Physical functioning in Ankylosing Spondylitis patients

van Weelij, S.F.E.

2015

Chapter 2

Reproducibility of performance measures of physical function based on the BASFI, in ankylosing spondylitis

Salima F.E. van Weely
J. Christiaan van Denderen
Irene E. van der Horst-Bruinsma
Mike T. Nurmohamed
Ben A.C. Dijkmans
Joost Dekker
Martijn P.M. Steultjens

Published in: Rheumatology (Oxford) 2009, 48(10): 1254-60.

Abstract

Objective: To establish the test–retest reproducibility of performance measures of physical function based on the Bath Ankylosing Spondylitis Functional Index (BASFI) questionnaire in patients with AS.
Methods: Data were obtained from 65 AS patients. They were tested on two occasions by one assessor with a 1-week interval. Physical function was assessed via eight performance measures based on items used in the BASFI questionnaire, representing activities of daily life, which AS patients frequently report to be problematic. For each activity, a performance score was determined. Pain and exertion were measured using a 10-cm horizontal visual analogue scale (VAS) and Borg's modified scale, respectively. Test-retest reproducibility was assessed for all measurements using intraclass correlation coefficients (ICCs) and by calculating the standard error of measurement (SEM).
Results: Adequate intrarater reliability was found. For performance scores, ICCs ranged from 0.73 to 0.96. Measurements of exertion and pain also showed adequate intrarater reliability, with the exception of one performance measure, namely the test for the ability to look over one's shoulder. For this test, the ICCs were 0.66 and 0.69 for exertion and pain, respectively. The remaining ICCs for exertion ranged from 0.71 to 0.88 and for pain from 0.74 to 0.83. The SEM for performance scores ranged from 4 to 9% of the observed score. The SEM for exertion ranged from 8 to 11% and for pain from 10 to 15%.
Conclusions: Performance measures of physical function based on the BASFI questionnaire have adequate to excellent test–retest reproducibility. Due to the presence of measurement error, measurements are accurate for group assessment; repeated measurements are advised for an adequate assessment of individual patients.

Introduction

Ankylosing Spondylitis (AS) is a chronic, progressive, inflammatory disease that mainly affects the axial skeleton and the SI joints, causing characteristic inflammatory back pain, spinal stiffness and fusion. Patients with AS gradually experience a reduction in physical functioning and quality of life over time [1–4].

For the assessment of physical function, two approaches can be used: self-report and performance measures [5–7]. Both measurements have their advantages and disadvantages. Self-report measures are brief, inexpensive and easy to administer [6, 8–10], capable of evaluating multiple aspects of physical function in a single test [10], are not influenced by observer bias [11] and reflect patients' point of view. Self-reported measures are traditionally used to establish the effect of interventions on physical function in AS. The ASAS Society (Assessment of SpondyloArthritis international Society) recommends the Bath Ankylosing Spondylitis Functional Index (BASFI) [12] or Dougados Functional Index (DFI) [13] for the assessment of the domain of physical function [14, 15]. Both BASFI and DFI are self-administered, disease-specific instruments. The clinimetric properties of these instruments have been shown to be adequate [12, 16–19].

Performance measures offer the potential for greater reproducibility and responsiveness [20]. Performance measures are considered to be less influenced than self-report measures by discrepancies between perceptions of a person's ability and their true performance (underestimation or overestimation). These discrepancies can occur due to personality traits, depression [20], poor cognitive function, language, educational level [5, 6, 8, 20–22], expectations and pain [11]. Thus, performance tests offer the potential to overcome some of the limitations of self-report measures. However, the subject's motivation to participate is a potential limitation of performance measures of physical function [20]. Furthermore, performance measures provide little information about the person's ability to adapt to his or her own environment [6, 8, 11, 20, 23].

Interestingly, clinical observations of functional disability in AS are scarce. Calin et al. [12] and Hidding et al. [24] reported on the external validation of the BASFI or showed the importance of self-reported measurements, respectively. No information on reproducibility of performance tests was reported.

Based on the items of the BASFI questionnaire, we developed performance measures of physical function in patients with AS. The aim of this study was to establish the reproducibility (intrarater reliability and agreement) [25, 26] of performance measures of physical function in patients with AS.

Patients and methods

Patients were recruited from a large outpatient centre for rheumatology and rehabilitation, the Jan van Breemen Institute in Amsterdam. Enrolment took place from May 2006 until June 2007. The following inclusion criteria were applied: diagnosis of AS according to the modified New York criteria [27], ≥18 years of age and sufficient command over the Dutch language. Patients with pulmonary, cardiovascular or neurological comorbidity affecting the patient's ability to do daily activities were excluded. The medication had to be stable during the previous 3 months before inclusion.

