Measuring Disability in Ankylosing Spondylitis: Comparison of Bath Ankylosing Spondylitis Functional Index with Revised Leeds Disability Questionnaire

SOPHIE EYRES, ALAN TENNANT, LESLEY KAY, ROBIN WAXMAN, and PHILIP S. HELLIWELL

ABSTRACT. ***Objective.*** Disability has been identified as a core outcome measure in ankylosing spondylitis (AS). The Dougados Functional Index (DFI) and the Bath Ankylosing Spondylitis Functional Index (BASFI) have been selected as core measures of function in this disease. However, neither of these instruments has undergone rigorous psychometric testing.

Methods. The psychometric properties of 2 measures of disability, the BASFI and the revised Leeds Disability Questionnaire (RLDQ), were compared in a cohort of 208 outpatients with AS. Rasch analysis was used to examine the properties of each measure and to compare them on a common scale. Test-retest was assessed in a cohort of 149 subjects who completed each instrument twice over an interval of 2 weeks.

Results. Both instruments gave an even spread of scores across the study group, but BASFI responses were positively skewed and RLDQ responses negatively skewed. There was a highly significant difference between perceived severity groups for both instruments (Kruskal-Wallis chi-squared: RLDQ, 75.1; BASFI, 80.4; both $p < 0.0001$). Both instruments gave acceptable test-retest scores (RLDQ ICC = 0.95, 95% CI 0.93–0.97; BASFI ICC = 0.94, 95% CI 0.92–0.96). Both instruments were found to be unidimensional according to the Rasch model, but the BASFI had more items displaying differential item functioning. Category disordering was apparent with the BASFI but not the RLDQ. However, both instruments displayed disordered item thresholds. Neither instrument can be used as an interval measure. Both measures had "towers" of thresholds whereby several thresholds were marking the same point on the underlying disability construct. This was particularly notable in the case of the BASFI.

Conclusion. Both the BASFI and RLDQ provide a unidimensional measure of function in AS that is in accord with patient perception of disease severity. Neither instrument can be used as an interval measure. Changing the way that the instruments are scored, for example by collapsing categories, may improve their performance. (J Rheumatol 2002;29:979–86)

Key Indexing Terms:
ANKYLOSING SPONDYLITIS
DISABILITY
FUNCTIONAL ASSESSMENT
RASCH ANALYSIS

Ankylosing spondylitis (AS) is a chronic inflammatory rheumatic condition affecting primarily the spinal column and the large peripheral joints, resulting in stiffness and deformity of the affected skeleton. The pattern and rate of disease progression are variable, but deterioration often proceeds independently of disease duration[1]. Although major advances have occurred in recent years in the understanding of the disease pathogenesis, the precise natural history and optimal strategy for treatment are still unknown. Disease onset is generally in late adolescence or early adulthood, and consequently the effects are present for a majority of the patient's life.

From the Rheumatology and Rehabilitation Research Unit, University of Leeds, Leeds; and Freeman Hospital, Newcastle-upon-Tyne, England.

S. Eyres, BSc, Research Fellow; A. Tennant, PhD, Professor of Rehabilitation Studies; R. Waxman, MPH, Research Fellow; P.S. Helliwell, MD, PhD, Senior Lecturer in Rheumatology, Rheumatology and Rehabilitation Research Unit, University of Leeds; L. Kay, MD, Consultant Rheumatologist, Freeman Hospital.

Address reprint requests to Dr. P. Helliwell, Rheumatology and Rehabilitation Research Unit, University of Leeds, 36 Clarendon Road, Leeds, UK LS2 SNZ. E-mail: p.helliwell@leeds.ac.uk

Submitted December 28, 2000; revision accepted November 25, 2001.

Instruments currently available for AS focus on symptoms (impairment) and function (disability) and are used to assess outcome in these terms. Recommended core measures for disability include the Bath Ankylosing Spondylitis Functional Index (BASFI) and the Dougados Functional Index (DFI)[2].

In this study the BASFI is compared to a previously published but infrequently used measure, the Leeds Disability Questionnaire, using a technique called Rasch modeling, after the Danish mathematician Georg Rasch[3,4]. The Rasch model is one of many within item response theory. Item Response Theory (IRT) is a general statistical theory about item (question or task) and scale performance and how that performance relates to the attribute measured by the items in the scale[5]. The Rasch model assumes that the probability of a given respondent affirming a particular item is a logistic function of the relative distance between the item location parameter (or difficulty) and the respondent location parameter (or ability)[6]. Namely, the probability that a person will mark any particular category on a disability item ($p_i(\theta)$), for example, is dependent only upon the difference between, first, the person's level of disability (θ) and,

second, the level of disability expressed by that particular item category (b). This relationship is expressed through a formula:

$$p_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

Data can be fitted to the Rasch model by using one of a number of Rasch computer programs (e.g., for one parameter models Winsteps[7] or RUMM[8]). The analysis calibrates ability and item difficulty onto a single common metric scale. Where the item is scored as, for example, 0–1–2–3, or in the case of the visual analog scale, 0–100, the analysis deconstructs each item into a series of thresholds (i.e., the threshold denotes the transition from 0 to 1, 1 to 2, 2 to 3, and so on). Results of the Rasch transformation are reported in logits, which represent the distance along the line of the variable that increases the odds of observing the event (i.e., passing the threshold) by a factor of 2.718. The application of the Rasch model ensures that the fundamental scaling properties of the instrument (for example, unidimensionality and level of measurement) are assessed in addition to the traditional psychometric assessments of reliability and construct validity. When data do fit the model and, uniquely within IRT, the Rasch model confirms the sufficiency of the raw score as an estimator of person ability and provides a transformation to interval level measurement.

