ARTHRITIS & RHEUMATISM
Vol. 50, No. 10, October 2004, pp 3296–3305
DOI 10.1002/art.20549
© 2004, American College of Rheumatology

Development and Validation of the Health Assessment Questionnaire II

A Revised Version of the Health Assessment Questionnaire

Frederick Wolfe,[1] Kaleb Michaud,[2] and Theodore Pincus[3]

***Objective.* The Health Assessment Questionnaire (HAQ) has become the most common tool for measuring functional status in rheumatology. However, the HAQ is long (34 questions, including 20 concerning activities of daily living and 14 relating to the use of aids and devices) and somewhat burdensome to score, has some floor effects, and has psychometric problems relating to linearity and confusing items. We undertook this study to develop and validate a revised version of the HAQ (the HAQ-II).**

***Methods.* Using Rasch analysis and a 31-question item bank, including 20 HAQ items, the 10-item HAQ-II was developed. Five original items from the HAQ were retained. We studied the HAQ-II in 14,038 patients with rheumatic disease over a 2-year period to determine its validity and reliability.**

***Results.* The HAQ-II was reliable (reliability of 0.88, compared with 0.83 for the HAQ), measured disability over a longer scale than the HAQ, and had no nonfitting items and no gaps. Compared with the HAQ, modified HAQ, and Medical Outcomes Study Short Form 36 physical function scale, the HAQ-II was as well correlated or better correlated with clinical and outcome variables. The HAQ-II performed as well as the HAQ in a clinical trial and in prediction of mortality and work disability. The mean difference between the HAQ and HAQ-II scores was 0.02 units.**

***Conclusion.* The HAQ-II is a reliable and valid 10-item questionnaire that performs at least as well as the HAQ and is simpler to administer and score. Conversion from HAQ to HAQ-II and from HAQ-II to HAQ for research purposes is simple and reliable. The HAQ-II can be used in all places where the HAQ is now used, and it may prove to be easier to use in the clinic.**

The Health Assessment Questionnaire (HAQ) is the most important and widely used functional status questionnaire in rheumatology. Developed by Fries et al in 1980 (1,2), it is used in most clinical trials and observational outcome studies (3), and it has been translated into most languages in the industrialized countries (3,4).

The HAQ is the best predictor of mortality (5), work disability (6), joint replacement (7), and medical costs (8). It is effective in rheumatoid arthritis (RA), osteoarthritis (OA), and other rheumatic conditions. The US Food and Drug Administration accepts it as a measure for evaluation of the prevention of disability.

Despite its extraordinary success, there are reasons to consider its revision (9–12). The HAQ is long. It is composed of 20 questions concerning activities of daily living (ADLs) and 14 questions relating to the use of aids and devices. In addition, its scoring is not simple. Subsequently, a modified HAQ (M-HAQ) with 8 ADLs was developed to address the length and scoring problems (13). A further modification, the multidimensional HAQ (MD-HAQ), added more complex ADLs (14). Like the HAQ, the M-HAQ predicts important long-term outcomes (15–17).

The HAQ also has something of a "floor" problem, in that many persons with physical disability can

The National Data Bank for Rheumatic Diseases has received grant support from Amgen, Aventis, Bristol-Myers Squibb, Centocor, Pfizer, and Pharmacia pharmaceutical companies.

[1]Frederick Wolfe, MD: Arthritis Research Center Foundation, Wichita, Kansas, and University of Kansas School of Medicine, Wichita; [2]Kaleb Michaud, MS: Arthritis Research Center Foundation, Wichita, Kansas; [3]Theodore Pincus, MD: Vanderbilt University, Nashville, Tennessee.

Address correspondence to Frederick Wolfe, MD, National Data Bank for Rheumatic Diseases, 1035 North Emporia Avenue, Suite 230, Wichita, KS 67214. E-mail: fwolfe@arthritis-research.org.

Submitted for publication November 3, 2003; accepted in revised form June 22, 2004.

have normal HAQ scores. In addition, the HAQ is not a linear scale; a 0.25 difference at one level of disability (e.g., a HAQ score of 0.50) may not mean the same as that at another level (e.g., a HAQ score of 1.75) (18). Previous analyses have also suggested that some of the individual questions are not being answered correctly or are being misunderstood by patients (9).

Given the track record of the HAQ and its modified versions, the development of a new version should not be undertaken lightly. A new questionnaire should not only be shorter, and better on a theoretical basis, but it must also be shown to be at least as good as the original HAQ in terms of construct validity, discriminant validity, predictive validity, and reliability. In addition, it must have mean scores that are similar to those produced by the HAQ so that there can be interconversion of the questionnaires. In this report, we describe validation studies of a revised HAQ, the HAQ-II, that was developed using an item bank and Rasch analysis, an item response theory model for measurement (11,19–28).

