficulties. The measurement span was relatively short (5.63 logits).

One item of the SF-36-PF had high fit statistics. Two items, walking several blocks and climbing one flight of stairs, had low fit statistics that indicated high in-built redundancy. The SF-36-PF may be improved by removing one of these items. Among the 4 indexes, the SF-36-PF had the best item separation, 6.99, and the longest measurement span, 9.03 logits.

Unidimensionality was confirmed with PCA. There was no evidence of a second factor in all indexes when the "Rasch factor" was removed. The observed variances explained by the Rasch model for BASFI, Dougados-FI, HAQ, and SF-36-PF were 78%, 76%, 68%, and 89%, respectively. The percentages of residual variance accounted for by the first contrast were 18%, 13%, 22%, and 21%, respectively, which were within acceptable limits.

Individual t-tests further confirmed unidimensionality, except for the Dougados-FI. There were 3.7%, 9.26%, 8.33%, and 5.66% of individual t-tests that were significant at the 5% level for BASFI, Dougados-FI, SF-36-PF, and HAQ, respectively. The 95% binomial proportions confidence interval of the BASFI, SF-36-PF, and HAQ covered 5%. This indicated no violation of unidimensionality. The Dougados-FI illustrated a mild issue of multidimensionality according to the result of this test.

Differential item functioning. One hundred four radiographs of sacroiliac joints were available for DIF analysis. Thirty-five patients (33.7%) had sacroiliitis according to the New York grading system. Items in each index that displayed DIF are shown in Table 2. Items displaying DIF are biased by different subgroups, meaning that at a level of disability, factors other than disability alone are determining patients' responses. As the BASFI was developed for patients with AS, it may work differently in PsA patients with or without spinal involvement. However, only one item in the BASFI, climb steps, displayed DIF in subgroups with or without sacroiliitis. This means that PsA patients with or without sacroiliitis responded similarly to the BASFI.

Distribution of item threshold, category, and step disordering. The distribution of item thresholds on a common underlying scale derived from the Rasch model for the 4 indexes is shown in Figure 2. For each item of each index, the transitions from category 0 to 1, 1 to 2, 2 to 3 and so on are expressed as probability thresholds on an underlying metric scale. The item threshold for category 0 to 1, for example, marks the disability level at which a response of 1 becomes more probable than a response of 0. This made 100 thresholds (10 thresholds for each of 10 items) for BASFI and 60, 24, and 20 thresholds for Dougados-FI, HAQ, and SF-36-PF, respectively.

The span of item threshold was the smallest for BASFI, although it had the largest number of thresholds. Therefore, the "towers" of thresholds were prominent. "Towers" means that several thresholds are marking the same point on the underlying disability construct, and it is easier to gain points in one area of the scale than the other. This means the BASFI was narrow in domains in the assessment of disability. Dougados-FI and HAQ had less evenly distributed thresholds and existence of "towers." On the other hand, the SF-36-PF displayed fewer "towers" and maintained a relatively even distribution over its span, except for the larger "space" observed in the upper end. "Spaces" were observed in all 4 indexes. These "spaces" reflect that scale precision might be compromised in a certain area. A recent study has also illustrated that many concepts of the International Classification of Functioning, Disability and Health (ICF) are not adequately covered by existing instruments in a PsA cohort[31]. If the item thresholds of an index follow a logarithmic pattern, the index would be functioning as an interval measure. However, none of the indexes could serve as interval measures. The item thresholds in SF-36-PF were the most evenly distributed among the 4 indexes, with fewer "towers" and "spaces," which might confer a certain superiority.

Whether the responses to the items are consistent with the metric estimate of the underlying construct is indicated by an ordered set of response thresholds for each of the items. No category or step disordering was found in Dougados-FI, HAQ, and SF-36-PF. Some category disordering and noticeable step disordering were observed in BASFI (Table 3). Normally, we expect the average person measure to progress in an order from category 0 to 10. However, the average measure of category 9 failed to follow this order (category disordering). The reason was that only 1% of the observed data fell in this category. On the other hand, we expect the item thresholds to progress in an order. Disordered item thresholds (or step disordering) imply that there is no ability level at which a particular category becomes the most likely response. The BASFI had marked step disordering in every item, indicating the presence of excessive category levels. Combining categories may be beneficial in improving the performance of the BASFI. No significant step disordering was observed in other indexes.