The study was approved by the medical ethical committee of the Slotervaart Hospital, Jan van Breemen Institute and BovenIJ Hospital. All patients gave written informed consent according to the Declaration of Helsinki [28].

Design

The aim of the study was to determine test–retest reproducibility for eight performance measures of physical function based on the items of the BASFI questionnaire. Patients were measured twice with a 1-week interval between measurements (i.e., on Days 1 and 8). Measurements were taken at a similar time of the day on both the occasions. It was assumed that during this period, the patient's physical condition remained unchanged. Therefore, any observed performance differences between the two sessions were considered to be a result of measurement error. The same assessor measured all subjects on both the occasions, but was blinded regarding Day 1 results for the second occasion. Performance measures of physical functioning The following tests were derived from the items in the BASFI questionnaire [12] and were carried out by the patients in both sessions and in the subsequent order. The time to perform Tests 1–6 and 8 was measured in seconds. In advance, it was assumed that Tests 4, 5 and 6 would have more variance. To minimize this effect, these tests were repeated three times. The mean value of three repetitions was used for analysis.

Test 1: Climbing stairs. Patients stood 25 cm in front of a flight of 12 steps (height*depth = 15*30 cm) and were instructed to climb the steps without using the handrail or a walking aid by placing one foot on each step. Both feet had to be placed on the 12th step.

Test 2: Bending. A shelf was placed at the same height as the iliac crest on the right side of the patient. Six pens were placed on the floor in front of the patient (with a distance of 50 cm from the heel of the patient's foot to the centre of the pen). Patients were instructed to bend forward from the waist and pick up the six pens from the floor without an aid and place them on the shelf one by one using either one of the hands. All six pens had to be placed on the shelf.

Test 3: Reaching. Above the shelf, as mentioned in Test 2, a second shelf was placed at patient's height +15%. Six pens were placed on the lower shelf. Patients faced the shelves (at a distance of 60 cm from the heel of the foot to the nearby edge of the lower shelf). Patients were instructed to place all six pens on the highest shelf without help or aids (e.g., 'helping hand').

Test 4: Putting on socks. A table and chair were placed 35 cm on the left and right side of the patient, respectively. Test socks in the patient's size were provided. Patients stood with the socks in one hand between the table and chair and were instructed to put on the socks without help or aids.

Test 5: Reclining and declining from a chair. A chair with a 44 cm sitting height and 78 cm back height was used. Patients were instructed to stand up and sit down three times in a row from the chair without using their hands or any other assistance. Patients started and ended the test sitting down.

Test 6: Getting up from the floor. A mat of 100*200 cm was used. A table and chair were placed to the right and left side of the mat, respectively. Patients commenced the test lying supine on the mat and were instructed to get up without help. Upon standing, the patient had to be in a position in front of the mat.

Test 7: Looking over the shoulder. A chair with a 44 cm sitting height and 78 cm back height was used. A horizontal board with numbers and characters was placed at a height of 110 cm and at a distance of 120 cm on the left or right side of the chair, respectively. The numbers and characters on the bar represented an increasing scoring system. Patients were instructed to look over their shoulder without turning their body, and read the numbers and characters as much as possible. The ability to look over the shoulder was measured by the compound movement of the rotation of the neck and the field of vision. A score between 1 and 40 points expressed the result of this compound movement. This test was performed on both sides. The lowest score was used for analysis.

Test 8: Physically demanding activities. Two pylons were placed 10m apart. Patients were provided with a heart rate monitoring device and were instructed to do the shuttle walk test. The patients had to walk continuously around the two pylons. The walking speed was gradually increased every minute (as instructed by the assessor). Patients were instructed to walk, or if necessary to run to keep up with the pace given by the assessor, as long as possible. The test was stopped if the patient's heart frequency (HF) exceeded 80% of the HF max; if the patient could not keep up with the pace as instructed by the assessor; or if the patient wanted to stop. The maximum performance time was measured in seconds.

Patients were instructed to perform all tests as quickly as possible, although in a safe manner. Before each test, the patient was uniformly instructed as to how to execute the test, including if it was to be performed one or three times. The assessor counted down from three before every test to prepare the patient for the start of the test. Patients were instructed to rest between tests if they were feeling tired. Between Tests 4 and 5, a standard break of 10 minutes was built in to minimize the effect of fatigue on performance. The ordering of the tests was identical for all patients and on both occasions. A complete testing session lasted for 50 min.

Pain and exertion experienced during each performance test were measured using a 10-cm horizontal visual analogue scale (VAS) and Borg's modified scale [29], respectively. The Borg scale is a rating of perceived exertion and was developed to describe a person's perception of exertion during exercise. For pain, the anchors were 0 cm for no pain and 10 cm for worst imaginable pain. Exertion ranged from 0 for no exertion at all to 10 for maximal exertion.