MATERIALS AND METHODS

All patients on a disease morbidity register were included after appropriate local ethics committee approval. All participating patients fulfilled the modified New York criteria for AS and made up the complete cohort of a morbidity register at the Freeman Hospital, Newcastle-upon-Tyne. Patients were sent a package consisting of a letter of introduction, a demographic questionnaire, the BASFI[9] and the RLDQ[10], and a reply-paid envelope. Patients who completed and returned the first pack were sent a similar package timed to arrive 2 weeks later. The demographic questionnaire included a question on patient perceived severity of illness. To facilitate interpretation of the results, the BASFI and RLDQ are given as an Appendix.

Rasch analysis. The Rasch computer program Winsteps was used in this study[7]. The fit of the data to the model is expressed in 2 ways. First, the mean-square information-weighted statistic (INFIT) provides information about responses given to items around the same difficulty level as the person's ability. Second, the outlier-sensitive statistic (OUTFIT) refers to items whose difficulty level is remote 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. It is usual to see an INFIT/OUTFIT range of 0.7–1.3 to denote adequate fit of the data to the model[11]. However, the magnitude of the fit statistics is affected by sample size and, in the case of the unweighted fit statistic (OUTFIT), by the number of items being summated. To have a consistent Type I error rate of approximately 0.05, a critical value for the upper limit of OUTFIT would be 1.3 with 150 persons, 1.2 with 500 persons, and 1.1 with 1000 person samples[12].

A poor item fit statistic can indicate poorly constructed or understood items or, when a scale score is assigned by a professional (as with, for example, many outcome measures used in the rehabilitation process), lack of reliability in assignment. Otherwise, poor fit may indicate problems with unidimensionality, that is, the item does not "belong" to the construct or attribute being measured.

Items may also display evidence of differential item functioning (DIF), or item bias. Items can be examined for DIF by comparing item performance for different subgroups using t tests. Specifically, this is achieved by analyzing person-response residuals for each item, which mark the extent to which each person diverges from the expected response for their particular ability level. Where divergence is common to a particular subgroup (for example, males may diverge more than females), item bias is suggested. A scale should work in the same way, irrespective of which group is assessed. Thus, the hierarchical ordering of items along the measurement construct should remain the same for males and females, for young or old, for different clinical groups (cross-diagnostic validity), and across culture (cross-cultural validity). Failure to do so would imply that the scale works differently across such groups, and that the data are not directly comparable.

The operation of categories within each item can also be investigated. In the first instance, we would want the location (or difficulty level) marked by each category to progress in the order intended, from lowest to highest or vice versa. This can be investigated through reference to the "average measure" (or average "ability") of patients in each category. Second, an analysis of the item thresholds marking transition between each category is necessary. Disordering in these would indicate that a particular category is never the most probable and that collapsing the number of categories may be beneficial.

Finally, by plotting the item thresholds for each measure, it is possible to determine the width of the construct covered (in log-odds units) by each measure, and the manner in which the thresholds mark that construct. Specifically, if a measure/scale were functioning at the interval level, its item thresholds would be distributed symmetrically from the center of the scale. As in a logarithmic function, spaces between item thresholds should increase by a common factor, as the distance from the scale center increases.

RESULTS

Questionnaires were sent to 288 people. Respondents to the first questionnaire numbered 208 (149 men, 59 women) and to the second 157 (109 men, 48 women). Demographic data for the initial respondents were as follows: median age 46 years (range 19–81), median duration of disease 18 years (range 1–62), median age at diagnosis 30 years (range 5–70).

The RLDQ was scored by 2 methods. The original method (RLDQ 4) required the instrument to be scored similarly to the Health Assessment Questionnaire: within each of the 4 domains the maximum score was recorded; these scores were summed and the total score was divided by the number of domains answered, thus giving a range of scores from 0 to 3[10]. The alternative scoring method (RLDQ 16) assigns a score of 0–3 for each item and simply sums all item scores, giving a score range of 0–48. Although the former scoring method allows a simple solution to the question of missing data, Rasch analysis indicated that the latter scoring method provided better psychometric properties for the instrument (see below).

The profile of scores for both the RLDQ and the BASFI are given in Figure 1. Both instruments gave an even spread of scores across the study group. BASFI responses were skewed toward the more disabled end of the score range, while responses to the RLDQ were skewed toward the less disabled end.

