The Academic Medical Center Linear Disability Score Item Bank: Psychometric Properties of a New Generic Disability Measure in Rheumatoid Arthritis

NADINE WEISSCHER, CARLA A. WIJBRANDTS, ROB de HAAN, CEES A.W. GLAS, MARINUS VERMEULEN, and PAUL PETER TAK

ABSTRACT. ***Objective.*** To determine the psychometric properties of the Academic Medical Center (AMC) Linear Disability Scale (ALDS) item bank in a population of patients with rheumatoid arthritis (RA).

Methods. 129 patients with RA completed the ALDS and Health Assessment Questionnaire Disability Index (HAQ-DI) at baseline, and after 8 and 16 weeks of anti-tumor necrosis factor-α treatment. Disease activity assessments at these timepoints included serum levels of C-reactive protein, Disease Activity Score 28, morning stiffness, and visual analog scales for global disease activity and fatigue.

Results. Reliability of the ALDS was excellent (homogeneity, Cronbach's $\alpha = 0.95$; test-retest, intraclass correlation coefficient = 0.93). The ALDS results at baseline were strongly correlated with the HAQ-DI ($r = -0.75$). With regard to known group validity, both instruments discriminated between higher and lower disease activity (ALDS, $p < 0.0001$; HAQ-DI, $p = 0.002$) and between non-, moderate, and good responders (ALDS, $p = 0.002$; HAQ-DI, $p < 0.0001$), indicating that both instruments differentiate between groups. The ALDS was moderately to highly responsive to changes between baseline and after 8 weeks and 16 weeks of treatment (standardized response mean, range = 0.71–1.19). No substantial floor or ceiling effects were found.

Conclusion. Our results show that the ALDS is a promising new instrument, with at least equivalent psychometric properties compared to the HAQ-DI. Advantages of the ALDS item bank are its linear structure and an item bank that can be adapted depending on the ability level of the patient. (J Rheumatol 2007;34:1222–8)

Key Indexing Terms:
DISABILITY
OUTCOME ASSESSMENT
RHEUMATOID ARTHRITIS
LINEAR DISABILITY SCALE

The influence of a disease on the patient's level of activities of daily living (ADL) is generally considered an important outcome measure in clinical studies[1,2]. At present a lot of effort is put into the development of new measures to assess patients' daily functioning and perception of illness. Developments such as item response theory (IRT) based item banks and computer adaptive testing for the measurement of patient-reported outcomes are of increasing interest in the rapidly developing field of outcome assessment in rheumatology[3].

The Health Assessment Questionnaire Disability Index (HAQ-DI) has become the most frequently used and validated functional disability scale in rheumatology[4,5]. Although the HAQ-DI has been shown to be an effective, reliable, and valid tool that is sensitive to change, a few issues remain. The HAQ-DI is a fixed-length instrument, meaning that all the items of the scale must be administered to all patients to calculate a total score irrespective of their level of disability, which means that able patients as well as more disabled patients will be presented the same items. This impractical approach may lead to ceiling and floor effects[6-8]. Another disadvantage is that the HAQ-DI uses an ordinal instead of a linear scale, thus a similar change in HAQ-DI scores represents a different amount of change in function depending on where the patient is situated on the scale (e.g., a functional health change in HAQ-DI from 0.5 to 1.0 is not the same as a change from 2.0 to 2.5)[9].

Interest is currently moving from sum score-based methods toward the more flexible framework offered by item banks in conjunction with IRT[10]. An item bank is a collection of items, for which the measurement properties of each item

From the Department of Neurology, Division of Clinical Immunology and Rheumatology, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam; and Department of Educational Measurement, University of Twente, Enschede, The Netherlands.

N. Weisscher is supported by a grant from the Academic Medical Center — the Anton Meelmeijer Fonds. C.A. Wijbrandts is supported by a grant (no. 945-02-029) from The Netherlands Organization for Health Research and Development (ZonMw), and the Dutch Arthritis Association.

N. Weisscher, MSc, Department of Neurology; C.A. Wijbrandts, MD; P.P. Tak, MD, PhD, Division of Clinical Immunology and Rheumatology; R.J. de Haan, PhD, Department of Clinical Epidemiology and Biostatistics; M. Vermeulen, MD, PhD, Department of Neurology, Academic Medical Center, University of Amsterdam; C.A.W. Glas, PhD, Department of Educational Measurement, University of Twente.