Therefore, each performance test received three scores: the performance score and a score for the pain and exertion associated with the test. Pain and exertion were recorded immediately following each test.

Statistical analyses

For the measurement of physical function performed, time (Tests 1–6 and 8) or score (Test 7) were used for analyses. For Tests 4, 5 and 6 (putting on socks, reclining and declining from a chair and getting up from the floor), the mean of the three repeated tests was used in analyses. For Test 7 (looking over the shoulder), the lowest score was used for analyses. Subject and time of measurement (Days 1 and 8) were used as sources of variance for assessing reproducibility. Intrarater reliability and agreement for measurement of performance, exertion and pain were calculated for all tests.

Reproducibility concerns the degree to which repeated measurements provide similar results [30]. Reproducibility can be divided into reliability and agreement parameters.

Reliability concerns the degree to which patients can be distinguished from each other [31]. Reliability of the measurements of performance, exertion and pain were expressed with intraclass correlation coefficients (ICCs 2.1.A) [25, 26, 30]. The ICC is the proportion of variance between observations due to between-subject variability in the true scores. The ICC has the capability to differentiate between patients, taking into account both systematic errors between two measurements and random measurement error [25]. An ICC of >0.70 is required for the comparison of groups, whereas an ICC >0.90 is recommended for individual evaluation.

Agreement concerns the degree to which scores on repeated observations correspond with each other [31]. The standard error of measurement (SEM), minimal detectable difference (MDD) [26] and Bland and Altman plots [32, 33] were used to assess agreement (i.e., measurement error) between the two sessions. The SEM is a standard for expressing the absolute error of a measurement. The SEM was assessed by taking the square root of the sum of the variation due to the measurement and the variation due to error. Using the SEM, the MDD can be computed by SEM*√2*1.96. When measurements are used on an individual level, the MDD is used for defining change. The method of Bland and Altman was used to express systematic differences between the two test occasions. Agreement parameters are expressed on the actual scale of measurement (i.e., seconds or points).

Furthermore, to assess the relationship between performed physical function and self-reported physical function, Pearson correlation coefficients between the performance tests and BASFI questionnaire were calculated. Statistical analyses were done using SPSS software, version 15.0 (SPSS, Chicago, IL, USA).

Results

The study population consisted of 67 AS patients. Data for two patients who failed to be retested were excluded from analyses. Accordingly, data for 65 patients were included in the analyses. Table 1 displays the patients' characteristics.

Table 2 lists the descriptive statistics for all measurements. The average performance measurements for Tests 3, 4 and 5, exertion for Test 3 and all measurements of pain, were slightly skewed to the right for both Days 1 and 8. All other scores showed normal distributions, without floor or ceiling effects for performance scores.

The results of reproducibility analyses are featured in Table 3. The test–retest reliability of performance measurements was adequate to excellent, with 0.73 and 0.96 the lowest and highest ICCs, respectively. The test–retest reliability for measurements of exertion showed ICCs ranging from 0.66 to 0.88. The ICCs for measurement of pain showed values between 0.69 and 0.83. For both exertion and pain, Test 7 (looking over the shoulder) was the only test for which the ICC was below the cut-off point for adequate reliability (i.e., < 0.70). Table 4 shows the SEM and MDD for all measurements and as a percentage of the total range of each score.

Table 1: Patient characteristics (*n*=65)

Characteristics	
Percentage of men	72
Age, mean ± S.D., years	48 ± 11
Disease duration, mean±S.D., years	15 ± 10
Symptom duration, mean±S.D., years	22 ± 11
Medication, %	
NSAIDs	74
Biological	14
DMARDs	17
Percentage of HLA-B27+	89
ESR, mean±S.D., mm/h	18 ± 17
Extra spinal symptoms,%	
Psoriasis	3
Uveitis	45
IBD	3
Arthritis	34
BASFI, mean ± S.D., (0-10)	4.2 ± 2.2
BASDAI, mean ± S.D., (0-10)	4.3 ± 2.4

BASDAI: Bath Ankylosing Spondylitis Disease Activity Index.

Table 2: Descriptive statistics of performance measures of physical function (n=65)