PATIENTS AND METHODS

Development of the HAQ-II. For background in understanding the development of the HAQ-II that preceded this report, we present the following information. In January 2001, the National Data Bank for Rheumatic Diseases (NDB) mailed surveys to participants in its long-term outcomes studies, as previously described (8,29). In addition to the standard HAQ that was mailed to all participants, one-third of participants received 1 of 3 test questionnaires. All of the test questionnaires contained 31 test questions and none of the questions on aids and devices that are present as modifiers for the HAQ. The questions were as follows:

Are you able to:
Walk a mile?
Walk 2 or more miles?
Go up a flight of stairs?
Go up 2 or more flights of stairs?
Open a previously unopened jar?
Vacuum in the house?
Do outside work (such as yard work)?
Wait in a line for 15 minutes?
Lift heavy objects?
Move heavy objects?
Change the bedding?
Dress yourself, including shoelaces and buttons?
Shampoo your hair?
Stand up from a straight chair?
Get in and out of bed?
Cut your meat?
Lift a full cup or glass to your mouth?
Open a new milk carton?
Walk outdoors on flat ground?
Climb up 5 steps?
Wash and dry your body?
Take a tub bath?
Get on and off the toilet?
Reach and get down a 5-pound object (such as a bag of sugar) from just above your head?
Bend down to pick up clothing from the floor?
Run errands and shop?
Get in and out of a car?
Do chores such as vacuuming or yard work?
Open car doors?
Open jars which have been previously opened?
Turn faucets on and off?

The first test questionnaire gave 4 choices: 1) without any difficulty, 2) with some difficulty, 3) with much difficulty, and 4) unable to do. The second test questionnaire changed the wording slightly to 1) with no difficulty, 2) with a little difficulty, 3) with a lot of difficulty, and 4) unable to do. The third test questionnaire used only 3 categories headed by these instructions: "The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?" The categories were 1) yes, limited a little; 2) yes, limited a lot; and 3) no, not limited at all.

The 31 questions, which included 20 from the HAQ, were used as a test item bank. The first two formats were chosen to see whether discrimination between "much difficulty" and "unable" could be improved by changing wording, since previous analyses had shown problems in the discrimination between these response levels. The third format sought to solve the discrimination problem by eliminating one of the categories.

Selection of HAQ-II items. The goal of the questionnaire development was to obtain a reliable, statistically valid, unidimensional scale that captured as much of the disability continuum as possible. Using Rasch analyses, we used an iterative procedure which balanced 4 concerns: removal of misfitting items, maximizing scale length, elimination of items with overlapping difficulties, and removal of gaps along the disability–difficulty continuum. The alternative wordings did not improve the psychometric properties of the potential questionnaires, and they were discarded. The 10-item HAQ-II questionnaire (Table 1) that emerged from the analyses of the 31 items best balanced the concerns of item fit, scale length, and evenly spaced items. The HAQ-II contains 5 of the original HAQ questions and 5 new questions.

In addition, we examined the existing HAQ questionnaires for scale length, reliability, misfitting items, and gaps in the scale. These analyses included 2,229 patients completing the HAQ, 7,289 completing the M-HAQ, 8,065 completing the MD-HAQ, and 2,374 completing the final version of the HAQ-II (Tables 1 and 2).

In the Rasch model of disability, functional ability is considered to lie upon a linear "ruler," similar to an ordinary ruler, where no disability is the anchor at one end and maximum disability is the anchor at the other end. The range of disability is expressed in logits, a completely linear measure (Table 2). The longer the scale length is in logits, the better the scale is in representing disability. As suggested above, the scale is anchored at one end by the task that is easiest to do and at the other end by the task that is most difficult to do (Table 1, thresholds). An item (question) difficulty (threshold) represents the position in logits that the item occupies on the linear disability scale. Since each task is composed of 4 levels (0, 1, 2,

Table 1. Wording of the revised Health Assessment Questionnaire (HAQ-II) and its category percentages and item thresholds in 2,374 patients with rheumatoid arthritis*

	Category description (score)				
	Without any difficulty (0)	With some difficulty (1)	With much difficulty (2)	Unable (3)	Item threshold†
We are interested in learning how your illness affects your ability to function in daily life. Place an X in the box which best describes your usual abilities *over the past week.* Are you able to:					
Get on and off the toilet?	64.0	31.8	4.0	0.2	2.254
Open car doors?	60.1	35.0	4.3	0.6	1.936
Stand up from a straight chair?	47.8	45.3	5.9	1.1	1.466
Walk outdoors on flat ground?	52.6	37.2	8.6	1.6	1.271
Wait in a line for 15 minutes?	37.1	41.5	16.5	5.0	0.244
Reach and get down a 5-pound object (such as a bag of sugar) from just above your head?	32.1	45.3	12.7	9.9	−0.241
Go up 2 or more flights of stairs?	19.6	39.8	29.8	10.9	−0.876
Do outside work (such as yard work)?	13.3	45.6	22.0	19.1	−1.452
Lift heavy objects?	5.2	40.1	30.1	24.7	−2.172
Move heavy objects?	3.4	39.6	31.1	25.9	−2.430

* Except where indicated otherwise, values are percentages of patients selecting category 0, 1, 2, or 3 for each item. The score of the questionnaire is the sum of the individual item scores divided by 10 or the mean of the item scores if 8 or 9 items are completed. The HAQ-II is not to be scored if fewer than 8 items are completed.

† Item thresholds are derived from Rasch analysis. The more negative the threshold, the more difficult it is to perform the task. See Patients and Methods for additional details.

3), a 10-item questionnaire actually addresses 40 levels of difficulty. In the Rasch analyses, however, we used only 30 levels of difficulty, with each level of an item (HAQ question) representing the point or threshold where the probability is 0.5 of being in 0 as opposed to 1, 1 as opposed to 2, and 2 as opposed to 3. In a "perfect" scale, the 30 thresholds lie equidistant from each other on the disability continuum. In the current analyses, we define gaps in the disability continuum scale as spaces between item threshold locations that are ≥1 logit in length. Duplicated thresholds are defined as threshold locations occupied by ≥2 item thresholds. The presence of gaps, duplicated locations, and nonlinear spacing of thresholds correlates with decreased precision in the assessment of disability.