Address reprint requests to N. Weisscher, Department of Neurology, H2-236, Academic Medical Center, 1105 AZ Amsterdam, The Netherlands. E-mail: n.weisscher@amc.uva.nl

Accepted for publication February 8, 2007.

are known[11,12]. Using psychometric methods based on IRT we developed an outcome scale that is not ordinal but linear and that identifies the full range of disability[13,14]. The Academic Medical Center Linear Disability Score (ALDS) is a generic, non-disease-specific item bank consisting of a large number of ADL items hierarchically ordered from simple to complex activities. By using a small number of items, tailored to the ADL level of patients, a sufficiently detailed clinical picture can be obtained. Even if different sets of items are used for different groups of patients, ALDS scores can still be compared within or between medical specialties.

The methodology used to develop the ALDS item bank has been analyzed in depth[14-18]. We used subsets of items from the ALDS item bank and have expanded the psychometric evaluations to patients with rheumatoid arthritis (RA).

MATERIALS AND METHODS

Patients. The patients with RA in our study were participants in a clinical trial with anti-tumor necrosis factor-α (TNF-α) therapy (infliximab). As part of this study the course of functional status in relation to clinical response to anti-TNF-α therapy was studied. The data presented here were obtained at baseline assessment and followup after 8 weeks and 16 weeks of treatment. All patients were included in the study after failing at least 2 disease modifying antirheumatic drugs (DMARD) including methotrexate (MTX), and having active disease defined as a Disease Activity Score 28 (DAS28) score ≥ 3.2[19]. Patients were taking maximal tolerable MTX treatment (5–30 mg/week), which had to be stable for at least 4 weeks prior to baseline. Oral corticosteroids (≤ 10 mg/day) and nonsteroidal antiinflammatory drugs were allowed if stable for at least 1 month prior to baseline. The study protocol was approved by the Medical Ethics Committee of the Academic Medical Center, University of Amsterdam. All patients gave written informed consent and were interviewed by trained research nurses.

Assessment. Demographic variables (age and sex) and disease status in terms of disease duration, presence of erosions, and rheumatoid factor positivity were registered at baseline. Disease activity was assessed by the DAS28 score, erythrocyte sedimentation rate (ESR), and serum levels of C-reactive protein (CRP). The DAS28 was based on the following 4 core variables: swollen joint count, tender joint count, ESR, and a 100 mm patient's global disease activity visual analog scale (VAS)[19]. Other variables assessed were the HAQ-DI[5,20], the ALDS, early morning stiffness in minutes, and VAS for fatigue. All variables were obtained at baseline and after 8 weeks and 16 weeks of anti-TNF-α therapy.

The ALDS item bank. The ALDS item bank was developed to quantify functional status in terms of the ability to perform ADL using a 2-parameter logistic IRT framework[14,21] (see Appendix 1 for methodological details). The current version of the item bank consists of 77 items, ranging from relatively easy to difficult (Appendix 2). Each patient was assessed with one of 4 increasingly difficult item sets, depending on the ability level of the patient. Beforehand the 4 different item sets were chosen by 2 of the authors (NW, RdH) and contained 14 items on average. Mildly disabled patients were presented with more difficult item sets because those were likely to provide more information. On the other hand, easier item sets lower on the scale were presented to more disabled patients. The items were administered by trained nurses and had 2 response options: "I can carry out the activity" and "I cannot carry out the activity." Participants were asked to indicate whether they could perform the activities now in the same way that they would at home or in the location where they were residing. If an activity had never before been performed by a patient (e.g., "travel by local bus or tram"), or the patient's response was "I do not know," "not applicable" was recorded. The score ranges from 0 to 100, with lower scores representing more disability.

Psychometric evaluation. IRT was used to construct the ALDS item bank, but in our study the psychometric properties of the ALDS were studied and presented in classical terms of reliability (i.e., internal consistency and test-retest reliability), validity (i.e., construct and known group), responsiveness, and the presence of ceiling and/or floor effects.

