Improving Scoring Precision and Internal Construct Validity of the Bath Ankylosing Spondylitis Disease Activity Index Using Rasch Measurement Theory

Alice Heaney , Stephen P. McKenna , Peter Hagell , and Raj Sengupta

ABSTRACT. *Objective.* To test the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) according to Rasch Measurement Theory and investigate whether measurement precision can be improved.

Methods. Secondary analysis of a BASDAI database. The data had been collected from individuals starting an Ankylosing Spondylitis Exercise Course at the Royal National Hospital for Rheumatic Diseases in Bath, UK.

Results. Data were available for 250 participants (23.6% female) aged between 18 and 85 years (mean 52.8, SD 14.6). Initial fit of the data to the Rasch model appeared good and item thresholds were consistent, but local item dependence (LID) was identified. After addressing the LID, a unidimensional measure was achieved. The Person Separation Index (reliability) was 0.83 and the location of the items was well matched to that of the respondents. A transformation table was generated to convert total raw BASDAI scores into linearized Rasch transformed scores that form an interval scale. The Smallest Detectable Difference improved from 2 to 1.2. This finding suggests that a change score of > 1.2 points on the modified BASDAI is required to achieve meaningful change.

Conclusion. Applying the Rasch transformed scores simplifies completion and scoring of the measure and confirms internal construct validity. It also ensures linear measurement and justifies the use of parametric statistical analyses when analyzing datasets. The transformation table can be used with existing BASDAI datasets to allow direct comparisons of disease activity scores with those generated from future studies. (First Release October 15 2019; J Rheumatol 2020;47:354–61; doi:10.3899/jrheum.180943)

Key Indexing Terms:
ANKYLOSING SPONDYLITIS
DISEASE ACTIVITY

EPIDEMIOLOGY
OUTCOME ASSESSMENT

Ankylosing spondylitis (AS) is a chronic rheumatic disease that targets the axial skeleton, causing inflammatory back pain. Characteristic symptoms of AS include spinal stiffness and loss of spinal mobility¹. The prevalence of AS is estimated between 0.2% and 0.5%². In the context of radiographic disease, about twice as many men as women are affected by the condition³.

In AS clinical trials, 3 key disease components are commonly assessed: disease activity, physical functioning, and structural damage⁴. Disease activity in AS has not been clearly defined because of the variation in the clinical picture for different patients⁵.

The most frequently used questionnaire for assessing disease activity in AS clinical trials is the Bath Ankylosing

Spondylitis Disease Activity Index (BASDAI)⁶. It was developed in Bath, England, by a multidisciplinary team of rheumatologists, physiotherapists, and researchers. It is completed by patients and has used a range of different response formats. The version used in our present study has an 11-point numerical rating scale (NRS). The BASDAI is quick and easy to complete, demonstrates good psychometric properties, and has been found to be sensitive to change⁶. Calin, *et al* used factor analysis to define disease activity in AS using the BASDAI⁷. They concluded that all items could be summed to provide a total score.

The National Institute for Health and Care Excellence (NICE) guidance recommends that the BASDAI be used to evaluate selection and response to biologic treatments for prescribing purposes⁴. A BASDAI score ≥ 4 is one of the eligibility criteria for initiation of biologic treatment. This requirement also applies in Australia⁸. A score of 4 or more is considered by the Assessment of Spondyloarthritis international Society Working Group to indicate the presence of active disease and is the cutoff value generally required for individuals to be included in clinical trials⁹.

Franchignoni and colleagues, using the Rasch measurement theory (RMT), found that the number of response

From Galen Research, Manchester, UK.

A. Heaney, BSc, Galen Research; S.P. McKenna, PhD, Galen Research and The University of Manchester; P. Hagell, PhD, Kristianstad University; R. Sengupta, MD, Royal National Hospital for Rheumatic Diseases.

Address correspondence to A. Heaney, Galen Research, B1 Chorlton Mill, 3 Cambridge St., Manchester M1 5BY, UK.

E-mail: aheaney@galen-research.com

Accepted for publication May 2, 2019.