The overall reliability of a scale can be estimated by examining the person separation statistics and the person model reliability. Separation is the measure of spread in the test sample expressed in units of the test error (30). Reliability is the ratio of the true measure variance to the observed measure variance and is the same as Cronbach's alpha (31). Reliabilities of ≥0.85 are satisfactory. Rasch analysis produces two additional statistics. The mean square INFIT (INFIT) and mean square OUTFIT (OUTFIT) statistics are measures of "signal to noise" that allow one to determine how well an item or an individual level (0, 1, 2, 3) of an item fits the Rasch model (32). An item that has a high INFIT or OUTFIT statistic (>1.3) may not fit the model because it is "noisy" (not understood well or ambiguous) or is measuring a second

Table 2. Characteristics of versions of the Health Assessment Questionnaire (HAQ)*

Scale (no. of patients analyzed)	Separation	Reliability	Misfitting items†	INFIT	OUTFIT	Reversed thresholds	Scale length, logits	Duplicated thresholds, %‡	Item gaps§
HAQ (2,229)	2.33	0.83	Hygiene	1.30	1.54	Hygiene, dressing¶	7.2	25	1
M-HAQ (7,289)	2.05	0.81	Turn faucets	1.21	1.34		7.1	29	1
MD-HAQ (8,065)	2.41	0.85	Participate in sports	1.29	1.76	Participate in sports; walk 2 miles	9.3	30	2
			Walk 2 miles	1.25	1.39				
			Turn faucets	1.20	1.34				
HAQ-II (2,374)	2.75	0.88		None	None		10.0	17	0

* M-HAQ = modified HAQ; MD-HAQ = multidimensional HAQ; HAQ-II = revision of HAQ tested in the present study (see Patients and Methods for definitions of column headings not given below).

† Those with INFIT or OUTFIT statistic >1.3. The maximum misfit statistic for the HAQ-II was 1.17.

‡ Threshold locations occupied by ≥2 item thresholds. Values refer to the number of duplications divided by the total number of item thresholds.

§ Spaces between item threshold locations that are ≥1 logit in length.

¶ Reversed threshold for hygiene was due to the items "shampoo hair" and "take a tub bath." HAQ refers to categories rather than items.

dimension. For example, if a person was asked to evaluate his/her ability to perform tasks with which he/she had little current experience, the replies might be inaccurate measures of actual ability and would be identified as being "noisy" by the Rasch model. A further indication of this problem is often seen with reversed thresholds. A reversed threshold occurs in these analyses when being "unable" to do an activity (scored as "3" on the HAQ/HAQ-II) appears to be easier to do than doing the activity with "much difficulty" (scored as "2").

Although the HAQ has 20 items, these collapse into 8 categories after scoring. In the analyses shown in Table 2, we report results at the item level as well as at the category level.

Validation studies. After development of the HAQ-II, the new questionnaire together with the HAQ was used in 4 consecutive biannual surveys mailed to participants in the NDB. The data from these assessments were then used in the validation studies that form the basis of this report. In addition to the HAQ and HAQ-II data, the NDB collects further data in its biannual, detailed 28-page questionnaire. At each assessment, demographic variables are recorded, including age, sex, ethnic origin, education level, current marital status, medical history, work status, and total family income. Data concerning disease status and activity variables were collected using the following instruments: the M-HAQ (13); visual analog pain, global disease severity, and fatigue scales (33); the Arthritis Impact Measurement Scales anxiety and depression scales (34,35); the Rheumatology Distress Index; the Rheumatoid Arthritis Disease Activity Index (36–38); and the Work Limitations Questionnaire, a 25-item, self-administered questionnaire measuring the degree to which health problems interfere with ability to perform job roles (39). Patients also completed the Medical Outcomes Study Short Form 36 (SF-36), from which the physical function (PF) scale was calculated (40,41). Utilities were measured using the EuroQol (42–44) and Short Form 6D (45). Analyses based on the above data were restricted to the 14,038 persons who completed all the HAQ, HAQ-II, M-HAQ, and PF scales.

A second set of data was available from 693 consecutive patients who were identified in 2003 from the clinical practices of 40 US and Canadian rheumatologists, the Rheumatoid Arthritis Evaluation Study (RAES) cohort. This data set was used to examine the correlation between the HAQ and HAQ-II and physical examination findings and laboratory and Disease Activity Score values (46).

The MD-HAQ (14) was not a part of the biannual NDB assessments and therefore was not included in the validation report. However, separate data from the NDB, where the MD-HAQ was collected as part of screening evaluations, were available for 15,543 patients. These data are presented briefly to describe the floor effects of this questionnaire.

The HAQ and HAQ-II questionnaires were also used simultaneously in an open-label clinical trial of 837 RA patients in community practice starting a new disease-modifying antirheumatic drug (DMARD). Pretest and posttest data from this study were analyzed after 3 months of therapy.

Statistical analysis. Rasch analysis was performed using Winsteps version 3.31 (Winsteps, Chicago, IL) (47) and RUMM 2010 version 3.3 (RUMM Laboratory, Duncraig, Australia) (48). Validation analyses, including correlation analysis, generalized estimating equations (GEE), and linear and Cox regression analysis, were performed using Stata version 8.1 (Stata Corporation, College Station, TX) (49). Comparison of correlation coefficients was made using the Fisher z transformation. The *t*-test was used to compare the HAQ and the HAQ-II in the clinical trial. The Bland-Altman limits of agreement procedure was used to assess the 2-SD difference between questionnaires administered at the same time to the same patients (50).