DISCUSSION

Modern psychometric theory makes further demands of a health status instrument. Rasch analysis has become pivotal in evaluation and production of instruments. We compared the use of the BASFI, Dougados-FI, HAQ, and SF-36-PF in

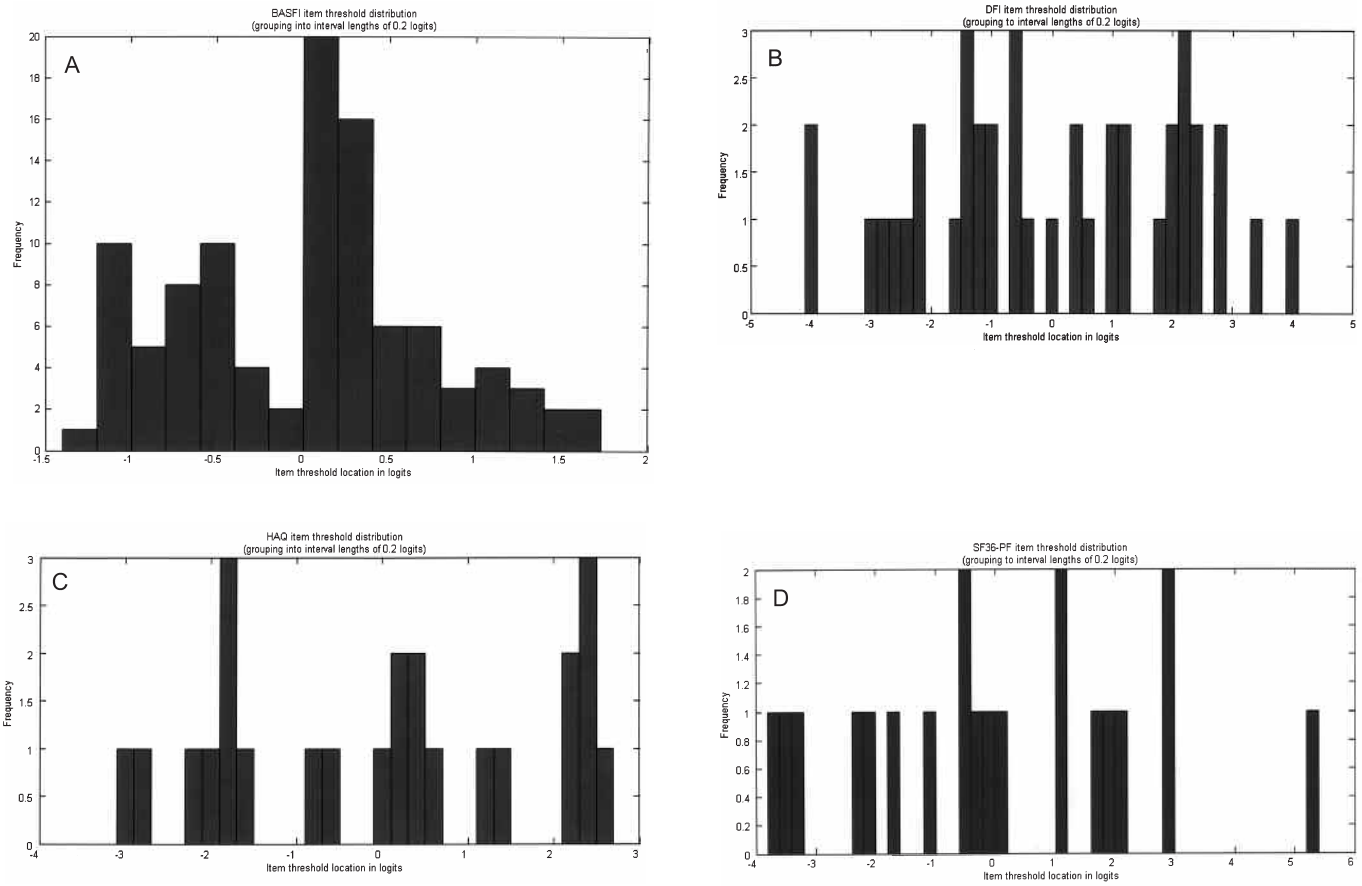

Figure 2. Item threshold distributions: (A) Bath Ankylosing Spondylitis Functional Index, (B) Dougados Functional Index, (C) Health Assessment Questionnaire, and (D) Medical Outcomes Short-Form 36, physical functioning.

Table 3. Category and step order for Bath Ankylosing Spondylitis Functional Index. Item threshold did not follow the numerical order, and this illustrated marked step disordering.