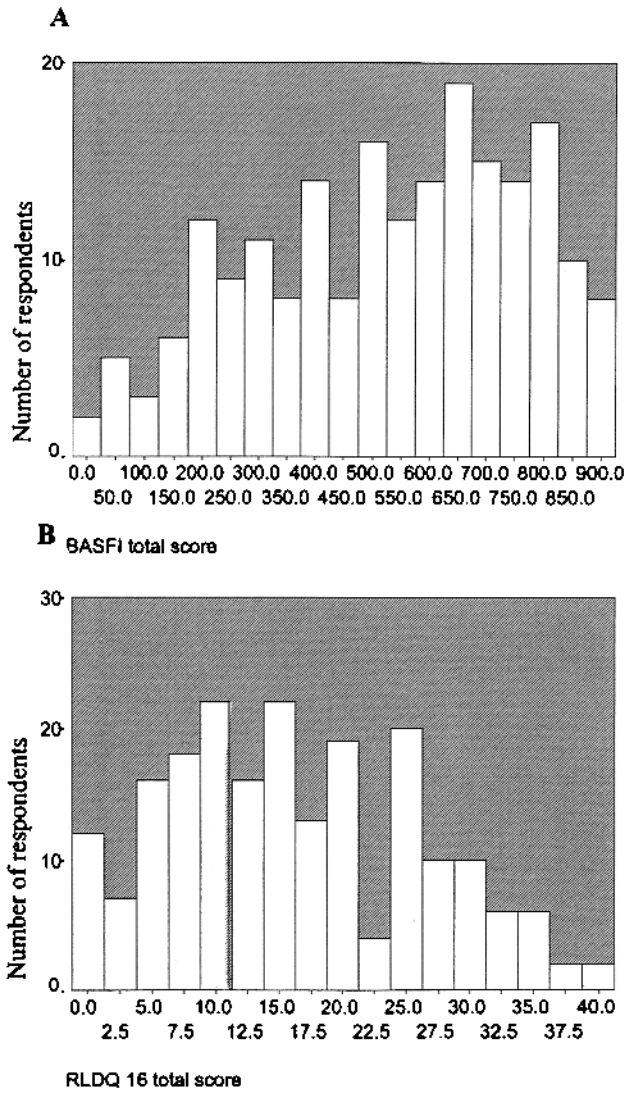


Figure 1. Histogram of scores for BASFI (A) and RLDQ 16 (B).

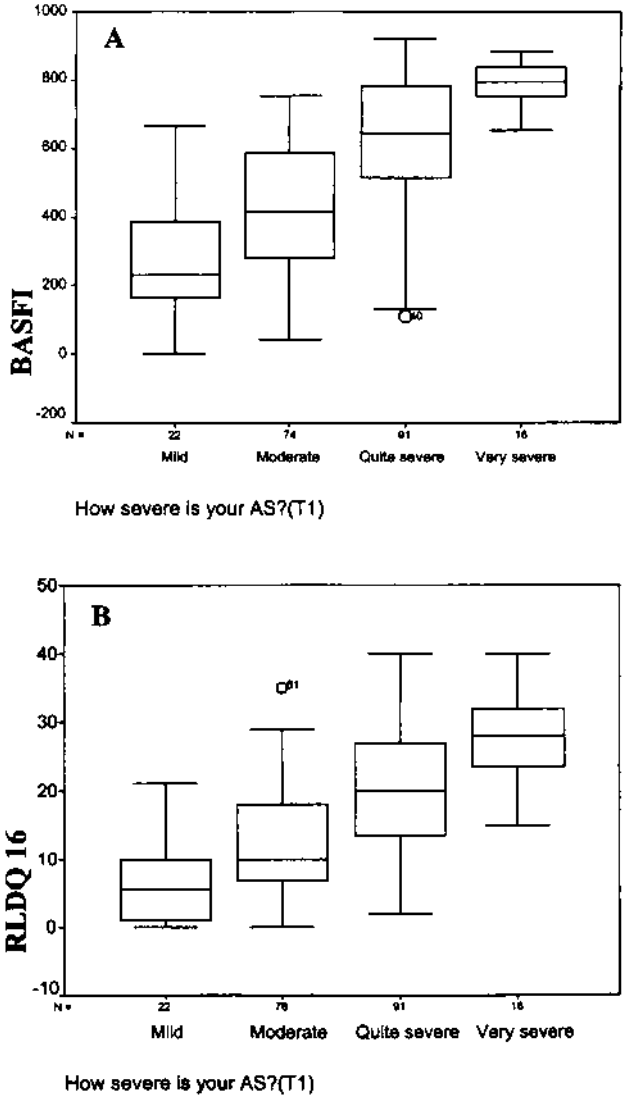


Figure 2. Boxplot (median, range and interquartile range) of BASFI (A) and RLDQ 16 (B) scores with patient's own perception of disease severity.