RESULTS

Questionnaire analysis. Rasch analysis was used to categorize the 4 HAQ questionnaires (Tables 1 and 2). The most difficult task was "move heavy objects," which had an item threshold of –2.430 (Table 1). At the other end of the spectrum, "get on/off toilet" was the easiest task to perform (threshold 2.254). Other items held intermediate positions. The differences among items in regard to their difficulty can also be seen in the percentages of patients selecting each category of a given item. These percentage results paralleled the Rasch item thresholds. The HAQ-II had the longest scale (Table 2), as measured in logits, indicating that it captured more of the continuum of disability than did the other questionnaires. The MD-HAQ also had a long scale, by virtue of the difficult items "participate in sports and games" and "walk 2 miles." However, these items misfit the Rasch model, indicating a lack of unidimensionality and/or inaccurate assessment. The HAQ also had items that did not fit the Rasch model. Within the HAQ hygiene category, the items "take a tub bath" and "shampoo hair" misfit the model. This, in turn, led to the misfitting of the hygiene category. We also noted gaps in the scales of all the HAQ family questionnaires except for the HAQ-II. Duplicated thresholds were least common in the HAQ-II. These data indicate that the HAQ-II had the most favorable psychometric characteristics as measured by reliability, fit, scale length, reversed thresholds, and item gaps.

The floor (scores of 0) and ceiling (scores of 3) effects and the mean scores for the HAQ questionnaires that were studied in the validation sample are shown in Table 3. Data from the SF-36 PF scale are included for

Table 3. Mean scores and floor and ceiling effects for validation study questionnaires (n = 14,038 patients)*

Scale	Score, mean ± SD	Patients at floor, %	Patients at ceiling, %
SF-36 PF	47.08 ± 28.4	3.4	3.0
HAQ-II	1.07 ± 0.66	5.8	0.1
HAQ	1.09 ± 0.72	10.1	0.2
M-HAQ	0.51 ± 0.49	24.5	0.2

* SF-36 PF = Medical Outcomes Study Short Form 36 physical function scale (see Table 2 for other definitions).

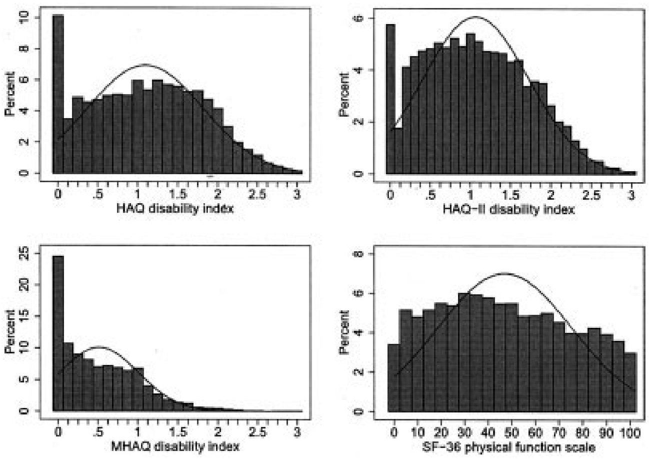

Figure 1. Distribution characteristics of 4 functional status questionnaires in 14,038 patients with rheumatic disease. Floor effect is noted by the percentage of values at 0 for each questionnaire. Ceiling effect for the Medical Outcomes Study Short Form 36 (SF-36) physical function scale is indicated by the percentage of values at 100. Curved lines are superimposed normal distribution curves. HAQ = Health Assessment Questionnaire; HAQ-II = revision of HAQ tested in the present study; M-HAQ = modified HAQ.

comparison. The HAQ-II had the least floor effect of the HAQ family of questionnaires (5.8%), as would be expected from its long logit length. The HAQ had a greater floor effect (10.1%), and the M-HAQ had the greatest floor effect (24.5%). Although data on the MD-HAQ were not available in the validation sample, 15,543 screening questionnaires using the MD-HAQ were available in the NDB. In this sample, 4.4% of patients were at the floor in the MD-HAQ.

Results of validation studies. *Diagnostic groups*. There were 14,038 persons who completed all of the HAQ, HAQ-II, M-HAQ, and PF scales. Of these, 10,916 (77.8%) had RA, 2,478 (17.7%) had OA, and 644 (4.6%) had fibromyalgia.

Distribution characteristics. The disability indexes differed in their distributions (Figure 1). The HAQ and HAQ-II tended toward normal distributions except at the floor (lowest levels). The PF scale was not normally distributed, with more values at the tails than might be expected. The M-HAQ floor effect was profound and contributed to its non-normality.

As noted above, among the HAQ family of questionnaires, the floor effect was least for the HAQ-II (5.8%) and greatest for the M-HAQ (24.5%) (Table 3). The PF scale had combined ceiling and floor effects of 6.4%. Therefore, compared with the combined floor and ceiling effects of the HAQ-II (with combined effects of 5.9%), the PF represented a shifting of the distribution curve to the right, compared with the HAQ-II scale.