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2020. All rights reserved.

categories used with the BASDAI exceeds the number of levels of a construct that participants can discriminate¹⁰. Further, the BASDAI was developed without the application of RMT, which offers a more powerful paradigm for the measurement of latent traits¹¹. RMT¹² is a simple logistic unidimensional measurement model that satisfies fundamental measurement requirements¹³. RMT is applied when a set of items in a scale are intended to be summed together to represent a common unidimensional latent variable¹⁴. Unless unidimensionality has been established, it is not valid to add together the scores for any set of items¹⁵.

The aim of our study was to determine whether the BASDAI meets the requirements of RMT, and if so, whether the measure could be simplified without losing measurement precision.

MATERIALS AND METHODS

The Bristol UK local research ethics committee (LREC) for National Health Service research approved the study, and all participants provided written informed consent (The Bath Spondyloarthritis Biobank; REC reference: 13/SW/0096).

Participants. BASDAI data were available from participants starting an Ankylosing Spondylitis Exercise Course at the Royal National Hospital for Rheumatic Diseases in Bath, UK. A target sample size of 250 was selected for the survey because that is within the recommended range for use with the RUMM software¹⁶ and gives 99% confidence that RMT item parameter estimates are within half a logit of the stable value¹⁷.

Details about the BASDAI. The BASDAI is a patient-reported outcome measure (PROM) designed to represent disease activity in AS⁶. It consists of 6 items: severity of fatigue, spinal pain, peripheral joint pain, localized tenderness (enthesitis), and severity and duration of morning stiffness. An 11-point NRS (scored 0–10) is used for each of the questions. A higher score on the BASDAI indicates greater disease activity. To score the questionnaire, the mean of the 2 items assessing morning stiffness is calculated. This value is added to the sum of the scores on the remaining items. The total is divided by 5 to give a score ranging from 0 to 10.

Rasch analyses. Internal measurement validity of the BASDAI was tested by RMT¹². RMT assumes that the probability of affirming an item is a logistic function of the relative distance between the item location parameter and the respondent location parameter. Regarding disease activity, the item location parameter refers to the severity represented by the item and the respondent location parameter refers to the patient's AS activity. RMT provides a method of transforming raw scores at the ordinal level into interval-level latent trait units, referred to as logits¹⁸. Interval-level measurement implies that differences between scores are equal along the whole measurement continuum (i.e., the difference in disease activity between 1 and 2 is the same as that between 6 and 7 or between 9 and 10). Where the data fit the assumptions of RMT, conceptual support for a common latent variable is provided and the scale can be used to produce interval-level measures rather than ordinal raw scores.

The person separation index (PSI) was used to assess internal reliability of the scale. A value of 0.7 is the minimum acceptable PSI level¹⁴. RMT model fit was assessed by a range of approaches based on the accordance between observed data and model expectations. The chi-square interaction statistic is used to indicate overall fit to the model. A significant chi-square statistic ($p < 0.05$) indicates lack of fit to the model. Individual item fit was investigated by standardized fit residuals. These are expected to fall within the range ± 2.5 . Individual item level chi-square-based statistics and item level ANOVA tests of residuals across class intervals are further tests of item level misfit that were also examined. Following Bonferroni corrections, if $p < 0.05$, there is item misfit.

A principal component analysis (PCA)-based method was used to provide further support for unidimensionality. This requires the identification of 2 item sets from a PCA of residuals that load most differently on the first component. A series of t tests is then conducted to assess whether the 2 separate sets of person measures produce significantly different estimates for each participant. Where fewer than 5% of these tests are statistically significant, or the lower bound of a binomial 95% CI overlaps 5%, unidimensionality can be inferred¹⁹. However, it must be noted that this should not be considered an ultimate test for unidimensionality²⁰.