Internal consistency was assessed at the baseline measurement using Cronbach's α. The Cronbach's α coefficient[22] is based on the (weighted) average correlation of items within a scale. The criterion used for good overall internal consistency was $\alpha \geq 0.80$[23]. Test-retest reliability was assessed using intraclass correlation coefficients (ICC). Since test-retest reliability is preferably assessed in stable patients, the HAQ-DI was chosen to identify patients whose responses did not change across baseline and 8 weeks later. A delta HAQ-DI score of < 0.22 was considered stable[24,25]. ICC [and the corresponding 95% confidence interval (CI)] was calculated for the ALDS score at the 2 timepoints for the stable patients only. The threshold for this type of reliability was defined as ICC ≥ 0.70[23].

Construct validity was assessed at baseline by measuring the extent to which the ALDS correlates with a measure addressing the same concept (HAQ-DI) or measures (CRP, VAS fatigue, morning stiffness, VAS disease activity, DAS28) that reflect different aspects of health. We labeled the strength of the association: absolute values of r = 0.00–0.19 were regarded as very weak, 0.20–0.39 as weak, 0.40–0.59 as moderate, 0.60–0.79 as strong, and 0.80–1.0 as very strong correlation[26]. We assumed that for the ALDS to be valid, the ALDS scores had to correlate at least moderately with the HAQ-DI. In addition, we would expect the ALDS scores to show lower associations with the clinical measures of disease activity (DAS28, VAS disease activity, morning stiffness, and VAS fatigue) and even lower with laboratory measures of disease activity (CRP).

A scale demonstrates known group validity if it discriminates between groups of patients with known differences in clinical status. Group differences were determined by comparing ALDS scores between more or less disease activity (DAS28 dichotomized on 5.1) and non-, moderate, and good responders according to the European League Against Rheumatism (EULAR) criteria after 16 weeks of anti-TNF-α therapy.

Responsiveness was investigated by calculating the standardized response mean (SRM). An SRM value between 0.5 and 0.80 is considered moderate, and ≥ 0.80 as high responsiveness[27]. Mean ALDS values at 8 and at 16 weeks of followup were compared with the mean value at baseline. Wolfe[9] stated that the HAQ-DI cannot reliably detect differences in patients below a HAQ-DI score of 0.24. Therefore, we additionally calculated the SRM between baseline and 8 weeks of treatment for a subpopulation with a HAQ-DI score ≤ 0.24.

With regard to floor and/or ceiling effects, the percentages of patients with a maximal or minimal score at all timepoints were presented for the ALDS. The HAQ-DI was used as benchmark and therefore, with the exception of the reliability analysis, all the above described psychometric analysis (construct and known group validity, responsiveness, and ceiling/floor effects) were also done for the HAQ-DI.

Statistical analysis. Patient characteristics, outcome scores, and ceiling or floor effects were analyzed using descriptive statistics. Scores for the ALDS item bank were calculated using previously published item measures[15] and algorithms implemented in Bilog-MG version 3.0[28] and SPSS 12.0 for Windows. Patients' ability measures were estimated with maximum likelihood methods. ALDS items to which a patient did not respond or gave a response in the "not applicable" category were treated as if they had not been offered to that patient[16]. Values of Cronbach's α were obtained using a specific IRT method that allows for missing item responses[22,29] implemented in Testfact (version 4.0)[28]. Associations between the ALDS scores and the other measures were expressed in Pearson's (r) or Spearman's correlation coefficients (r_s). Differences between mean ALDS and HAQ-DI scores were analyzed using an unpaired t-test or analysis of variance (ANOVA) and Tukey's HSD post hoc analysis, when appropriate. The SRM was calculated by dividing the mean change scores by the standard deviation of the change scores. The original ALDS logits were used in all analysis, but for clarity only the linearly transformed ALDS scores (0–100) are presented.

RESULTS

Patient characteristics. One hundred twenty-nine patients with RA were enrolled between April 2001 and May 2004. The study group consisted of a predominantly female (73%) population with a mean age of 55 (range 23-85) years. On average, patients had failed treatment with 2.2 ± 1.5 DMARD before inclusion in the study. Oral corticosteroids were taken by 34 (26%) patients, mean dose 2.1 ± 3.7 mg/day. Baseline patient characteristics are summarized in Table 1.

Reliability. The internal consistency of the ALDS was good (Cronbach's $\alpha = 0.95$). Twenty-seven percent (n = 34) reported a stable disability level (delta HAQ-DI score < 0.22) and were included in the analysis of test-retest reliability. The ICC for the baseline and 8 week followup ALDS scores was 0.93 (95% CI 0.83–0.97), well above the 0.70 threshold.