Correlates of functional status questionnaires. The HAQ-II results were correlated with clinical and outcomes variables at levels similar to those of the HAQ, M-HAQ, and PF scale (Tables 4 and 5). For the

Table 4. Correlations between results of functional status questionnaires and clinical and outcome variables in the National Data Bank for Rheumatic Diseases Sample (n = 14,038 patients)*

Variable	HAQ-II	HAQ	M-HAQ	SF-36 PF scale
HAQ-II (0–3 scale)	1.00	0.91	0.84	−0.85
HAQ (0–3 scale)	0.91	1.00	0.86	−0.80
SF-36 PF scale (0–100)	−0.85	−0.80	−0.72	1.00
M-HAQ (0–3 scale)	0.84	0.86	1.00	−0.72
EuroQol utility (0–1 scale)	−0.67	−0.64	−0.69	0.62
RADAI score (0–10)	0.65	0.63	0.66	−0.61
Rheumatology Distress Index (0–100 scale)	0.61	0.59	0.61	−0.58
Global disease severity (0–10 VAS)	0.61	0.58	0.59	−0.59
Pain (0–10 VAS)	0.61	0.59	0.61	−0.57
Fatigue (0–10 VAS)	0.56	0.54	0.52	−0.53
SF-6D utility (0–1 scale)	−0.56	−0.54	−0.48	0.60
Work Limitations Questionnaire index (0–100 scale)	0.56	0.54	0.55	−0.53
QOL scale (0–100 VAS)	−0.54	−0.51	−0.52	0.53
AIMS depression scale (0–10)	0.44	0.42	0.47	−0.42
Sleep disturbance (0–10 scale)	0.41	0.40	0.42	−0.38
AIMS anxiety scale (0–10)	0.38	0.36	0.41	−0.36
Social security disability, last 6 months (%)	0.34	0.32	0.34	−0.30
GI severity (0–10 scale)	0.33	0.31	0.34	−0.30
Total direct medical costs, $	0.24	0.23	0.20	−0.22
Total joint replacement, %	0.18	0.20	0.13	−0.18

* SF-36 PF = Medical Outcomes Study Short Form 36 physical function; RADAI = Rheumatoid Arthritis Disease Activity Index; VAS = visual analog scale; SF-6D = Short Form 6D; QOL = quality of life; AIMS = Arthritis Impact Measurement Scales; GI = gastrointestinal (see Table 2 for other definitions).

Table 5. Correlations between results of functional status questionnaires and clinical practice variables from the Rheumatoid Arthritis Evaluation Study (n = 693 patients)*

Variable	HAQ-II	HAQ	M-HAQ
HAQ-II (0–3 scale)	1.00	0.92	0.85
HAQ (0–3 scale)	0.92	1.00	0.84
M-HAQ (0–3 scale)	0.85	0.84	1.00
Pain (0–10 VAS)	0.66	0.66	0.67
Patient's assessment of global disease severity (0–10 VAS)	0.62	0.60	0.61
Fatigue (0–10 VAS)	0.57	0.56	0.55
DAS28	0.51	0.54	0.50
Physician's assessment of global disease severity (0–10 VAS)	0.48	0.50	0.50
Disability (stopped work)	0.41	0.42	0.35
Tender joint count (range 0–28)	0.37	0.39	0.40
ESR, mm/hour	0.25	0.27	0.22
Swollen joint count (range 0–28)	0.24	0.27	0.25
Joint surgery, no/yes	0.20	0.23	0.11

* VAS = visual analog scale; DAS28 = Disease Activity Score in 28 joints; ESR = erythrocyte sedimentation rate (see Table 2 for other definitions).

14,038-patient NDB data set (Table 4), HAQ-II correlations were greater than HAQ correlations in 15 of 16 instances. Compared with M-HAQ correlations, HAQ-II correlations were greater for 7 variables and less for 6 variables. Compared with PF scale correlations, HAQ-II correlations were greater for 14 variables, less for 1 variable, and equal for 1 variable. Although these differences achieved statistical significance owing to the large sample size, they were clinically insignificant. The results of the analyses should be considered to show no important difference between the questionnaires. Correlation levels were similar in a data set of serial patients from clinical practice—the RAES cohort (Table 5)—but with the smaller sample size (n = 693), there were no significant differences in the correlations among the questionnaires at the 0.05 probability level.

Correlates of functional status questionnaires for different diagnostic groups. Correlations between questionnaire results and clinical and outcome variables were not significantly different for the different diagnostic groups ($P > 0.05$) (Table 6).

Clinical trial results: *comparison of HAQ and HAQ-II*. To assess the ability of the HAQ and HAQ-II to perform in a clinical trial setting, 837 RA patients who received a DMARD over a 3-month period in an open-label clinical trial were studied. At the start of therapy, the HAQ score was 1.50 and the HAQ-II score was 1.41. Effect sizes were calculated for the before–after difference for the HAQ and HAQ-II. The effect size for the HAQ-II was 23.0 (95% confidence interval [95% CI] 18.4–27.4). The effect size for the HAQ was 24.8 (95% CI 20.0–29.5). These differences were not significant ($P = 0.298$).