Respondents were asked to rate the severity of their AS on a 4 point Likert scale (mild, moderate, quite severe, very severe). Scores on both the RLDQ and the BASFI are given in Figure 2. For both instruments there was a highly significant difference between severity subgroups (Kruskal-Wallis: RLDQ, chi-squared 75.1; BASFI, chi-squared 80.4; both p values < 0.0001).

Paired data were available for reliability analysis in 149 respondents. Both instruments gave acceptable test-retest reliability with the following scores for the intraclass correlation coefficient: RLDQ ICC = 0.95, 95% CI 0.93–0.97; BASFI ICC = 0.94, 95% CI 0.92–0.96.

Rasch analysis. A basic assumption underlying the Rasch model is that items belong to a single underlying construct (unidimensionality). It is first necessary to test whether these assumptions have been met, by examining the extent to which responses to each instrument fit the Rasch model (Table 1). This is denoted by 2 fit statistics, INFIT and OUTFIT. For the number of cases in this study, INFIT and OUTFIT values within the range 0.7 to 1.3 represent adequate fit to the Rasch model[12].

Three RLDQ 16 items, "posture: coughing and sneezing," "posture: sleep on stomach," and "posture: sleep on back," were above the required fit range for OUTFIT. Specifically, a high OUTFIT on a frequently affirmed item (such as "posture: sleep on stomach," for example) implies that although this item indicates lower (or initial) levels of disability, highly disabled people do not find this item particularly difficult (or vice versa). Two items, "bending down: taking socks on/off" and "neck: open window," were below the required fit range, in that responses to this item are too predictable from the overall pattern of responses. These items may be redundant or (at worst) have an inbuilt dependency on the others.

Table 1. Summary of Rasch analysis (items in order of increasing difficulty). Fit statistics within the range 0.7–1.3 denote adequate fit to the model.

Item	INFIT, Mnsq	OUTFIT, Mnsq	Item Calibration, Logits
RLDQ 16			
2A Bending toilet	0.99	1	1.03
1C Mobility getting up	1.01	1.05	0.74
4B Posture cough/sneeze	**1.6**	**1.78**	**0.73**
1B Mobility car	0.83	0.81	0.65
3A Neck open window	*0.79*	*0.66*	*0.35*
3D Neck drink from glass	0.98	0.86	0.33
3B Neck cross road	0.88	0.84	0.18
1D Mobility rolling bed	0.9	0.89	0.16
2B Bending socks	*0.69*	*0.67*	*–0.01*
1A Mobility bath	0.79	0.75	–0.02
2C Bending laces/shoes	0.72	0.7	–0.17
4A Posture walk on heels	0.98	0.9	–0.2
3C Neck reach shelf	1.02	1.08	–0.47
4C Posture sleep on back	**1.62**	**1.76**	**–0.58**
2D Bending cut nails	0.81	0.81	–0.8
4D Posture sleep stomach	**1.56**	**1.78**	**–1.92**
RLDQ 4			
1. Mobility	1.03	1.03	0.80
2. Bending	0.96	0.96	0.54
3. Neck	0.97	0.98	0.31
4. Posture	1.05	0.92	–1.66
BASFI			
1 Put on socks	1.15	1.15	0.06
7 Climb steps	0.8	1.01	0.04
3 Reach to shelf	0.8	0.85	0.03
4 Up from chair	*0.69*	*0.66*	*0.01*
2 Bend to floor	1.04	0.94	0
6 Stand unsupported	**1.36**	**1.56**	**–0.01**
10 Full days activities	0.8	0.98	–0.02
5 Up from floor	0.86	0.83	–0.03
8 Look over shoulder	**1.55**	**1.94**	**–0.03**
9 Demanding activities	0.98	1.02	–0.04

For the RLDQ 4, all 4 "items" displayed adequate fit to the Rasch model.

In the case of the BASFI, 2 items ("look over shoulder" and "stand unsupported") were above the required fit range for OUTFIT and one item ("up from chair") was below the required fit range.

Comparison of RLDQ scoring methods. Figure 3 depicts the distribution of item thresholds (derived from Rasch analysis) for each version of the RLDQ as a whole, on a common underlying scale. For each item of each measure, the transition from categories 0 to 1, 1 to 2, and 2 to 3 is expressed as a probability threshold on an underlying metric scale, giving 48 thresholds for the RLDQ 16 (3 thresholds for each of 16 items) and 12 thresholds for the RLDQ 4 (3 thresholds for each of 4 subscales). Thus, the item threshold for categories 0–1, for example, will mark the disability level at which a response of 1 becomes more probable than a response of 0 (Figure 3).

A number of psychometric features can be noted in Figure 3. First, in terms of measurement span, the RLDQ 16 has a slightly wider scale width (7.03 logits) than the RLDQ 4 (6.36 logits). Second, both RLDQ versions have "towers" of thresholds, indicating that several thresholds are marking the same point on the underlying disability construct. This could lead to spurious levels of responsiveness, whereby it is easier to gain points in one area of the scale as opposed to another. Finally, for the RLDQ 4 there are a number of spaces on the underlying construct, meaning that scale precision may be compromised in certain areas. As a consequence, it was decided that the RLDQ 16 would be used for further comparisons with the BASFI.