Validity. Table 2 summarizes the scale scores of the measures studied at baseline assessment. Convergent validity was confirmed by a strong correlation between the ALDS and the HAQ-DI. Additionally, the ALDS showed weak to moderate associations with the clinical measures of disease activity (VAS fatigue, morning stiffness, VAS disease activity, DAS28) and a very weak association with the laboratory measure of disease activity (CRP). The correlations of the HAQ-DI with the above mentioned measures were in about the same range.

Table 3 shows the results with regard to the known group validity; both instruments discriminated between different levels of disease activity (score ≤ 5.1 and score > 5.1) according to the DAS28 (ALDS: t = 3.97, p < 0.0001; HAQ-DI: t = –3.22, p = 0.002). ANOVA showed statistically significant differences (ALDS: F = 6.56, p = 0.002; HAQ-DI: F = 16.96, p < 0.0001) between non-, moderate, and good responders, indicating that both instruments differentiate between different groups. Post hoc analysis, however, showed no significant difference in ALDS or HAQ-DI scores between moderate and good responders (ALDS: Tukey's HSD, p = 0.70; HAQ-DI: Tukey's HSD, p = 0.57).

Responsiveness. The SRM between baseline and 8 weeks of treatment indicated that the ALDS was moderately (SRM = 0.71) and the HAQ-DI (SRM = 0.88) was highly responsive. Both measures were highly responsive to change between baseline and 16 weeks of treatment (ALDS, SRM = 0.85; HAQ-DI, SRM = 1.03). Additionally, the SRM for the subpopulation with HAQ-DI score ≤ 0.24 at baseline and 8 weeks of treatment (n = 16) was 1.19 for the ALDS and 0.98 for the HAQ-DI, respectively.

Floor and ceiling effect. At all timepoints the ALDS showed no floor (0%) or substantial ceiling effect (baseline = 0.9%; 8 and 16 weeks = 0.8%). The same was true for the baseline scores of the HAQ-DI (floor = 0.8%, ceiling = 2.3%). In addition, the HAQ-DI showed no considerable floor effect at both followup measurements (8 wks: 1.6%; 16 wks: 0.8%), whereas the percentage patients scoring at the ceiling was substantially higher (8 wks: 9.3%; 16 wks: 13.2%).

Table 1. Baseline characteristics.

Characteristic	Total (n = 129)
Demographics	
Age, yrs (range)	55 (23–85)
Female, n (%)	94 (73)
Disease status	
Disease duration, mo (range)	123 (7–519)
Erosive disease, n (%)	100 (78)
Rheumatoid factor-positive, n (%)	95 (74)
DAS28, score (range)	5.7 (3.4–8.0)
ESR, mm/h (SD)	32 (± 23)
CRP, mg/l (SD)	21 (± 27)
Drug treatments	
Previous DMARD (SD)	2.2 (± 1.5)
Methotrexate, mg/wk (SD)	18.7 (± 8.3)
Corticosteroids, n (%)	34 (26)
NSAID, n (%)	63 (49)

DAS28: Disease Activity Score; ESR: erythrocyte sedimentation rate; CRP: C-reactive protein; DMARD: disease modifying antirheumatic drugs; NSAID: nonsteroidal antiinflammatory drugs.

Table 2. Descriptive statistics of the instruments and construct validity at baseline assessment.

Instrument	Score	Correlation Coefficients ALDS (n = 108) (score 76.2 ± 14.3)	Correlation Coefficients HAQ-DI (n = 122) (score 1.4 ± 0.7)
CRP*	11 (3–164)	–0.18†	0.15†
VAS fatigue	56.7 (± 21.8)	–0.39	0.35
Morning stiffness*	45 (0–1440)	–0.43	0.41
VAS disease activity	59.7 (± 21.8)	–0.45	0.40
DAS28	5.9 (± 1.1)	–0.47	0.39
HAQ-DI		–0.75	

Score values are mean (± SD); correlation values are Pearson's correlation coefficient calculated with the ALDS logits. * Due to skewed data, median and range as well as Spearman's rho correlation coefficient are presented. † All correlations are significant at the 0.01 level, except both correlations with CRP. Since higher ALDS scores indicate better functioning, signs of the coefficients are negative. ALDS: AMC Linear Disability Score; HAQ-DI: Health Assessment Questionnaire Disability Index; VAS: visual analog scale.