Table 6. Correlations between results of functional status questionnaires and clinical and outcome variables according to diagnostic group*

	Patients		
Variable	RA (n = 10,916)	OA (n = 2,478)	Fibromyalgia (n = 644)
HAQ-II (0–3 scale)	1.00	1.00	1.00
HAQ (0–3 scale)	0.91	0.89	0.89
SF-36 PF scale (0–100)	−0.86	−0.84	0.82
M-HAQ (0–3 scale)	0.85	0.82	0.85
EuroQol utility (0–1 scale)	−0.68	−0.66	0.64
RADAI score (0–10)	0.66	0.65	0.64
Rheumatology Distress Index (0–100 scale)	0.62	0.61	0.56
Pain (0–10 VAS)	0.62	0.59	0.57
Global disease severity (0–10 VAS)	0.61	0.60	0.58
SF-6D utility (0–1 scale)	−0.57	−0.56	0.42
Fatigue (0–10 VAS)	0.57	0.58	0.48
Work Limitations Questionnaire index (0–100 scale)	0.55	0.55	0.49
QOL scale (0–100 VAS)	−0.54	−0.55	0.49
AIMS depression scale (0–10)	0.45	0.44	0.41
Sleep disturbance (0–10 scale)	0.42	0.40	0.35
AIMS anxiety scale (0–10)	0.39	0.38	0.35
Social Security disability, last 6 months (%)	0.34	0.30	0.38
GI severity (0–10 scale)	0.33	0.35	0.33
Total direct medical costs, $	0.24	0.26	0.28
Total joint replacement, %	0.20	0.11	0.05

* RA = rheumatoid arthritis; OA = osteoarthritis (see Tables 2 and 4 for other definitions).

Change in HAQ and HAQ-II scores over time. Using a population-averaged model (GEE) restricted to RA patients who had completed both questionnaires (n = 10,494), the change in HAQ score per year of disease duration was 0.014 (95% CI 0.013–0.016, Wald χ^2 = 519.1). The equivalent value for the HAQ-II score was 0.012 (95% CI 0.011–0.123, Wald χ^2 = 423.7).

Predictive ability: *mortality*. For 10,281 persons who completed more than one NDB questionnaire, Cox regression was used to estimate the ability of the HAQ and HAQ-II to predict future mortality. The hazard ratio (HR) for the HAQ was 2.28 (95% CI 1.89–2.75, likelihood ratio χ^2 = 77.2); the HR for the HAQ-II was 2.44 (95% CI 2.00–2.97, likelihood ratio χ^2 = 80.11).

Predictive ability: *Social Security disability awards*. For 6,472 patients age <65 years who were not receiving US Social Security benefits at their first assessments, Cox regression was used to estimate the ability of the HAQ and HAQ-II to predict future Social Security benefits. The HR for the HAQ was 5.47 (95% CI 4.72–6.35, likelihood ratio χ^2 = 552.6); the HR for the HAQ-II was 6.06 (95% CI 5.19–7.07, likelihood ratio χ^2 = 549.7).

Conversion of HAQ and HAQ-II scales. To understand how the HAQ-II might be substituted for the HAQ, as well as the reverse condition, we first graphed the relationship between the two variables using locally weighted scatterplot smoothing (lowess) regression and linear regression (Figure 2). Lowess regression will demonstrate the nonlinear aspects of the relationship between variables. Since the relationship shown in Figure 2 was essentially linear, we performed linear regression and described the relationship between variables by the regression intercept and coefficient. Based on the regression analyses of 14,038 observations, HAQ-II = 0.158 + 0.83 × HAQ and HAQ = 0.39 + 0.989 × HAQ-II (R^2 = 0.821). The M-HAQ and the PF scale differ too much from the HAQ and HAQ-II for useful conversions and are not described.

The strong relationship between the HAQ and the HAQ-II allows reliable interconversion of research data from the HAQ to the HAQ-II and from the HAQ-II to the HAQ, although this should be confined to adjustment of means and, perhaps, to distribution-independent analyses such as median regression. However, individual patient data cannot be converted, since the level of agreement, even with a correlation coefficient >0.9, is not high enough. Although the difference between the HAQ and HAQ-II mean scores was only 0.02 units, and Lin's concordance correlation was 0.902, the Bland-Altman 95% limits of agreement values were –0.567 and 0.622. The distance between these values is too great for substitution of one functional measure for another in an individual patient.

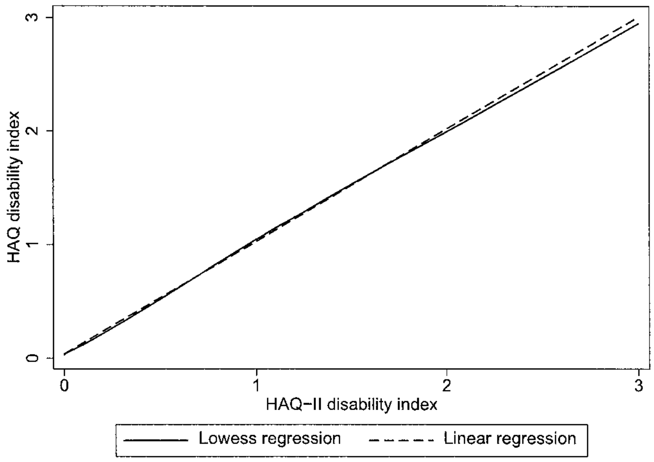

Figure 2. Regression of HAQ on HAQ-II in 14,038 patients with rheumatic disease. Locally weighted scatterplot smoothing (lowess) regression indicates graphically the nonlinear aspects of the HAQ and HAQ-II relationship compared with linear regression. The lines are virtually superimposable, indicating validity of a linear predictive model. See Figure 1 for other definitions.