An important requirement of RMT model fit is item invariance across groups. In the current analysis, an ANOVA of standardized residuals was used to examine differential item functioning (DIF). Uniform DIF occurs when different sample groups (e.g., males and females) have different response probabilities on a specific item, despite having the same level of the latent trait²¹. One way to resolve this is by treating items exhibiting DIF as different items for different groups. This results in DIF-free person estimates and is known as *splitting for DIF*. Nonuniform DIF is manifested when there is an interaction effect between subgroup affiliation and person location. Formally, an ANOVA p value of < 0.05 (Bonferroni corrections applied) indicates the presence of DIF.

Items in a measure should have local independence. This implies that correlations between items in the scale result from the latent trait (disease activity in this case). Where this is not the case, local item dependence (LID) is present. There are 2 types of LID: trait and response dependence (RD). The former suggests the presence of more than 1 latent trait, while the latter implies that the response on 1 item influences the response to 1 or several other items²². RD can be controlled by combining the affected items into a single item, which is referred to as a subtest¹⁴. LID should be considered relative to the average residual correlation for all items. A value of 0.2 above the average residual correlation is indicative of LID²³.

The ordering of the response thresholds for each item was examined to establish whether the response categories were consistent with the intended order¹⁴.

The Rasch Unidimensional Measurement Model (RUMM) 2030 program was used to conduct the analyses²⁴.

Prior to Rasch analysis, the BASDAI items were reviewed from a clinical perspective, according to their expected hierarchy. A high score on a BASDAI item indicates "high disease activity." "Easy" items on a scale are those for which it is easy for respondents to select high scores. For "difficult" items, it is difficult for respondents to select high scores. It was expected that the items relating to fatigue and spinal pain would represent "easy" items, because they are experienced by most people with AS^{25,26}. Up to half of the AS population is expected to have arthritis in peripheral joints or peripheral entheses at some stage of their disease²⁷. Consequently, at least half the current sample would find it very difficult to score highly on the peripheral pain item, making it a difficult item. This hypothesis is supported by Heuft-Dorenbosch and colleagues who found that disease activity, as measured by the BASDAI, was higher in patients with peripheral joint disease than those with disease limited to the axial skeleton²⁸. This was true even when the peripheral pain item was excluded from the BASDAI.

Additional analyses. A measure is responsive if it can detect real changes in the outcome being measured. An indication of potential responsiveness is the smallest detectable difference (SDD), i.e., the change in score on the measure for which anything smaller cannot be reliably distinguished from random measurement error²⁹. It is based on the Standard Error of Measurement (SEM), SD, and reliability and is calculated as follows:

$$\text{SDD} = \text{SEM} \times 1.96 \times \sqrt{2}$$

$$\text{Where SEM} = \text{SD} (\sqrt{(1 - \text{reliability})})$$

The SDD can also be expressed as a percentage of the full operational range of the scale under consideration. SDD is heavily dependent on the reproducibility/reliability of the outcome measure concerned.

RESULTS

Internal construct validity of traditional BASDAI. BASDAI data were available for 250 participants (23.6% female) aged between 18 and 85 years (mean 52.8, SD 14.6). In accordance with the traditional method of scoring the BASDAI, scores for the 2 items assessing morning stiffness were averaged and treated as 1 item. RMT fit statistics are shown in Table 1. Despite the PCA/t test approach inferring unidimensionality, insurmountable problems related to LID and item misfit were identified. Given the conceptual importance of the items that displayed misfit, their removal could not be justified.

Identifying a BASDAI scale that fits RMT. In applying RMT to represent a common latent variable, the scores for a set of items in a scale should be added together to produce a total score. Therefore, further analyses were undertaken in which all 6 items were treated as individual items, i.e., the morning stiffness items were not averaged. The results of the analysis revealed that all items showed acceptable fit indices according to RMT (fit residuals -2.2 to 2.0, $p > 0.05$) and there was no DIF by sex or age. Responses to the NRS had ordered thresholds for all 6 items, as shown in Figure 1. However, evidence of LID was found.

Analysis of the residual correlations indicated marked RD between the 2 items assessing morning stiffness. The duration of morning stiffness item was removed because it differs

from the other 5 items in that it assesses duration rather than severity. The inclusion of this item also complicates the scoring of the BASDAI, because of the need to take the mean of the responses to the 2 morning stiffness items. Removal of the item only had a marginal effect on the location of participants along the measurement continuum.