Table 3. Known group validity at baseline assessment.

	ALDS		HAQ-DI	
DAS28*				
≤ 5.1	83.3 (7.4)	p < 0.0001	1.1 (0.6)	p = 0.002
> 5.1	73.4 (15)		1.5 (0.7)	
EULAR criteria				
Non-responder	68.6 (18.2)	p = 0.002	1.7 (0.8)	p < 0.0001
Moderate responder	76.5 (12.2)		1.3 (0.6)	
Good responder	80.4 (12.6)	p = 0.70	1.3 (0.6)	p = 0.57

* DAS28 was dichotomized on a cutoff score of 5.1. Score distributions are presented as mean (± SD). Differences in mean logit and HAQ-DI scores are calculated using the unpaired t-test (DAS28) and ANOVA (EULAR response criteria). ALDS scores are presented after linear transformation of the original logits.

DISCUSSION

It is now widely acknowledged in the rheumatology field that assessment of patient disability should be part of the core outcome measures used in clinical trials. The HAQ was one of the first patient reported outcome measures to be used, and it has been extensively validated in a wide range of clinical trials[5,30]. Since patient reported outcomes have become increasingly important in rheumatology the development of novel outcome measures that might supersede the old are under way.

We have shown that subsets of items from the ALDS item bank used in RA can make up a reliable instrument showing good internal consistency and test-retest stability. The strong association between the ALDS and the HAQ-DI indicates that these scales largely focus on the same disability construct. The decreasing association between the ALDS scores and the clinical and laboratory measures demonstrates that the ALDS is less focused on other aspects of health, also indicating construct validity. Although no differences in ALDS scores exist between moderate and good responders to anti-TNF-α therapy, our findings that the ALDS scores differentiate between non- and moderate responders and between higher and lower level of disease activity indicate that the ALDS is sensitive to discriminate between different patient groups. The results concerning construct and known group validity are identical for the ALDS and the HAQ-DI. Compared to the HAQ-DI, the ALDS was shown to be sufficient, but less responsive between baseline and 8 weeks of treatment. Both instruments were highly responsive after 16 weeks of anti-TNF-α treatment. Moreover, the ALDS was shown to be less susceptible to ceiling effects.

The ALDS items can be used adaptively, meaning that more difficult items higher up the scale can be presented to relatively able patients, while the easier items can be presented to more disabled patients. Accordingly, it is possible to choose for a sample of items ranging from easy to difficult in case of unknown characteristics of the patient group. Items with a similar difficulty grade are interchangeable and can be used to assemble scales with items that may be of interest for different patient groups. Since the metric properties of all items have been established, ALDS scores can be compared within or between medical specialties, even when different sets of items are used. Another advantage of an IRT based item bank like the ALDS over fixed-length measures is that a linear instead of an ordinal scale is used.

The use of IRT to improve patient reported outcome received increased attention following the National Institutes of Health initiative to build a comprehensive Patient Reported Outcome Measurement System based on IRT (PROMIS, www.nihpromis.org). This initiative focuses not only on physical health but also mental health and social health. Examples of other physical-functioning item banks have been developed by Haley, *et al*[31] and Ware, *et al*[32]. However, to our knowledge item banks in use in the clinical field to date are scarce.

When validating this item bank in an RA population we were aware that a new instrument would only be interesting if it were able to rise above the qualities of the HAQ-DI, which has almost become a "gold standard." The ALDS has at least equivalent metric properties combined with attractive new and relevant features, for example, improving the clinical interpretation of scores and possibility to shorten the instrument. An advantage of the HAQ-DI is the availability of the questionnaire in several languages. The ALDS is still under development and investigation of its psychometric properties is continuing. At present versions in English and Dutch are available; German, Italian, and Spanish versions are in the process of validation.

The ALDS is an interesting new instrument that is simple to use for assessing level of disability in patients with RA. It has been suggested that a new questionnaire should not only be shorter, and better on a theoretical basis, but it must also have psychometric properties at least as good as the original HAQ-DI[33]. Our results show that use of different subsets of items from the ALDS item bank offers equivalent psychometric properties compared to the HAQ-DI. This initial study on the use of the ALDS in RA provides the basis for further testing. However, 2 advantages of the ALDS item bank compared to the original HAQ-DI are its linearity and that an IRT item bank can be used adaptively and will form a good foundation for computer adaptive testing, where the difficulty level is automatically adapted per question depending on the individual patient's ability to perform the requested activity.