Variables	Mean ± S.D. (min–max) Day 1	Mean ± S.D. (min–max) Day 8
(1) **Climbing stairs**		
Performance[a]	5.77 ± 2.09 (3.36–13.59)	5.94 ± 2.30 (3.59–13.82)
Exertion	2.33 ± 1.84 (0–8)	2.14 ± 1.62 (0–8)
Pain	1.88 ± 2.24 (0–8.6)	1.92 ± 2.18 (0–7.9)
(2) **Bending**		
Performance[a]	18.71 ± 9.79 (9.29–62.33)	16.92 ± 7.43 (9.09–42.48)
Exertion	3.34 ± 2.11 (0–10)	3.11 ± 1.95 (0–9)
Pain	2.92 ± 2.62 (0–10)	2.61 ± 2.3 (0–8.8)
(3) **Reaching**		
Performance[a]	10.67 ± 2.31 (7.43–17.33)	10.88 ± 2.91 (6.88–20.22)
Exertion	1.56 ± 1.52 (0–6)	1.92 ± 1.64 (0–6)
Pain	1.66 ± 2.14 (0–7.9)	1.82 ± 2.19 (0–8.5)
(4) **Putting on socks**		
Performance[a,b]	19.66 ± 12.74 (6.97–63.65)	18.49 ± 14.23 (5.52–84.66)
Exertion	3.05 ± 2.21 (0–9)	3.17 ± 2.21 (0–9)
Pain	2.54 ± 2.48 (0–9.1)	2.63 ± 2.54 (0–9.07)
(5) **Reclining and declining from a chair**		
Performance[a,b]	11.3 ± 6.36 (5.44–52.44)	11.93 ± 7.8 (5.39–64.32)
Exertion	2.99 ± 2.06 (0–10)	3.16 ± 2.04 (0–10.0)
Pain	2.4 ± 2.35 (0–9.93)	2.49 ± 2.42 (0–9.5)
(6) **Getting up from the floor**		
Performance[a,b]	8.1 ± 5.95 (2.51–38.11)	8.29 ± 8.27 (2.47–63.27)
Exertion	3.98 ± 2.31 (0–10)	3.92 ± 2.33 (0–10)
Pain	3.28 ± 2.64 (0–9.93)	3.26 ± 2.81 (0–10)
(7) **Looking over the shoulder**		
Performance[a,c]	22.24 ± 6.3 (5–36)	21.92 ± 6 (4–37)
Exertion	2.98 ± 1.92 (0–9)	3.16 ± 1.79 (0–7)
Pain	2.6 ± 2.28 (0–8.9)	2.78 ± 2.54 (0–8.7)
(8) **Physically demanding activity** (i.e., shuttle-walk test)		
Performance[a]	441.72 ± 90.13 (191.11–573.89)	450.95 ± 97.35 (182.07–660.00)
Exertion	4.47 ± 2.09 (0–9)	4.49 ± 1.96 (1–10)
Pain	3.25 ± 2.75 (0–8.9)	3.49 ± 2.81 (0–8.3)

[a]Performance in seconds for tests 1-6 and 8; in points for test 7. [b]Average time. [c]Lowest score

Table 3: Intra-rater correlation coefficients for performance, exertion and pain (n=65)

	Performance	95% CI		Exertion	95% CI		Pain	95% CI	
Test	ICC	Lower	Higher	ICC	Lower	Higher	ICC	Lower	Higher
1	0.87	0.8	0.92	0.79	0.68	0.87	0.83	0.74	0.89
2	0.73	0.59	0.83	0.86	0.78	0.91	0.81	0.70	0.88
3	0.77	0.65	0.85	0.71	0.55	0.81	0.74	0.61	0.84
4	0.94	0.87	0.96	0.88	0.81	0.92	0.81	0.67	0.88
5	0.94	0.91	0.97	0.85	0.77	0.91	0.83	0.74	0.90
6	0.86	0.77	0.91	0.81	0.70	0.88	0.82	0.73	0.89
7	0.96	0.94	0.98	0.66	0.49	0.78	0.69	0.54	0.80
8	0.90	0.83	0.94	0.73	0.59	0.82	0.78	0.67	0.86

Tests are identical to those in Table 2.

Figure 1 shows the Bland and Altman plots of performance measures for each test. The difference between the first and second measurements was plotted against the mean value of both measurements. Analysis showed that only the mean of the difference of performance score of Test 2 (Bending) differed significantly from zero. A difference of 1.9 s was found between the first and second measurements (paired t-test; p=0.017). Excluding outliers only had a minor effect on the level of significance. The other plots showed no systematic differences between the two test occasions for any of the other tests. In addition, Bland and Altman plots were made for exertion and pain (data not shown). The measurement of exertion of Test 3 (reaching) initially showed a systematic difference, but one outlier caused this. For these measures, no systematic differences were found.

Except for Test 8, a significant but modest association was found for the relation between the performance measures of physical function and the BASFI questionnaire. Pearson correlation coefficients are featured in Table 5 and varied from low to moderate, with 0.29 and 0.58 being the lowest and highest, respectively.