DISCUSSION

The validation results of this study suggest that the HAQ-II performs at least as well as the original HAQ. This should not be surprising, since 5 of the 10 HAQ-II items come directly from the HAQ. In addition, poorly fitting items of the HAQ were removed, and the overall item content of the HAQ-II was selected with careful attention to psychometric properties using Rasch analysis. Although the HAQ has 20 items (plus 14 aids and device modifiers), the method of scoring the HAQ reduces the questionnaire to 8 categories. In effect, the HAQ is an 8-item questionnaire, but one that gets some additional reliability from the redundancy of multiple questions in each category. Given the (de facto) 8-item HAQ and the 10-item HAQ-II, the HAQ-II, all things being equal, should perform as well as or better than the original HAQ.

The HAQ-II was developed using Rasch analysis and an item bank of questions in which each question has an intrinsic and measurable difficulty. For example, it is easier to walk on flat ground or get up from a chair than it is to walk up 2 flights of stairs or to walk 2 miles. If questions are selected properly, it is possible to select starting questions about actions that are very easy to do and to end with questions about actions that are very difficult to do. Each question, moreover, has sublevels of difficulty. Walking 2 miles can be done without difficulty, with some difficulty, with great difficulty, or not at all, and each level represents a separate measure of difficulty. Thus, a 10-item questionnaire can represent 4 × 10 separate levels of difficulty or 30 item thresholds. In developing a questionnaire, all of the levels must be considered. An ideal questionnaire would therefore space out the individual difficulties as evenly as possible. It is an axiom of proper questionnaire scaling that, on average, a person who can accomplish activities at a given level of difficulty can also accomplish all items that have lesser degrees of difficulty.

In addition to evenly spacing item difficulties, it is desirable to have a questionnaire that measures a long span of difficulties. It is relatively easy to capture the functional level of persons who are severely disabled (e.g., unable to walk or to arise), but it is much more difficult to measure items at the other end of the spectrum. That is the reason that floor effects are commonly seen in the HAQ series of questionnaires. The problem with questions at the floor end of the disability spectrum is that they often have to refer to activities that people do not often do or that are not necessarily a part of the unidimension of function as much as they are of dimensions such as the performance of athletic activities.

Rasch analysis provides statistical methods to identify items that do not "fit" the hypothesized unidimensional Rasch model or that are not answered accurately. The SF-36 PF scale, which otherwise has superb psychometric properties, has items that do not fit the Rasch model. Similarly, the MD-HAQ questions regarding participation in sports and walking 2 miles do not satisfy the fit criteria. In general, items that are not clearly understood or are not completed add noise (inaccuracy) to the measurement scale, since persons guess at their ability to perform these activities. A further example of this problem can be found in the HAQ question regarding bathing. Because many people use showers instead of bathtubs, arthritis patients' responses indicate that it is more "difficult" to take a bath "with difficulty" than it is to be "unable" to take a bath at all.

A questionnaire with evenly spaced, well-fitting items can provide a good measurement tool, much as a ruler can. However, if the integers on the ruler are not evenly spaced or tend to clump together, the ruler will be less useful as a measurement tool. Furthermore, it is possible to design a "perfect" scale and yet have a scale that is not clinically useful or that is insensitive to change. The validation studies of the HAQ-II show that it performs as well as the "gold standard" HAQ in identifying treatment effect and predicting important outcomes such as mortality or work disability. In addition, it is as strongly related to clinical and outcome variables as is the HAQ, or even more so.

The 10-item scale is easier than the HAQ to use and score in the clinic and in research studies. Because the scales are so closely allied (Figure 2) and have mean scores that differ by only 0.02 units, it is relatively easy to substitute one scale for another. The very large sample size of this study (n = 14,038) provides assurance of the accuracy of the process of converting research data from the HAQ to the HAQ-II and vice versa.

Although we have indicated above that the HAQ and HAQ-II cannot be substituted in individual patients, that warning applies only to contiguous observations, for example, observations 2 and 3. However, if the substitution is continued to observation 4, then the new scale that now has 2 observations can take over from the old one. As with all such changes, experience and thoughtful use of the questionnaire will allow substitution.

The structure of the HAQ-II may seem strange, since it does not use ADL categories. The HAQ places its 20 questions into 8 ADL categories. Each category has its own score, a score that is based only on the most abnormal answer in the category. Ideally, the overall HAQ score would be a measure of functional disability averaged over all of the ADL categories. One problem with this visualization is that categories would somehow have to be weighted, either to be equal in difficulty or to represent some known, expected weight or value for the category. However, there are no known weights, nor is there evidence that equality of categories is rational or correct. In practice, the situation is worse. The HAQ hygiene category, for example, has a Rasch difficulty of −0.82 compared with a difficulty of −0.68 for "activities" (9). Hygiene, which should be much easier than "activities," is not, and is driven almost entirely by the very difficult "take a bath" question. It is therefore the case that the actual item difficulties, rather than their categorization, are what drive the HAQ score. The HAQ-II ignores ADL categorization, as does the SF-36, in order to build a psychometrically valid questionnaire. This may not be a loss, since it is difficult to express ADL category performance based on a single question within a category. Clinicians who require detailed information regarding specific categories or activities (e.g., hand function) should consider the use of activity- or area-specific questionnaires.

There has been increasing recognition of the conceptual importance of separating functional limitations and disability (51–53). Among the limitations of both the HAQ and the HAQ-II is that they mix items measuring functional limitations with items measuring disability. Nine of the 10 HAQ-II items assess functional limitations; only one ("doing outside work") is a measure of disability. It would be ideal if both instruments only assessed functional limitations. Future functional and disability assessments are likely to have increasing sophistication as the interactions among illness, function, disablement, and society become increasingly recognized (54).