Comparison of RLDQ and BASFI. Overall fit to the model for the BASFI and RLDQ has been noted previously (Table 1). Each measure was further examined for evidence of differential item functioning, disordered categories, range, and level of measurement.

Differential item functioning (DIF). Table 2 details BASFI and RLDQ items displaying DIF, or item bias by different subgroups. This means that at given levels of function (or disability), factors other than function (or disability) alone are determining subject responses. This is illustrated for the RLDQ with reference to different age groups (age has been

Table 2. Items displaying differential item functioning (DIF). Both variables (age and perceived duration) were split into 2 categories at the median score.

Subgroup	Items Displaying DIF	
	RLDQ	BASFI
Age	(3b) Neck: crossing road	(6) Stand unsupported (8) Look over shoulder
Perceived duration	(3c) Neck: reaching shelf	(3) Reach to shelf (6) Stand unsupported (8) Look over shoulder (9) Physically demanding activities

Table 3. Category and step order for RLDQ items 1a and 4a. Low scores indicate low disability (high ability).

Item	Average Patient Measure	Item-Threshold
Category		
1a Mobility: bath		
0	–2.45	None
1	–0.82	–1.81
2	0.32	0.48
3 (unable to do)	0.68	1.28
4a Posture: walk on heels		
0	–2.12	None
1	–0.67	–1.26
2	–0.15	2.56
3 (unable to do)	0.55	–1.89

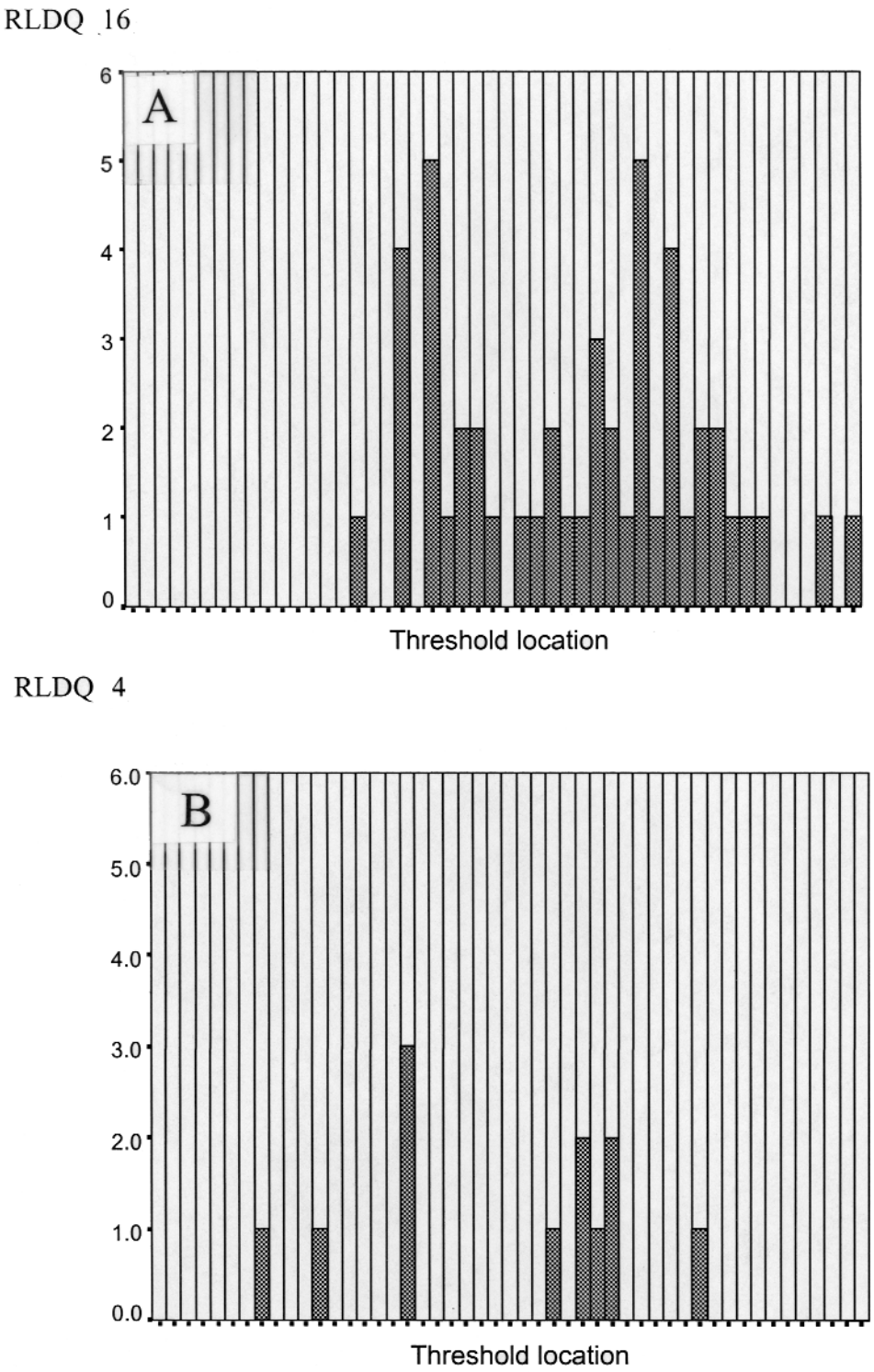
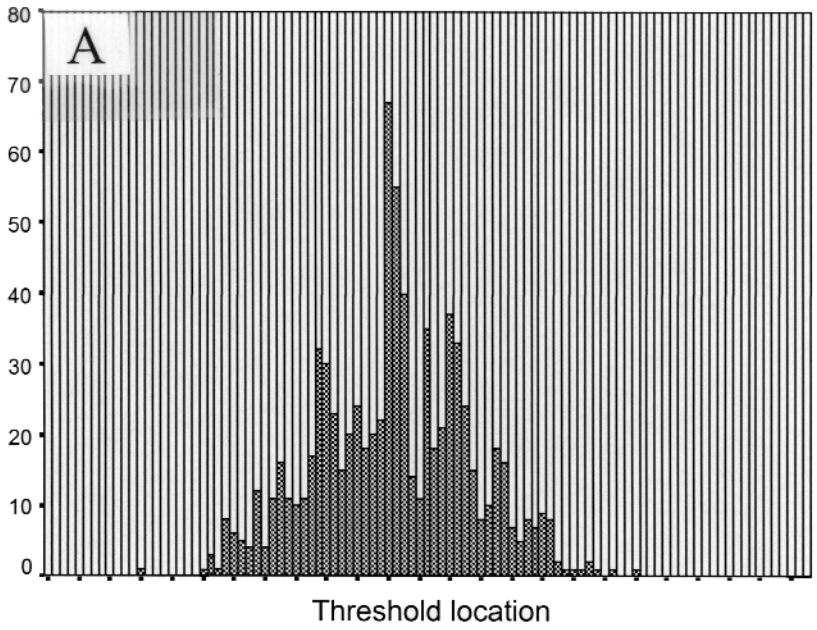