However, removal of the duration of morning stiffness item resulted in mild uniform DIF by sex becoming apparent on the peripheral joint pain item ($p < 0.05$, Bonferroni adjusted). At all levels of disease activity, males reported more peripheral joint pain than females. A paired samples t test indicated that there was virtually no difference between unsplit person estimates ($M = -0.43$, $SD\ 0.86$) and split person estimates [$(M = -0.43$, $SD\ 0.87$); $t(248) = -0.09$, $p = 0.93$, $d = 0.01$]. Further investigation indicated that the DIF did not cause misfit to the Rasch model at the scale level. Consequently, it was decided that the peripheral joint pain item could remain in the scale unchanged without a clinically meaningful effect on BASDAI-based person measures.

Reanalysis of the residual correlations identified 2 further pairs of items displaying RD; the items assessing spinal pain and morning stiffness and those addressing peripheral joint pain and enthesitis. Each of these pairs of items was combined into a subtest.

Resolving the RD reduced the reliability (PSI) of the scale (Table 1). Following the creation of the 2 subtests, the measurement range and SD of the scale also declined. The PCA/t test approach revealed that unidimensionality can be inferred from the modified BASDAI scale.

All 5 items in the modified BASDAI had acceptable fit indices (fit residuals -1.0 to 1.1, $p > 0.05$). Targeting of the items in the new scale was excellent, as can be seen in the person-item threshold distribution shown in Figure 2. The locations of respondents are displayed in the top half of the figure, with the locations of items shown in the bottom half. The items were well targeted to the severity range exhibited by respondents.

Inspection of the item locations revealed that those assessing severity of fatigue and spinal pain represented “easy” items to affirm. The peripheral joint pain item was the

Table 1. Summary of Rasch analyses and person locations of traditional and modified BASDAI.

Variables	Traditional 6-item BASDAI*	Modified 5-item BASDAI
PSI	0.86	0.83
Chi-square interaction	0.46	0.48
PCA/t test (CI)	0.07 (0.05–0.10)	0.04 (0.01–0.07)
Person locations, 0–10		
Mean	4.50	4.68
SD	1.00	1.05
Range	0–7.47	0–7.71

*Items 5 and 6 have been averaged and treated as 1 item. BASDAI: Bath Ankylosing Spondylitis Disease Activity Index; PSI: person separation index; PCA: principal component analysis.

Figure 1. Response threshold map. All items showed ordered thresholds and are displayed in location order.

Figure 2. Location of items and respondents on the modified BASDAI scale. BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.

most “difficult,” confirming the earlier predictions. This item ordering was maintained following the creation of the subtests.

Total BASDAI raw scores range from 0 to 10. To generate the true BASDAI score, the logit value for each total raw score was transformed to range from 0 to 10. The Rasch

transformation scores for the modified BASDAI are presented in Appendix 1.

The relation between total raw scores and Rasch transformed (true) scores are shown in Figure 3. These vary markedly, particularly at the extremes. A raw score of 1 corresponds to a true score of 3.1 and a raw score of 9 to a true

Figure 3. Relation between total raw scores and Rasch transformed scores for the modified BASDAI. BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.

score of 7.1. It is also clear from the figure that a person whose raw score changes from 8 to 2 has a true change score of only 2 points.

Smallest detectable difference of the BASDAI. The SDD of the traditional BASDAI based on raw scores (traditional scoring) was 2. Using the Rasch transformed scores, the SDD was reduced to 1.2. These values equate to 20% and 12% of the measurement range, respectively.

DISCUSSION

The BASDAI is a widely used questionnaire for assessing disease activity in AS. Given that it was developed without the application of RMT, the instrument performed well in the current analyses. Despite this, improvements to the traditional BASDAI were necessary to meet the requirements of RMT. This resulted in the identification of a modified 5-item BASDAI measure that demonstrated good precision and internal validity.