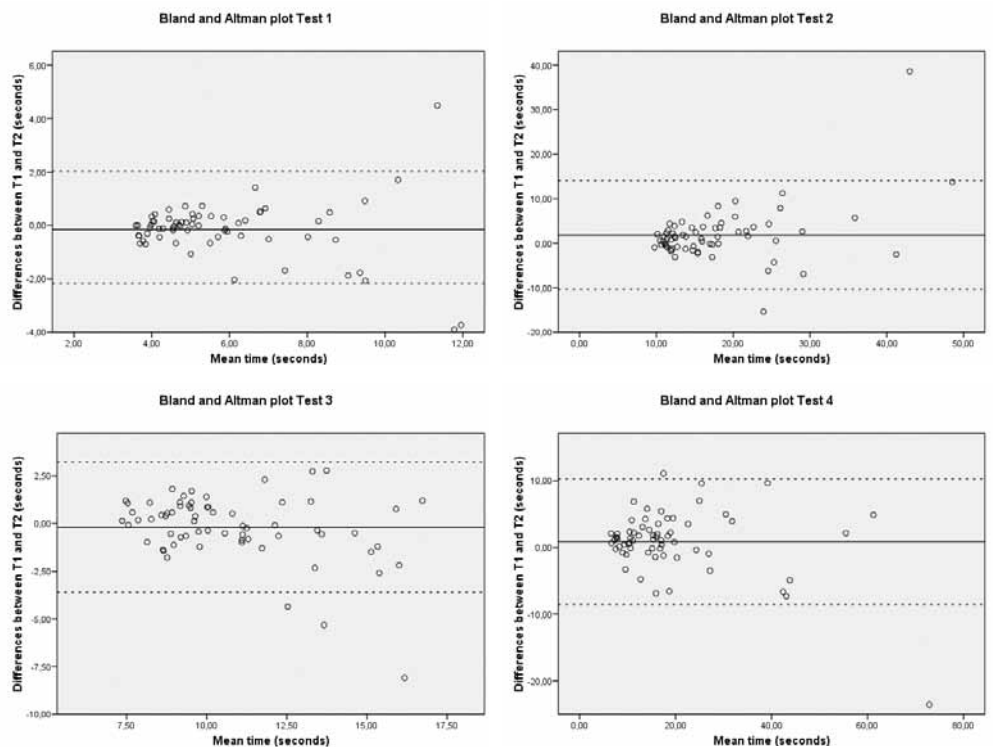

Figure 1: Bland and Altman plots for performance of all tests. Continuous line: mean difference; dashed line: 95% limits of agreement (±1.96 S.D.)

Table 4: Intra-rater agreement parameters

	Performance			
Test[a]	SEM[a]	MDD[a]	SEM % range	MDD % range
1. Climbing stairs	0.8	2.2	7.8	21.5
2. Bending	4.6	12.7	8.7	23.9
3. Reaching	1.3	3.5	7.4	20.1
4. Putting on socks	3.4	9.5	6.0	16.8
5. Reclining and declining form a chair	1.7	4.7	3.6	10.0
6. Getting up from the floor	2.7	7.6	7.6	21.3
7. Looking over the shoulder	1.2	3.4	3.9	11.0
8. Physically demanding activity	30.5	84.4	7.4	20.4

SEM and MDD expressed in percentages of the total range (n=65). [a]In seconds for test 1-6 and 8; in points for test 7.