Figure 3. RLDQ 16 (A) and RLDQ 4 (B) item-threshold imprints.

divided by the median score into 2 groups): the response to item 3b (neck: crossing road) may be influenced by age as well as level of function. As an example for the BASFI, 2 patients at the same overall level of function may mark different points on "reach to shelf" according to their disease duration. If we profess to measure function (or disability) alone, such items can compromise the unidimensionality of the scale. All items functioned consistently across sex, perceived severity, and time-points.

Category and step disordering. For each of the 16 RLDQ

BASFI



RLDQ

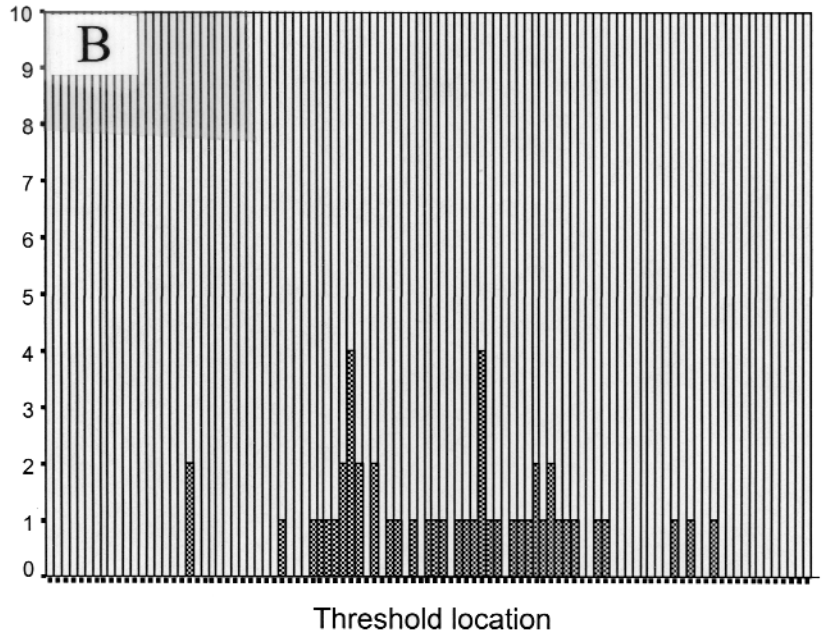


Figure 4. BASFI (A) and RLDQ 16 (B) item-threshold imprints.

items, average measures progressed in the order expected from 0 to 3. That is, for all items, the mean "ability" of patients in category 3 ("unable to do") was lower than those in category 2 ("only able to do using unusual movements"). For the BASFI, all 10 items displayed disordered categories and the mean "ability" of patients did not decrease progressively through categories 0 ("easy") to 100 ("impossible").

Similarly, we would hope that each item threshold progresses in a hierarchical order. Disordered thresholds would imply that there is no ability (or disability) level at

which a particular category becomes the most likely response; that category is not being fully utilized. This may suggest that excessive category options have been presented to respondents, or that a given category defines too narrow a portion of the underlying construct.