It was surprising to find that the response thresholds for all items were ordered, given Franchignoni, *et al*’s claim that 11 categories are too difficult to discriminate between¹⁰. An attempt to collapse the central responses led to considerable disordering of the thresholds. It has been argued that this is because collapsing categories is only justified if the discrimination at the threshold between 2 categories is 0³⁰. Therefore, the decision was taken to maintain the original 11 response categories. The traditional method of scoring the BASDAI revealed that the PCA/t test approach met the requirements for unidimensionality, despite evidence of misfit to the Rasch model. This highlighted the argument presented by Hagell²⁰ in his demonstration of why this test should not be considered to provide definite evidence of the unidimensionality of a scale. It was hardly surprising to find that a unidimensional version of the BASDAI could not be identified according to the traditional scoring methods. There is no conceptual basis for combining the morning stiffness items, given that they assess 2 different outcomes (severity and duration). Further, each item in a “unidimensional” scale should contribute to the total score, and so averaging the responses absorbs the LID that emerged when each item was treated separately.

Removal of the item enquiring into duration of morning stiffness makes the scale simpler to administer and score without loss of information. However, removal of the item led to DIF by sex on the peripheral joint pain item. Further investigation indicated that this DIF had minimal effect on the scores of respondents and it was therefore decided to maintain this item. RD identified in 2 pairs of items was addressed by producing a separate subtest for each pair. This finding confirmed the importance of checking for LID in existing clinical PROM and paying attention to relative, as opposed to absolute, residual correlations³¹. For the 5-item BASDAI this is accounted for in the Rasch transformed scores. The new method of scoring the BASDAI improves

the measurement precision of the scale and produces valid interval level scores for all participants. Without making these corrections, the reliability and dispersion of the scale would be overestimated²². Further advantages of using the 5-item modified BASDAI are summarized in Table 2.

It is possible to reanalyze existing BASDAI datasets using the new scoring system. This will allow the results of future studies to be compared with previous work, in which scores for individual BASDAI items are available.

The predicted ordering of item difficulty was confirmed by the initial Rasch analysis. The finding that severity of fatigue and spinal pain were the “easiest” items to affirm and peripheral joint pain the most “difficult” supports previous research^{25,26,27,28}. This confirmation of the *a priori* expectations provides further evidence of the scale’s construct validity.

An advantage of RMT is that measures producing data that fit the Rasch model can be transformed to interval, rather than ordinal level measurement³². This allows the application of parametric statistics, providing reliable and accurate comparisons of change within and between individuals³³. The importance of using Rasch transformed scores was illustrated in Figure 3. Change in raw score along the measurement continuum does not result in equivalent change in true score¹⁸.

As can be seen from the figure, raw scores and changes in raw scores should be interpreted with caution. This is because low raw BASDAI scores underestimate true disease activity. Similarly, high raw BASDAI scores overestimate true disease activity. This has important implications for the calculation of SDD and minimum clinically importance difference (MCID) scores. SDD and MCID calculations based on ordinal data understate true change at the margins and overstate the importance of changes in the midrange of a scale³⁴. Despite the importance of using interval level data, inappropriate statistical tests are still widely applied to ordinal level data, resulting in inaccurate conclusions being drawn³⁵.

In the United Kingdom, NICE requires a reduction in BASDAI score to 50% of the pretreatment value or by 2 or more units and a reduction of 2 cm or more on a 10-cm spinal

Table 2. Advantages of using the 5-item modified BASDAI.

The Modified BASDAI:

- Increases precision of measurement
- Reduces burden on patients (1 less item to complete)
- Reduces burden on scorers (no averaging of responses to 2 items)
- Transformation table converts raw scores to Rasch transformed scores
- Measurement at the interval rather than ordinal level
- Allows valid calculation of means and SD
- Allows application of parametric statistical analyses (such as t tests and ANOVA)
- Reduces sample sizes needed for clinical studies and trials

BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.

pain visual analog scale (VAS) for treatment with tumor necrosis factor- α (TNF- α) inhibitors to be continued⁴. A change in raw score of 2 units on the BASDAI represents different amounts of change along the measurement scale. An improvement in raw score from 2 to 0 is a true improvement of 4. An improvement from 6 to 4 represents a true improvement of < 0.5. In one study, a pain VAS was found to overestimate true responsiveness by 59%³⁶. The authors concluded that raw pain VAS data should not be used as a primary outcome in clinical studies. Considering these BASDAI and pain VAS findings, it is recommended that NICE amend their requirements for continuation of TNF- α inhibitor treatment.