ACKNOWLEDGMENT

We thank research nurses M.P. Colombijn and N. Cassin for scoring the ALDS.

REFERENCES

1. International classification of impairment, disabilities, and handicaps. Geneva: World Health Organization; 1980.
2. International classification of functioning, disability, and health. Geneva: World Health Organization; 2001. Available at: http://www3.who.int/icf/icftemplate.cfm [accessed March 23, 2007].
3. Fries JF, Bruce B, Cella D. The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. Clin Exp Rheumatol 2005;23 Suppl 39:S53-S57.
4. Bruce B, Fries J. The Stanford Health Assessment Questionnaire: dimensions and practical applications. Health Qual Life Outcomes 2003;1:20.
5. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
6. Freeman JA, Hobart JC, Thompson AJ. Does adding MS-specific items to a generic measure (the SF-36) improve measurement? Neurology 2001;57:68-74.
7. Gurka JA, Felmingham KL, Baguley IJ, Schotte DE, Crooks J, Marosszeky JE. Utility of the functional assessment measure after discharge from inpatient rehabilitation. J Head Trauma Rehabil 1999;14:247-56.
8. Hall KM, Cohen ME, Wright J, Call M, Werner P. Characteristics of the functional independence measure in traumatic spinal cord injury. Arch Phys Med Rehabil 1999;80:1471-6.
9. Wolfe F. The psychometrics of functional status questionnaires: room for improvement. J Rheumatol 2002;29:865-8.
10. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care 2000;38 Suppl:II28-II42.
11. Bode RK, Lai JS, Cella D, Heinemann AW. Issues in the development of an item bank. Arch Phys Med Rehabil 2003;84 Suppl 2:S52-S60.
12. McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. Ann Intern Med 2003;139:403-9.
13. de Haan RJ, Vermeulen M, Holman R, Lindeboom R. Measuring the functional status of patients in clinical trials using modern clinimetric methods [Dutch]. Ned Tijdschr Geneeskd 2002;146:606-11.
14. Holman R, Lindeboom R, Glas CAW, Vermeulen M, de Haan RJ. Constructing an item bank using item response theory; The AMC Linear Disability Score Project. Health Serv Outcomes Res Methodol 2003;4:19-33.
15. Holman R, Weisscher N, Glas CAW, et al. The Academic Medical Center Linear Disability Score (ALDS) item bank: item response theory analysis in a mixed patient population. Health Qual Life Outcomes 2005;3:83.
16. Holman R, Glas CAW, Lindeboom R, Zwinderman AH, de Haan RJ. Practical methods for dealing with 'not applicable' item responses in the AMC Linear Disability Score Project Health Qual Life Outcomes 2004;2:29.
17. Holman R, Lindeboom R, de Haan R. Gender and age based differential item functioning in the AMC linear disability score project. Quality of life newsletter 2004;32:1-4.
18. Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. Control Clin Trials 2003;24:390-410.
19. Prevoo ML, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LB, van Riel PL. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. Arthritis Rheum 1995;38:44-8.
20. Siegert CE, Vleming LJ, Vandenbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305-9.
21. Lindeboom R, Vermeulen M, Holman R, de Haan RJ. Activities of daily living instruments: optimizing scales for neurologic assessments. Neurology 2003;60:738-42.
22. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.
23. Nunnally J. Psychometric theory. 2nd ed. New York: McGraw-Hill; 1978.
24. Wolfe F, Michaud K, Strand V. Expanding the definition of clinical differences: from minimally clinically important differences to really important differences. Analyses in 8931 patients with rheumatoid arthritis. J Rheumatol 2005;32:583-9.
25. Wells GA, Tugwell P, Kraag GR, Baker PR, Groh J, Redelmeier DA. Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. J Rheumatol 1993;20:557-60.
26. Swinscow TDV. Correlation and regression. Statistics at square one. 9th ed. Southampton: BMJ Publishing Group; 1997.
27. Cohen J. Statisical power analysis for the behavioural sciences. New York: Academic Press; 1977.
28. du Toit ME. IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact. Lincolnwood, IL: Scientific Software International; 2003.
29. Harvey W. Estimation of variance and covariance components in the mixed model. Biometrics 1970;26:485-504.
30. 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.
31. Haley SM, Pengsheng NI, Hambleton RK, Slavin MD, Jette AM. Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. J Clin Epidemiol 2006;59:1174-82.
32. Ware J Jr, Sinclair S, Gandek B, Bjorner B. Item response theory in computer adaptive testing: implications for outcomes measurement in rehabilitation. Rehabil Psychol 2005;50:71–8.
33. Wolfe F, Michaud K, Pincus T. Development and validation of the Health Assessment Questionnaire II: a revised version of the Health Assessment Questionnaire. Arthritis Rheum 2004;50:3296-305.