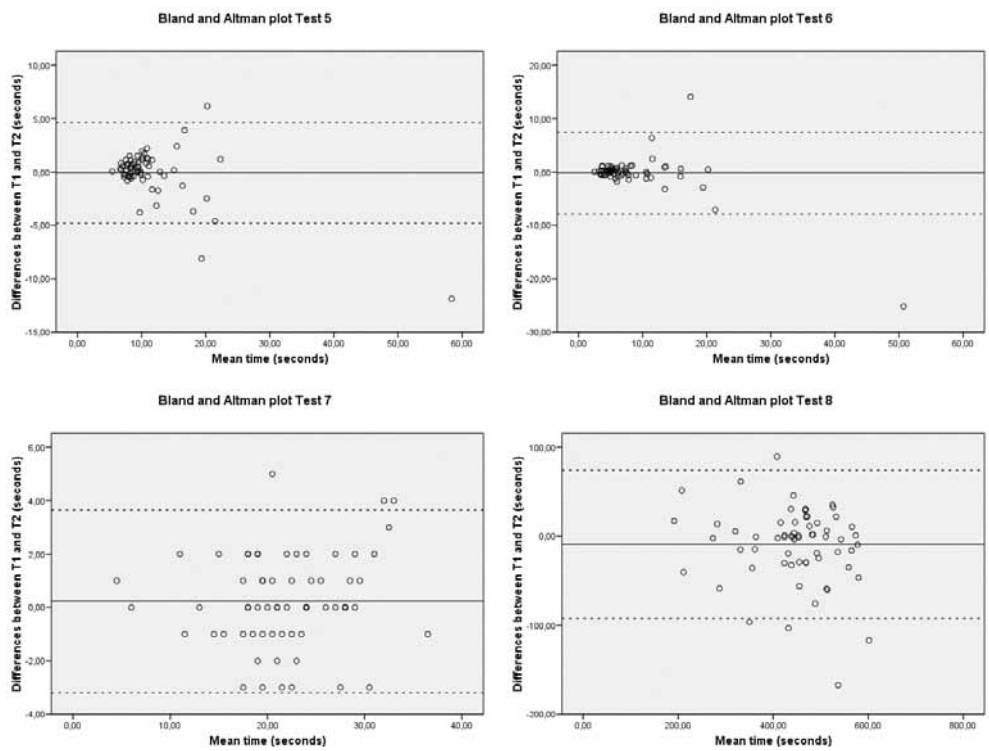

Figure 1: (continued): Bland and Altman plots for performance of all tests.

Table 4: (continued) Intra-rater agreement parameters

	Exertion				Pain			
Test	SEM	MDD	SEM % range	MDD % range	SEM	MDD	SEM % range	MDD % range
1. Climbing stairs	0.8	2.2	10.0	27.5	0.9	2.5	10.5	29.1
2. Bending	0.8	2.1	8.0	21.0	1.1	3.0	11.0	30.0
3. Reaching	0.9	2.4	15.0	40.0	1.1	3.0	13.9	38.0
4. Putting on socks	0.8	2.1	8.9	23.3	1.1	3.0	12.1	33.0
5. Reclining and declining form a chair	0.8	2.2	8.0	22.0	1.0	2.7	10.1	27.3
6. Getting up from the floor	1.0	2.8	10.0	28.0	1.1	3.1	11.1	31.3
7. Looking over the shoulder	1.0	3.0	11.1	33.3	1.3	3.7	14.6	41.6
8. Physically demanding activity	1.0	2.9	11.1	32.2	1.3	3.6	14.6	40.4

Table 5: Relationships between the BASFI questionnaire and performance measures of physical functioning (n=65)

Test	Pearson correlation coefficient
1. Climbing stairs	0.36[a]
2. Bending	0.42[a]
3. Reaching	0.46[a]
4. Putting on socks	0.44[a]
5. Reclining and declining form a chair	0.29[b]
6. Getting up from the floor	0.36[a]
7. Looking over the shoulder	0.58[a]
8. Physically demanding activity (i.e., shuttle-walk test)	0.20

Tests are identical to those in Table 2. [a] p=0.01. [b]p=0.05.

Discussion

The ICCs for intrarater reliability of performance measures based on the BASFI varied between adequate and excellent. The ICCs for intrarater reliability of measures of exertion and pain were similar, and indicated adequate to good reproducibility. One test (looking over the shoulder) showed moderate reliability for exertion and pain. The intrarater agreement parameters showed values varying from 3.6 to 8.7% of the observed range for the SEM and from 10.0 to 23.9% of the observed range for the MDD, respectively. Analyses of the intrarater agreement according to the method of Bland and Altman showed no systematic differences in measurements of performance, exertion or pain. The only exception was the performance score for Test 2 (bending).

Our results indicate that the reliability of all tests is satisfactory for group assessment (research). However, to pass a reliable judgement on an individual's level of physical functioning, information at group level is not sufficiently reliable. A higher level of reliability is desirable when evaluating individual patients (i.e., in clinical practice). When changes within individual patient scores do not exceed the MDD, such changes must be interpreted as being attributable to the measurement error [25]. Higher reliability can be achieved by repeating the measurements more often within a test protocol. Calculating an average score of the repeated measurements will decrease the absolute measurement error. Consequently, this will result in increased reliability and decrease the MDD, which is used for defining the change on an individual basis. This approach is advocated for tests with relatively high measurement error.

Similar tests for climbing stairs, bending, reaching, reclining and declining from a chair and walk tests have been used in various populations. In patients with chronic low back pain [34, 35], rheumatic disorders [36], chronic obstructive pulmonary disease [37, 38], knee OA

[39, 40] and in healthy elderly [21, 41], performance tests have been shown to have adequate reproducibility. Such data support good levels of reproducibility of performance tests in general; however, they give no information on the reproducibility of performance measures of physical function in an AS population. To our knowledge, this study was the first to report test–retest reliability and agreement of performance measures of physical function in AS patients.

The presented performance tests were derived from the BASFI. The reliability of the BASFI questionnaire has been established [16]. Auleley et al. [16] conducted an international, multicenter reliability study on the BASFI. The ICCs for the reliability of the BASFI (0–100 mm) varied between 0.88 and 0.97 [16]. Comparable levels of reliability were shown for performance measures of physical function derived from the BASFI.