The SDD for the BASDAI improved from 2 to 1.2. To put these values into context, Table 3^{37,38,39,40,41,42,43,44} shows SDD values for other frequently used PROM in rheumatology (applying their recommended scoring procedures). The SDD data are estimated from test-retest reliabilities and SD reported in previous studies. It can be seen from the table that the change in percentage of measurement range required to achieve SDD is as high as 83% for the Medical Outcomes Study Short Form-36 survey and 51% for the EQ-5D. Consequently, most of these scales would be unlikely to detect true changes in score. Only the Ankylosing Spondylitis Quality of Life score and Health Assessment Questionnaire approach the SDD value achieved by the Rasch transformed BASDAI. Given the results of our present study, an improvement in score in the region of or > 1.2 points anywhere along the modified BASDAI interval scale would provide a more accurate indication of effective treatment.

Adapting the BASDAI by removing the item on duration of morning stiffness, creating 2 subtests, and applying Rasch transformed scores simplifies completion and scoring of the measure and enables it to provide interval-level scores, suitable for analysis using parametric statistical analyses. The low SDD value achieved also indicates that it is more likely to be responsive to true changes in disease activity. However, formal testing of responsiveness would require an intervention study.

REFERENCES

- Braun J, Sieper J. Ankylosing spondylitis. *Lancet* 2007;369:1379-90.
- Reveille JD. Epidemiology of spondyloarthritis in North America. *Am J Med Sci* 2011;341:284-6.
- Feldtkeller E, Khan MA, van der Heijde D, van der Linden S, Braun J. Age at disease onset and diagnosis delay in HLA-B27 negative vs. positive patients with ankylosing spondylitis. *Rheumatol Int* 2003;23:61-6.
- National Institute for Health and Care Excellence. TNF-alpha inhibitors for ankylosing spondylitis and non-radiographic axial spondyloarthritis. NICE technology appraisal guidance TA383. [Internet. Accessed August 29, 2019.] Available from: <https://www.nice.org.uk/guidance/ta383>
- Spoorenberg A, van Tubergen A, Landewé R, Dougados M, van der Linden S, Mielants H, et al. Measuring disease activity in ankylosing spondylitis: patient and physician have different perspectives. *Rheumatology* 2005;44:789-95.
- Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. *J Rheumatol* 1994;21:2286-91.
- Calin A, Nakache JP, Gueguen A, Zeidler H, Mielants H, Dougados M. Defining disease activity in ankylosing spondylitis: is a combination of variables (Bath Ankylosing Spondylitis Disease Activity Index) an appropriate instrument? *Rheumatology* 1999;38:878-82.
- Lim I. PBS criteria to access TNF inhibitors for ankylosing spondylitis in Australia. [Internet. Accessed November 21, 2018.] Available from: <http://www.bjchealth.com.au/connected-care/pbs-criteria-to-access-tnf-inhibitors-for-ankylosing-spondylitis-in-australia>
- Braun J, Pham T, Sieper J, Davis J, van der Linden S, Dougados M, et al; ASAS Working Group. International ASAS consensus statement for the use of anti-tumour necrosis factor agents in patients with ankylosing spondylitis. *Ann Rheum Dis* 2003; 62:817-24.
- Franchignoni F, Salaffi F, Ciapetti A, Giordano A. Searching for optimal rating scales in the Bath Ankylosing Spondylitis Functional Index (BASFI) and Bath Ankylosing Spondylitis Disease Activity Index (BASDAI). *Rheumatol Int* 2014;34:171-3.
- Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technol Assess* 2009;13:1-177.
- Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
- Andrich D. Rasch models for measurement. Beverly Hills, CA: Sage; 1988.