APPENDIX 1. Development of the ALDS item bank.

Methodology. The items were obtained from a systematic review of generic and disease-specific functional health instruments. Each item describes an activity of daily life. During the calibration process, an incomplete, anchored calibration design was used (Holman, *et al*). Six targeted sets of items were offered to different groups in the sample. The items in common between any 2 sets of items are known as anchors. This design allows all items and patients to be calibrated on the same scale. Respondents have to rate if they could carry out the activity at present using 2 response options: "I can carry out the activity" and "I cannot carry out the activity." Participants were asked to indicate whether they could perform the activities now in the same way that they would at home or in the location where they were residing. If a patient had never experienced an activity (e.g., "travel by local bus or tram") "not applicable" was recorded.

By constructing an item bank by use of IRT, it is not essential for all respondents to be examined using all items, since IRT centers on the measurement properties of individual items rather than the instrument as a whole. Clinicians can select items from the item bank that are applicable to the population they are investigating. This means that the item bank can be used to assess patients with a wide range of conditions and levels of functional status, without placing an undue strain on the patient. Regardless of the ALDS items

used, it is possible to compare the level of functional status between patients and populations.

The original units of the ALDS scale are (logistic) regression coefficients, expressed in logits (see Appendix 2). To make the results easier to interpret, the logit scores are linearly transformed into values between 0 and 100, higher scores representing a higher level of functional status (complete item bank, psychometric properties of the individual items, and scoring rule are available from the authors on request).

Statistics. The ALDS item bank was constructed using the 2-parameter logistic item response theory model (Birnbaum). In this model, the probability, $P_{ik}(\theta_k)$, that patient k responds to item i in the category "can carry out" is

$$P_{ik}(\theta_{ik}) = \frac{\exp(\alpha_i(\theta_k + \beta_i))}{1 + \exp(\alpha_i(\theta_k + \beta_i))}$$

where θ_k represents the functional status of patient k. In addition, α_i denotes the discrimination value and β_i the difficulty value for item i. In our study, the values of α_i and β_i are assumed to be equal to the published values for a mixed patient population (Holman, *et al*).

Differential item functioning of the items was examined using the one-parameter logistic item response theory model (Rasch) by investigating if the item difficulty (β_i) was similar for male and female and for younger and older patients. The cutoff point between younger and older patients was the median age. Items were excluded from further analysis if the value was more than half the value of the standard deviation of the underlying distribution of ability value (θ).

Dimensionality of the item bank was examined using item response theory-based full information factor analysis (Bock, *et al*). An exploratory factor analysis was carried out on the 6 different item sets used. To examine the population as a whole, a confirmatory factor analysis was carried out using data from all respondents. In addition, Cronbach's alpha coefficient was calculated for each of the 6 item sets and for all the data. The statistical analysis has been described in detail (Holman, *et al*).

REFERENCES

Holman R, Weisscher N, Glas CAW, et al. The Academic Medical Center Linear Disability Score (ALDS) item bank: item response theory analysis in a mixed patient population. Health Qual Life Outcomes 2005;3:83.

Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In: Lord FM, Novick MR. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.

Rasch G. On general laws and the meaning of measurement in psychology. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of Chicago Press; 1980:321-34.

Bock RD, Gibbons RD, Muraki E. Full-information factor analysis. Appl Psychol Measurement 1988;12:261-80.

Appendix 2. The 77 ALDS items and their measurement properties.