As examples, Table 3 presents data from items 1a and 4a from the RLDQ. Data from BASFI are not included, as with 101 potential response categories this would be too lengthy and probably not informative, given the number of respondents in this study. The 2 items in Table 3 show exemplary progression in terms of difficulty, but item 4a shows disordered thresholds. Indeed, 8 of the RLDQ items displayed disordered item thresholds, in that one or more thresholds for these items did not increase with category number. Generally speaking, the aberrant thresholds were related to response 2 ("Only able to do using unusual movements or gadgets"). Combining categories, particularly categories 2 and 3, may be beneficial in these cases. For the BASFI, all 10 items also displayed disordered item thresholds. Again, collapsing categories and using, for example, an 11 point (0–10) numeric rating scale may be beneficial.

Range of measurement. To directly compare the measurement range of each scale, the BASFI and RLDQ were calibrated onto a common scale of 26 items (16 RLDQ and 10 BASFI items). In terms of overall fit, 6 of these 26 items were above the required OUTFIT range (BASFI items 1, 6, 7, 8, 9, and 10).

Figure 4 displays the distribution of item thresholds derived for the RLDQ (16 items) and BASFI (10 items) as a whole. In the case of the BASFI, there are 1000 thresholds (100 thresholds for each of the 10 items).

In terms of measurement span, the RLDQ 16 has a slightly wider scale width (5.33 logits) than the BASFI (5.01 logits). Second, both measures have "towers" of thresholds, whereby several thresholds are marking the same point on the underlying disability construct. This is particularly notable in the case of the BASFI.

In terms of measurement level, if either scale were functioning at the interval level, the item threshold imprints would follow a logarithmic pattern. Clearly, the 2 imprints do not follow such a pattern, and thus can only be deemed ordinal.

DISCUSSION

Both the BASFI and the Dougados Functional Index (DFI) have been selected as appropriate measures of functional outcome in AS[2]. Both these measures are in common use and have undergone assessment and validation studies[9,13]. However, direct comparative studies between the measures are few, and rigorous validation studies have not been performed. Both the BASFI and DFI correlate well with external measures of disease activity and damage[14], but the BASFI is reported to have greater sensitivity to change in trials of physical therapy[15].

In this report the BASFI has been compared with an alternative existing measure of functional outcome in AS — the RLDQ. In addition to the usual psychometric properties the instruments have been compared using Rasch analysis, which permits comparison between the measures against a common scale. Direct comparison between the measures shows both to have good test-retest reliability, acceptable external validation (Figure 2), and similar unidimensionality (Table 1). Neither instrument can be used as an interval scale (Figure 4). The RLDQ has fewer items displaying differential item functioning (Table 2) and, despite having many fewer categories, has a wider measurement span than the BASFI (Figure 4). One further notable difference between the measures is seen in the item threshold imprint (Figure 4). Here, the RLDQ demonstrates fewer redundant item thresholds, although clearly neither scale is ideal in this respect.

It is important to acknowledge the cross-sectional nature of this study. Although these instruments have been compared using item response theory in order to obtain more information about their fundamental scaling properties, the analysis tells us nothing about other important and, from a research point of view, practical properties such as sensitivity to change. Indeed both the BASFI and the RLDQ have shown an ability to record appropriate changes in response to known efficacious treatments[9,10].

The BASFI is unusual in scales of this kind in that it utilizes a visual analog scale as the response to each item of the questionnaire. This arguably enhances responsiveness, in that each item has a potential 101 responses (i.e., one response for each millimeter of the 0–100 mm scale). In practice, however, respondents seldom use the full repertoire, with a preference for the point two-thirds of the way along the scale[16] — as shown for the BASFI in Figure 1. The use of an 11 point numeric rating scale might improve the properties of the BASFI.

On the other hand the RLDQ appears to show a floor effect in Figure 1. This would imply that the instrument is unable to measure improvement at the lower end of the scale. Further, although the subject is constrained to select one of 4 responses, subjects have difficulty selecting the third response — "Able to do only using unusual movements or gadgets." This difficulty was identified during the development phase of the instrument but the response was retained in order to produce a scale that was scored similarly to the Health Assessment Questionnaire (HAQ)[17]. Like the BASFI, the current analysis suggests that consideration should be given to a change in the score options for this scale, such as combining the second and third responses, so that subjects are presented with 3 response possibilities: "Without difficulty," "With difficulty," and "Impossible to do." Rasch analysis further suggests that if the instrument is scored by summing all the individual items, rather than by scoring in a manner similar to the HAQ, the scale has superior measurement properties.

One further area of uncertainty is the handling of missing data. The RLDQ as originally designed was scored in a fashion similar to the HAQ — that is, the maximum score in a section was recorded, the section scores summed, and the total score divided by the total number of sections answered. In this case, if an item or section is not completed the final score will continue to be a number from 0 to 3. If the new scoring system is adopted for the RLDQ, both the RLDQ and the BASFI will face the same problem with incomplete items — how to handle these missing data. However, the advantage of having a scale that fits the Rasch model is that estimates of a person's ability (through the separation of parameters) are independent of which (sub)set of items are answered. Thus estimates are not affected by missing items, only the precision of the estimate.