However, the association between performed and self-reported physical function was shown only to be moderate to weak. This association is in concordance with the results seen in two other studies [5, 42] in which elderly and low back pain patients were observed. Our results do not correspond to those of Hidding et al. [24] in which relatively young AS patients with inflammatory back pain were observed. In their study, nine items of the DFI [13] were actually performed and compared with self-reported functional disability, which was measured using a VAS. Hidding and coworkers concluded that their AS patients and their observers agreed on the severity of functional disability. Hidding et al. [24] found a negligible difference between self-reported and observed functional disability. This variation in results indicates that further validation of the reported tests is needed.

Moreover, the moderate correlations found in this study raise issues on the convergent validity of self-reported and performance measures. The correlations suggest that self-reported functioning and actual performance are two related but distinct entities. Based on the results presented in this study, no conclusions can be drawn yet on the implications for the assessment of function in future. Two limitations of this study have to be considered. First, the test–retest interval was 1 week. Depending on the task, appropriate test–retest intervals can range from 1 h to a much longer time interval. Generally speaking, a retest interval of 2–15 days is considered to be appropriate [25]. The variability between test and retest can be explained by three factors: changes in disease activity, learning and recollection. In this study, these three factors do not seem to be of major importance.

A stable disease activity was assumed. However, this assumption might not be appropriate because patients can experience a day-to-day variability in disease symptoms. This variability would increase the error of the measurements of the physical performance. On the other hand, disease duration (mean±S.D. 15±10 years) was relatively long in our group and, therefore, these patients might show less variation in disease activity compared with the general population of AS patients. Therefore, this long disease duration might have resulted in a figure for higher reproducibility.

Although the eight tests were executed in the same order on both occasions, systematic differences in all but one score were absent. This makes it unlikely that learning effects may have played a role. Recollection is an important factor in explaining variability in reliability research of questionnaires. A recollection effect could have been present when patients

remembered their score on the Borg scale, VAS or the way they performed a test. However, it is unlikely that patients want to repeat their exact performance. Therefore, it is not likely that recollection may have influenced the results. The interval used in this study should have been long enough to exclude the effects of day-to-day variability in disease symptoms or recollection. Hence, the good levels of reproducibility reported could not be due to the short interval between the two test occasions.

A second limitation is that this study did not provide information on inter-rater reproducibility. Therefore, future research should focus on this topic.

In conclusion, performance measurements of physical function based on the BASFI questionnaire in AS patients show adequate to excellent test–retest reproducibility. These measurements are accurate for group assessment but for evaluation of individual patients, repeated measurements are advised. The inter-rater reproducibility, validity and responsiveness of performance measurements in AS are unknown and should be established in future studies.

Key messages

- Performance measures of physical function based on the BASFI questionnaire have adequate to excellent test–retest reproducibility.
- Performance measures are accurate for group assessment.
- Repeated measurements are advised for evaluation of individual patients.

References

1. Ward MM. Outcomes in ankylosing spondylitis: what makes the assessment of treatment effects in ankylosing spondylitis different? Ann Rheum Dis 2006, 65(Suppl. 3): iii25–8.
2. van Tubergen A, Landewe R, Heuft-Dorenbosch L et al. Assessment of disability with the World Health Organisation Disability Assessment Schedule II in patients with ankylosing spondylitis. Ann Rheum Dis 2003, 62: 140–5.
3. Sigl T, Cieza A, van der Heijde D, Stucki G. ICF based comparison of disease specific instruments measuring physical functional ability in ankylosing spondylitis. Ann Rheum Dis 2005, 64: 1576–81.
4. Braun J, Sieper J. Ankylosing spondylitis. Lancet 2007, 369: 1379–90.
5. Hoeymans N, Feskens EJ, van den Bos GA, Kromhout D. Measuring functional status: cross-sectional and longitudinal associations between performance and self-report (Zutphen Elderly Study 1990–1993). J Clin Epidemiol 1996, 49: 1103–10.
6. Guralnik JM, Branch LG, Cummings SR, Curb JD. Physical performance measures in aging research J Gerontol 1989, 44: M141–6.
7. Kempen GI, van Heuvelen MJ, van den Brink RH et al. Factors affecting contrasting results between self-reported and performance-based levels of physical limitation. Age Ageing 1996, 25: 458–64.
8. Sager MA, Dunham NC, Schwantes A, Mecum L, Halverson K, Harlowe D. Measurement of activities of daily living in hospitalized elderly: a comparison of self-report and performance-based methods. J Am Geriatr Soc 1992, 40: 457–62.
9. Steultjens MP, Roorda LD, Dekker J, Bijlsma JW. Responsiveness of observational and self-report methods for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 2001, 45: 56–61.
10. Stratford PW, Kennedy D, Pagura SMC, Gollish JD. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. Arthritis Rheum 2003, 49: 535–40.
11. Terwee CB, van der Slikke RMA, van Lummel RC, Benink RJ, Meijers WGH, de Vet HCW. Self-reported physical functioning was more influenced by pain than performance-based physical functioning in knee-osteoarthritis patients. J Clin Epidemiol 2006, 59: 724–31.
12. 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.
13. Dougados M, Gueguen A, Nakache JP, Nguyen M, Mery C, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1998, 15: 302–7.
14. 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 the ASAS Working Group. Assessments in Ankylosing Spondylitis. J Rheumatol 1999, 26: 951–4.
15. van der Heijde D, van der Linden S, Bellamy N, Calin A, Dougados M, Khan MA. Which domains should be included in a core set for endpoints in ankylosing spondylitis? Introduction to the ankylosing spondylitis module of OMERACT IV. J Rheumatol 1999, 26: 945–7.
16. Auleley GR, Benbouazza K, Spoorenberg A et al. Evaluation of the smallest detectable