Item	Item Content	Patients*	NA	"Can carry out"	Item Difficulty (logits)
Are you able to ...					
1	Ride a bike for at least 2 hours	30	0	18	–3.05
2	Vacuum a flight of stairs	53	3	31	–2.65
3	Carry a bag of shopping upstairs	30	0	22	–2.14
4	Clean a bathroom	53	1	42	–1.96
5	Vacuum a room and move light furniture	55	0	43	–1.88
6	Fetch groceries for 3–4 days	30	0	13	–1.63
7	Go for a walk in the woods	30	0	22	–1.50
8	Travel by local bus or tram	55	2	40	–1.23
9	Walk for more than 15 minutes	55	0	36	–0.82
10	Carry a tray	43	0	15	–0.80
11	Walk up a hill or high bridge	55	1	47	–0.78
12	Go shopping for clothes	55	0	51	–0.72
13	Cut your toenails	16	0	6	–0.66
14	Fill in an official form	5	0	3	–0.61
15	Go to a party	55	0	44	–0.56
16	Stand for 10 minutes	55	0	42	–0.53
17	Go to a restaurant	43	0	42	–0.48
18	Sweep the floor	55	1	46	–0.45
19	Hang and take in a load of washing	55	2	42	–0.44
20	Vacuum without moving any furniture	0	0	0	–0.35
21	Move a bed or table	53	0	37	–0.30
22	Use a washing machine	16	1	12	–0.23
23	Reach into a high cupboard	8	0	3	–0.23
24	Walk up a flight of stairs	8	0	2	–0.19
25	Go to the bank or post office	16	0	14	–0.13
26	Walk down a flight of stairs	8	0	5	–0.02
27	Go to the general practitioner	16	0	13	0.02
28	Use a dustpan and brush	2	0	1	0.08
29	Go for a short walk (15 min)	16	0	9	0.07
30	Write a letter	5	0	5	0.18
31	Change the sheets on a bed	55	0	31	0.21
32	Cross the road	2	0	1	0.22
33	Open and close a window	43	0	25	0.24

Appendix 2. Continued.

Item	Item Content	Patients*	NA	"Can carry out"	Item Difficulty (logits)
Are you able to ...					
34	Fetch a few things from the shop	43	0	37	0.29
35	Polish shoes	16	0	9	0.34
36	Have a shower and wash your hair	16	0	14	0.66
37	Fold up the washing	43	3	38	0.70
38	Dust	5	0	2	0.70
39	Put on/take off lace-up shoes	8	0	3	0.76
40	Clean a toilet	16	0	7	0.78
41	Make a bed	5	0	2	0.84
42	Cut your fingernails	3	0	2	0.90
43	Reach under a table	32	0	27	0.91
44	Heat tinned food	8	1	1	0.92
45	Make eggs or beans on toast	8	0	5	1.02
46	Reach into a low cupboard	2	0	0	1.09
47	Move between 2 low chairs	43	0	35	1.14
48	Pick something up from the floor	39	0	34	1.15
49	Clean a bathroom sink	16	0	12	1.18
50	Put the washing up away	8	2	6	1.26
51	Read a newspaper	5	0	5	1.28
52	Get in and out of a car	3	0	2	1.34
53	Make porridge	8	0	8	1.37
54	Clear the table after a meal	8	0	8	1.47
55	Peel and core an apple	3	0	1	1.49
56	Prepare breakfast or lunch	8	0	8	1.52
57	Clean the kitchen surfaces	8	0	7	1.76
58	Put a chair up to the table	2	0	1	1.77
59	Eat a meal at the table	3	0	2	1.79
60	Wash up	8	1	4	1.86
61	Put on/take off socks and slip on shoes	3	0	2	1.93
62	Sit up (from lying) in bed	0	0	0	1.95
63	Get a book off the shelf	1	0	0	2.11
64	Answer the telephone	0	0	0	2.15
65	Hang clothes up in a cupboard	5	0	2	2.19
66	Make coffee or tea	5	1	3	2.35
67	Put long trousers on	3	0	3	2.38
68	Make a bowl of cereal	5	0	4	2.28
69	Sit on the edge of a bed from lying down	0	0	0	2.67
70	Move between 2 dining chairs	1	0	0	2.72
71	Wash and dry your lower body	5	0	5	2.78
72	Put on/take off a coat	5	0	4	2.86
73	Wash/dry your face and hands	3	0	3	2.97
74	Get out of bed into a chair	2	0	1	2.99
75	Go to the toilet	5	0	5	3.08
76	Wash your lower body (at sink)	2	0	1	3.24
77	Put on and take off a T-shirt	4	0	3	3.49

* Number of patients presented with the item. NA: number of patients responding in the category "not applicable."