In summary, both instruments provide a measure of functional ability in ankylosing spondylitis that accords with patient perception of disease severity. Neither instrument can be used as an interval measure. The RLDQ shows a floor effect, while the BASFI exhibits a ceiling effect. The RLDQ might be improved by combining 2 of the response categories and the BASFI by using a numeric rating scale for the response options.

APPENDIX:

Items in the BASFI and RLDQ

BASFI. The BASFI comprises 10 questions to which the response is marked, on a 10 cm visual analog scale, from "easy" to "impossible." Respondents are asked to indicate their level of ability within the last week. The questions are: (1) Putting on your socks or tights without help or aids (e.g., sock aid); (2) Bending forward from the waist to pick up a pen from the floor without an aid; (3) Reaching up to a high shelf without help or aids (e.g., helping hand); (4) Getting up out of an armless dining room chair without using your hands or any other help; (5) Getting up off the floor from lying on your back without help; (6) Standing unsupported for 10 minutes without discomfort; (7) Climbing 12–16 steps without using a handrail or walking aid. One foot on each step; (8) Looking over your shoulder without turning your body; (9) Doing physically demanding activities (e.g., physiotherapy exercises, gardening or sports); (10) Doing a full day's activities whether at home or at work.

RLDQ. The RLDQ comprises 16 questions arranged in 4 domains. Respondents are asked to tick one of 4 responses: "Able to do without difficulty;" "Able to do with difficulty;" "Only able to do using unusual movements or gadgets," and "Unable to do." Respondents are asked to indicate their level of ability within the last week. The questions are : Mobility: (1a) Getting into and out of the bath; (1b) Getting into and out of the car; (1c) Getting up and out of bed in the morning; (1d) Rolling over in bed. Bending down; (2a) Wiping yourself right after using the toilet; (2b) Putting on and taking off your socks; (2c) Putting on your shoes and tying your laces; (2d) Cutting your toenails. Neck movements: (3a) Opening high windows; (3b) Looking both ways before crossing the road (e.g., do you have to move your feet); (3c) Looking at what you are reaching for on a high shelf; (3d) drinking from a small glass or can (e.g., do you have to bend your knees?). Posture: (4a) Walking on your heels; (4b) Coughing or sneezing; (4c) Sleeping on your back; (4d) Sleeping on your stomach.

REFERENCES

1. Lubrano E, Helliwell PS. Deterioration in anthropometric measures over 6 years in patients with ankylosing spondylitis: an initial comparison with disease duration and reported exercise frequency. Physiotherapy 1999;85:138-43.
2. Van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis. Progress report of ASAS Working Group. J Rheumatol 1999;26:951-4.
3. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1980.
4. Andrich D. Rasch models for measurement. Quantitative applications in the social sciences. No. 68. London: Sage Publications; 1988.
5. 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.
6. Cialdella PH, Figon G, Haugh MC, Boissel JP. Prescription intentions in relation to therapeutic information: A study of 117 French general practitioners. Soc Sci Med 1991;33:1263-74.
7. Linacre JM, Wright BD. A user's guide to Winsteps. Chicago: Messa Press; 1997.
8. Andrich D, Lyne A, Sheridon B. RUMM 2010. Version 3 for Windows. Duncraig: Rumm Laboratory Pty Ltd.: 2000.
9. Calin A, Garret 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.
10. Abbott CA, Helliwell PS, Chamberlain MA. Functional assessment in ankylosing spondylitis: Evaluation of a new self-administered questionnaire and correlation with anthropometric variables. Br J Rheumatol 1994;33:1060-6.
11. Prieto L, Alonso J, Lamarca R, Wright BD. Rasch measurement for reducing the items of the Nottingham Health Profile. J Outcome Meas 1998;2:285-301.
12. Smith RM, Schumacher RE, Bush MJ. Using item mean squares to evaluate fit in the Rasch model. J Outcome Meas 1998;2:66-78.
13. Dougados M, Boumier P, Amor B. Sulphasalazine in ankylosing spondylitis: a double blind controlled study in 60 patients. BMJ 1986;293:911-4.
14. Spoorenberg A, van der Heijde D, de Klerk E, et al. A comparative study of the usefulness of the Bath Ankylosing Spondylitis Functional Index and the Dougados Functional Index in the assessment of ankylosing spondylitis. J Rheumatol 1999;26:961-5.
15. Ruof J, Stucki G. Comparison of the Dougados Functional Index and the Bath Ankylosing Spondylitis Functional Index. A literature review. J Rheumatol 1999;26:955-60.
16. Dixon JS, Bird HA. Reproducibility along a 10 cm vertical visual analogue scale. Ann Rheum Dis 1981;40:87-9.
17. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: The Health Assessment Questionnaire, disability and pain scales. J Rheumatol 1982;9:789-93.