Table 3. Commonly used PROM SDD and change in percentage of measurement range required to achieve SDD.

Type of Outcome	PROM	SEM	SDD	Change in Measurement Range Required for SDD	Data Sources
NRS	Pain intensity	0.12	34.4	34.4%	37
Disability	HAQ	0.25	0.69	23.1%	38, 39
HRQOL	SF-36 sections (range)	6.9–30.2	19.1–83.7	19.1–83.7%	40
QOL	ASQoL	1.22	3.37	18.7%	41
Utility	EQ-5D	0.22	0.61	50.8%	42
	SF-6D	0.11	0.31	47.7%	43
	HUI-3	0.15	0.42	30.9%	44

PROM: patient-reported outcome measure; SEM: standard error of measurement; SDD: smallest detectable difference; NRS: numerical rating scale; HAQ: Health Assessment Questionnaire; HRQOL: health-related quality of life; SF-36: Medical Outcomes Study Short Form-36 survey; QOL: quality of life; ASQoL: Ankylosing Spondylitis Quality of Life questionnaire; SF-6D: Short Form-6 Dimensions; HUI-3: Health Utilities Index Mark 3.

14. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res* 2007;57:1358-62.
15. Streiner D, Norman G. Health measurement scales. Oxford: Oxford University Press; 1989.
16. Hagell P, Westergren A. Sample size and statistical conclusions from tests of fit to the Rasch model according to the Rasch Unidimensional Measurement Model (RUMM) program in health outcome measurement. *J Appl Meas* 2016;17:416-31.
17. Linacre JM. Sample size and item calibration stability. *Rasch Measurement Transactions* 1994;7:328.
18. Doucette A, Wolf AW. Questioning the measurement precision of psychotherapy research. *Psychother Res* 2009;19:374-89.
19. Smith EV Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *J Appl Meas* 2002;3:205-31.
20. Hagell P. Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: the primacy of theory over statistics. *Open J Stat* 2014;4:456-65.
21. Hagquist C, Andrich D. Is the sense of coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Pers Individ Dif* 2004;36:955-68.
22. Marais I, Andrich D. Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *J Appl Meas* 2008;9:200-15.
23. Christensen KB, Makransky G, Horton M. Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Appl Psychol Meas* 2017;41:178-94.
24. Rasch Unidimensional Measurement Model Version RUMM2030 [Computer software]. Perth: RUMM Laboratory; 2010.
25. Farren W, Goodacre L, Stigant M. Fatigue in ankylosing spondylitis: causes, consequences and self-management. *Musculoskeletal Care* 2013;11:39-50.
26. Bidad K, Gracey E, Hemington KS, Mapplebeck JCS, Davis KD, Inman RD. Pain in ankylosing spondylitis: a neuro-immune collaboration. *Nat Rev Rheumatol* 2017;13:410-20.
27. Taurog JD, Chhabra A, Colbert RA. Ankylosing spondylitis and axial spondyloarthritis. *N Engl J Med* 2016;374:2563-74.
28. Heuft-Dorenbosch L, van Tubergen A, Spoorenberg A, Landewé R, Dougados M, Mielants H, et al. The influence of peripheral arthritis on disease activity in ankylosing spondylitis patients as measured with the Bath Ankylosing Spondylitis Disease Activity Index. *Arthritis Rheum* 2004;51:154-9.
29. Lassere MN, van der Heijde D, Johnson KR, Boers M, Edmonds J. Reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for smallest detectable difference, minimal clinically important difference, and analysis of treatment effects in randomized controlled trials. *J Rheumatol* 2001; 28:892-903.
30. Andrich D. Understanding the response structure and process in the polytomous Rasch model. In: Nering ML, Ostini R, editors. *Handbook of polytomous item response theory models*. New York: Routledge; 2011:133-62.