Category Label	Average Person Measure	Item Threshold (relative to item difficulty)
0	–1.12	None
1	–0.74	0.18
2	–0.67	–0.90
3	–0.38	–0.44
4	–0.23	0.21
5	–0.05	–0.47
6	0.10	0.43
7	0.12	0.26
8	0.22	0.48
9	0.17	1.24
10	0.32	–0.99

a cohort of patients with PsA and confirmed that each of the 4 indexes measures a single construct of physical disability. All indexes had floor effects and could not be used as interval measures. The HAQ and SF-36-PF had fewer misfit items. SF-36-PF was superior in terms of having less floor effect, good item separation, long span of item threshold, better distributional properties, and less DIF. HAQ was limited by its floor effect, inadequate item separation, and relatively short measurement span. A recent study also demonstrated similar superior properties of the SF-36-PF in a PsA cohort[32].

The BASFI and Dougados-FI behaved similarly in patient subgroups with or without sacroiliitis. There is no superiority in using these indexes in the axial subgroup of PsA. In AS, significant step disordering was observed with the original BASFI using a VAS of 0–100[33]. We hoped that step disordering could be solved with the numerical rating scales from 0–10 in this study. However, marked step disordering with this numerical rating BASFI was still demonstrated. This indicates that the BASFI has far too many categories that confuse patients with PsA. Patients choosing category 1 or 2 may not have much difference in the level of disability. BASFI may be improved by collapsing categories. DFI, on the other hand, had many redundant items and may be improved by repeating Rasch analysis in a larger cohort and reducing redundant items one by one.

Good fit statistics were exhibited by the HAQ and SF-36-

PF. There were 2 items in SF-36-PF that had low fit statistics or in-built interdependence. These tasks may involve similar ability in this patient cohort. We preliminarily removed item 3e, "climbing one flight of stairs." The remaining 9 items were subjected to Rasch analysis and resulted in better fit statistics (data not shown). We propose that this item be removed in further studies using the SF-36-PF in PsA.

There are several limitations in our report. First, the sample size was relatively small, in particular, that for DIF measurement. However, unidimensionality in BASFI, HAQ, and SF-36-PF was observed. The DIF analyses were justified, with good fit statistics, the absence of a second factor in PCA analysis, and individual t-test analyses. There was a slight concern with multidimensionality in the Dougados-FI with the individual t-test approach, which is a more robust form of unidimensionality test. The overall fitness by the other tests, however, was largely fulfilled.

The second limitation was the cross-sectional study design. This does not allow test-retest reliability evaluation and does not provide information on the sensitivity to change after treatment. Larger multicenter studies over longer periods of time or after interventions are warranted to address these issues.

In conclusion, the BASFI and Dougados-FI confer no superiority in patients in the subgroup with axial PsA. The SF-36-PF is the best instrument for measuring functional disability in PsA, in terms of least floor effect, good item separation, long measurement span, less DIF, and better distributional properties.

ACKNOWLEDGMENT

We gratefully acknowledge Dr. Joyce Hui Wai Yee for reading the radiographs; and Tena Li and Lorraine Tseung for assistance in data collections.