difference in outcome or process variables in ankylosing spondylitis. Arthritis Rheum 2002, 47: 582–7.
17. Calin A, Nakache JP, Gueguen A, Zeidler H, Mielants H, Dougados M. Outcome variables in ankylosing spondylitis: evaluation of their relevance and discriminant capacity. J Rheumatol 1999, 26: 975–9.
18. 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.
19. 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.
20. Kivinen P, Sulkava R, Halonen P, Nissinen A. Self-reported and performance-based functional status and associated factors among elderly men: the Finnish cohorts of the Seven Countries Study. J Clin Epidemiol 1998, 51: 1243–52.
21. Hoeymans N, Wouters ER, Feskens EJ, van den Bos GA, Kromhout D. Reproducibility of performance-based and self-reported measures of functional status. J Gerontol A Biol Sci Med Sci 1997, 52: M363–8.
22. Elam JT, Graney MJ, Beaver T, El Derwi D, Applegate WB, Miller ST. Comparison of subjective ratings of function with observed functional ability of frail older persons. Am J Public Health 1991, 81: 1127–30.
23. Myers AM, Holliday PJ, Harvey KA, Hutchinson KS. Functional performance measures: are they superior to self-assessments? J Gerontol 1993, 48: M196–206.
24. Hidding A, van Santen M, De Klerk E et al. Comparison between self-report measures and clinical observations of functional disability in ankylosing spondylitis, rheumatoid arthritis and fibromyalgia. J Rheumatol 1994, 21: 818–23.
25. Streiner D, Norman GR. Health measurement scales. Oxford: Oxford University Press, 2003.
26. van der Esch M, Steultjens M, Ostelo RWJG, Harlaar J, Dekker J. Reproducibility of instrumented knee joint laxity measurement in healthy subjects. Rheumatology 2006, 45: 595–9.
27. 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.
28. Vollmann J, Winau R. Informed consent in human experimentation before the Nuremberg code. Br Med J 1996, 31: 1448–9.
29. Borg GA. Psychophysical bases of perceived exertion. Med Sci Sports Exerc 1982, 14: 377–81.
30. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006, 59: 1033–9.
31. Dekker J, Dallmeijer AJ, Lankhorst GJ. Clinimetrics in rehabilitation medicine: current issues in developing and applying measurement instruments 1. J Rehabil Med 2005, 37: 193–201.
32. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986, 1: 307–10.
33. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. Int J Epidemiol 1995, 24 (Suppl. 1): S7–14.
34. Smeets RJEM, Hijdra HJM, Kester ADM, Hitters MWGC, Knottnerus JA. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. Clin Rehabil 2006, 20: 989–97.
35. Simmonds MJ, Olson SL, Jones S et al.

Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine 1998, 23: 2412–21.

36. Archenholtz B, Ahlmen M, Bengtsson C et al. Reliability of articular indices and function tests in a population study of rheumatic disorders. Clin Rheumatol 1989, 8: 215–24.
37. Solway S, Brooks D, Lacasse Y, Thomas S. A qualitative systematic overview of the measurement properties of functional walk tests used in the cardiorespiratory domain. Chest 2001, 119: 256–70.
38. Skumlien S, Hagelund T, Bjortuft O, Ryg MS. A field test of functional status as performance of activities of daily living in COPD patients. Respir Med 2006, 100: 316–23.
39. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW. Get up and go test in patients with knee osteoarthritis. Arch Phys Med Rehabil 2004, 85: 284–9.
40. Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. Scand J Med Sci Sports 2001, 11: 280–6.
41. Reuben DB, Siu AL. An objective measure of physical function of elderly outpatients. The Physical Performance Test. J Am Geriatr Soc 1990, 38: 1105–12.
42. Wittink H, Rogers W, Sukiennik A, Carr DB. Physical functioning: selfreport and performance measures are related but distinct. Spine 2003, 28: 2407–13.