31. Marais I. Local dependence. In: Christensen KB, Kreiner S, Mesbah M, editors. *Rasch models in health*. London: Wiley; 2013:111-30.
32. Bond TG, Fox CM. *Applying the Rasch model: fundamental measurement in the human sciences*. New Jersey: Erlbaum; 2001.
33. Avery LM, Russell DJ, Raina PS, Walter SD, Rosenbaum PL. Rasch analysis of the gross motor function measure: validating the assumptions of the Rasch model to create an interval-level Measure. *Arch Phys Med Rehabil* 2003;84:697-705.
34. Doganay Erdogan B, Leung YY, Pohl C, Tennant A, Conaghan PG. Minimal clinically important difference as applied in rheumatology: an OMERACT Rasch Working Group systematic review and critique. *J Rheumatol* 2016;43:194-202.
35. Kahler E, Rogausch A, Brunner E, Himmel W. A parametric analysis of ordinal quality-of-life data can lead to erroneous results. *J Clin Epidemiol* 2008;61:475-80.
36. Kersten P, White PJ, Tennant A. Is the pain visual analogue scale linear and responsive to change? An exploration using Rasch analysis. *PLoS One* 2014;9:e99485.
37. Myles PS, Myles DB, Gallagher W, Boyd D, Chew C, MacDonald N, et al. Measuring acute postoperative pain using the visual analog scale: the minimal clinically important difference and patient acceptable symptom state. *Br J Anaesth* 2017;118:424-9.
38. Krishnan E, Tugwell P, Fries JF. Percentile benchmarks in patients with rheumatoid arthritis: Health Assessment Questionnaire as a quality indicator (QI). *Arthritis Res Ther* 2004;6:R505-13.
39. Thorsen H, Hansen TM, McKenna SP, Sørensen SF, Whalley D. Adaptation into Danish of the Stanford Health Assessment Questionnaire (HAQ) and the Rheumatoid Arthritis Quality of Life Scale (RAQoL). *Scand J Rheumatol* 2001;30:103-9.
40. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). *Br J Rheumatol* 1998;37:425-6.
41. Doward LC, Spoorenberg A, Reynolds S, Whalley D, Helliwell PS, McKenna SP, et al. The development of the ASQoL: a quality of life instrument specific to ankylosing spondylitis. *Ann Rheum Dis* 2003;62:20-6.
42. Shikier R, Willian MK, Okun MM, Thompson CS, Revicki DA. The validity and responsiveness of three quality of life measures in the assessment of psoriasis patients: results of a phase II study. *Health Qual Life Outcomes* 2006;4:71.
43. Ferreira LN, Ferreira PL, Pereira LN, Rowen D, Brazier JE. Exploring the consistency of the SF-6D. *Value Health* 2013;16:1023-31.
44. Gooch K, Feeny D, Wong RL, Kupper H, Pangan AL, Revicki DA, et al. Is the Health Utilities Index 3 valid for patients with ankylosing spondylitis? *Value Health* 2011;14:160-5.

APPENDIX 1. Modified BASDAI Rasch transformation table.

Raw Score	Rasch Transformed Score	Raw Score	Rasch Transformed Score
0	0.00	5.2	5.16
0.2	1.12	5.4	5.21
0.4	1.86	5.6	5.26
0.6	2.36	5.8	5.32
0.8	2.74	6	5.37
1	3.05	6.2	5.43
1.2	3.30	6.4	5.49
1.4	3.51	6.6	5.55
1.6	3.69	6.8	5.62
1.8	3.84	7	5.69
2	3.98	7.2	5.77
2.2	4.11	7.4	5.86
2.4	4.22	7.6	5.96
2.6	4.32	7.8	6.07
2.8	4.41	8	6.19
3	4.49	8.2	6.33
3.2	4.57	8.4	6.48
3.4	4.64	8.6	6.66
3.6	4.71	8.8	6.86
3.8	4.77	9	7.09
4	4.83	9.2	7.36
4.2	4.89	9.4	7.71
4.4	4.95	9.6	8.17
4.6	5.00	9.8	8.89
4.8	5.05	10	10.00
5	5.10	—	—

Raw scores for the modified BASDAI can be calculated by taking the average of the scores for items 1 to 5.
 BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.