REFERENCES

1. Husted JA, Gladman DD, Farewell VT, Cook RJ. Health-related quality of life of patients with psoriatic arthritis: a comparison with patients with rheumatoid arthritis. Arthritis Rheum 2001;45:151-8.
2. Husted JA, Tom BD, Farewell VT, Schentag CT, Gladman DD. Description and prediction of physical functional disability in psoriatic arthritis: a longitudinal analysis using a Markov model approach. Arthritis Rheum 2005;53:404-9.
3. Husted JA, Tom BD, Farewell VT, Schentag CT, Gladman DD. A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: Does the effect change over time? Arthritis Rheum 2007;56:840-9.
4. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcomes in arthritis. Arthritis Rheum 1980;23:137-45.
5. Blackmore MG, Gladman DD, Husted J, Long JA, Farewell VT. Measuring health status in psoriatic arthritis: the Health Assessment Questionnaire and its modifiction. J Rheumatol 1995;22:886-93.
6. Husted JA, Gladman DD, Cook RJ, Farewell VT. Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. J Rheumatol 1998;25:2146-55.
7. Antoni CE, Kavanaugh A, Kirkham B, et al. Sustained benefits of infliximab therapy for dermatologic and articular manifestations of psoriatic arthritis: results from the Infliximab Multinational Psoriatic Arthritis Controlled Trial (IMPACT). Arthritis Rheum 2005;52:1227-36.
8. Mease PJ, Gladman DD, Ritchlin CT, et al. Adalimumab for the treatment of patients with moderately to severely active psoriatic arthritis: results of a double-blind, randomized, placebo-controlled trial. Arthritis Rheum 2005;52:3279-89.
9. Daltroy LH, Larson MG, Roberts NW, Liang MH. A modification of the Health Assessment Questionnaire for the spondyloarthropathies. J Rheumatol 1990;17:946-50.
10. Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire II: a revised version of the Health Assessment Questionnaire. Arthritis Rheum 2004;50:3296-305.
11. Cardiel MH, Londono JD, Gutierrez E, Pacheco-Tena C, Vazquez-Mellado J, Burgos-Vargas R. Translation, cross-cultural adaptation, and validation of the Bath Ankylosing Spondylitis Functional Index (BASFI), the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) and the Dougados Functional Index (DFI) in a Spanish speaking population with spondyloarthropathies. Clin Exp Rheumatol 2003;21:451-8.
12. Husted JA, Gladman DD, Farewell VT, Long JA, Cook RJ. Validating the SF-36 health survey questionnaire in patients with psoriatic arthritis. J Rheumatol 1997;24:511-7.
13. Heikkila S, Viitanen JV, Kautianen H, Kauppi M. Evaluation of the Finnish versions of the functional indices BASFI and DFI in spondyloarthropathy. Clin Rheumatol 2000;19:464-9.
14. Hambleton R, Jones R. Comparison of classical test theory and item response theory and their applications to test development. Educ Meas Issues Practice 1993;12:38-47.
15. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1980.
16. Andrich D. Rasch models for measurement. Quantitative applications in the social sciences. No. 68. London: Sage Publications; 1988.
17. Doward LC, Spoorenberg A, Cook SA, et al. Development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. Ann Rheum Dis 2003;62:20-6.
18. Leong KP, Kong KO, Thong BYH, et al. Development and preliminary validation of a systemic lupus erythematosus-specific quality-of-life instrument (SLEQOL). Rheumatology Oxford 2005;44:1267-76.
19. Wolfe F, Kong SX. Rasch analysis of the Western Ontario McMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. Ann Rheum Dis 1999;58:563-8.
20. McKenna SP, Doward LC, Whalley D, Tennant A, Emery P, Veale DJ. Development of the PsAQoL: a quality of life instrument specific to psoriatic arthritis. Ann Rheum Dis 2004;63:162-9.
21. Taylor W, Gladman D, Helliwell P, Marchesoni A, Mease P Mielants H, CASPAR Study Group. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. Arthritis Rheum 2006;54:2665-73.
22. Lam CL, Gandek B, Ren XS, Chan MS. Tests of scaling assumptions and construct validity of the Chinese (HK) version of the SF-36 Health Survey. J Clin Epidemiol 1998;51:1139-47.
23. Calin A, Garrett S, Whitelock H, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994;21:2281-5.
24. Van Tubergen A, Debats I, Rvser L, et al. Use of a numerical rating scale as an answer modality in ankylosing spondylitis-specific questionnaires. Arthritis Rheum 2002;47:242-8.
25. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. Arthritis Rheum 1984;27:361-8.

26. Linacre JM, Wright BD. A user's guide to Winsteps. Chicago: Mesa Press; 1997.
27. Tennant A, Pallant JF. Unidimensionality matters! (A tale of two Smiths?). Rasch Measurement Transactions 2006;20:1048-51.
28. Silverstein F, Kilgore DJ, Harman JG, Harvey T. Applying psychometric criteria to functional assessment in medical rehabilitation, 1: exploring unidimensionality. Arch Phys Med Rehabil 1991;72:631-7.
29. Smith EV. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. J Appl Meas 2002;3:205-31.
30. Wright BD. Reasonable mean-square fit values. In: Wright BD, Linacre JM, editors. Rasch measurement transactions. Part 2. Chicago: Mesa Press; 1996:370.
31. Stamm TA, Nell V, Mathis M, et al. Concepts important to patients with psoriatic arthritis are not adequately covered by standard measures of functioning. Arthritis Rheum 2007;57:487-94.
32. Taylor WJ, McPherson KM. Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. Arthritis Rheum 2007;57:723-9.
33. Eyres S, Tennant A, Kay L, Waxman R, Helliwell PS. Measuring disability in ankylosing spondylitis: comparison of Bath Ankylosing Spondylitis Functional Index with revised Leeds Disability Questionnaire. J Rheumatol 2002;29:979-86.