In conclusion, the HAQ-II is a reliable and valid 10-item questionnaire that performs at least as well as the HAQ and is simpler to administer and score. Conversion from HAQ to HAQ-II and from HAQ-II to HAQ for research purposes is simple and reliable. The HAQ-II can be used in all places where the HAQ is now used, and it may prove to be easier to use in the clinic.

REFERENCES

1. 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.
2. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
3. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: a review of its history, issues, progress, and documentation. J Rheumatol 2003;30:167–78.
4. Ramey DR, Raynauld JP, Fries JF. The Health Assessment Questionnaire 1992: status and review. Arthritis Care Res 1992;5:119–29.
5. Wolfe F, Michaud K, Gefeller O, Choi HK. Predicting mortality in patients with rheumatoid arthritis. Arthritis Rheum 2003;48:1530–42.
6. Wolfe F, Hawley DJ. The long-term outcomes of rheumatoid arthritis: work disability: a prospective 18 year study of 823 patients. J Rheumatol 1998;25:2108–17.
7. Wolfe F, Zwillich SH. The long-term outcomes of rheumatoid arthritis: a 23-year prospective, longitudinal study of total joint replacement and its predictors in 1,600 patients with rheumatoid arthritis. Arthritis Rheum 1998;41:1072–82.
8. Michaud K, Messer J, Choi HK, Wolfe F. Direct medical costs and their predictors in patients with rheumatoid arthritis: a three-year study of 7,527 patients. Arthritis Rheum 2003;48:2750–62.
9. Wolfe F. Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ (DHAQ), and a rescored 20 item HAQ (HAQ20): analyses in 2,491 rheumatoid arthritis patients following leflunomide initiation. J Rheumatol 2001;28:982–9.
10. Wolfe F. A reappraisal of HAQ disability in rheumatoid arthritis. Arthritis Rheum 2000;43:2751–61.
11. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheumatol 1996;35:574–8.
12. Stucki G, Stucki S, Bruhlmann P, Michel BA. Ceiling effects of the Health Assessment Questionnaire and its modified version in some ambulatory rheumatoid arthritis patients. Ann Rheum Dis 1995;54:461–5.
13. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
14. Pincus T, Swearingen C, Wolfe F. Toward a multidimensional Health Assessment Questionnaire (MDHAQ): assessment of advanced activities of daily living and psychological status in the patient-friendly Health Assessment Questionnaire format. Arthritis Rheum 1999;42:2220–30.
15. Callahan LF, Pincus T, Huston JW III, Brooks RH, Nance EP Jr, Kaye JJ. Measures of activity and damage in rheumatoid arthritis: depiction of changes and prediction of mortality over five years. Arthritis Care Res 1997;10:381–94.
16. Pincus T, Brooks RH, Callahan LF. Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. Ann Intern Med 1994;120:26–34.
17. Callahan LF, Bloch DA, Pincus T. Identification of work disability in rheumatoid arthritis: physical, radiographic and laboratory variables do not add explanatory power to demographic and functional variables. J Clin Epidemiol 1992;45:127–38.
18. Wolfe F. The psychometrics of functional status questionnaires: room for improvement. J Rheumatol 2002;29:865–8.
19. Andrich D. Rasch models for measurement. Newbury Park (CA): Sage Publications; 1988.
20. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah (NJ): Lawrence Erlbaum Associates; 2001.
21. Daltroy LH, Logigian M, Iversen MD, Liang MH. Does musculoskeletal function deteriorate in a predictable sequence in the elderly? Arthritis Care Res 1992;5:146–50.
22. Fisher WP Jr. Physical disability construct convergence across instruments: towards a universal metric. J Outcome Meas 1997;1:87–113.
23. Linacre JM. Understanding Rasch measurement: estimation methods for Rasch measures. J Outcome Meas 1999;3:382–405.
24. McHorney CA, Haley SM, Ware JE Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10). II. Comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol 1997;50:451–61.
25. McHorney CA. Generic health measurement: past accomplishments and a measurement paradigm for the 21st century. Ann Intern Med 1997;127:743–50.
26. Raczek AE, Ware JE, Bjorner JB, Gandek B, Haley SM, Aaronson NK, et al, International Quality of Life Assessment. Comparison of Rasch and summated rating scales constructed from SF-36 physical functioning items in seven countries: results from the IQOLA Project. J Clin Epidemiol 1998;51:1203–14.
27. Stucki G, Daltroy L, Katz JN, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: the whole may not equal the sum of the parts. J Clin Epidemiol 1996;49:711–17.
28. Wright BD, Masters GN. Rating scale analysis: Rasch measurement. Chicago: Mesa Press; 1982.
29. Wolfe F, Michaud K. Heart failure in rheumatoid arthritis: rates, predictors, and the effect of anti-tumor necrosis factor therapy. Am J Med 2004;116:305–11.
30. Fisher WP Jr. Reliability statistics. In: Linacre JM, editor. Rasch measurement transactions part 2. Chicago: MESA Press; 1992. p. 238.
31. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.
32. Linacre JM, Wright BD. A user's guide to BIGSTEPS: Rasch model computer program (version 2.8). Chicago: Mesa Press; 1997.

33. Wolfe F, Hawley DJ, Wilson K. The prevalence and meaning of fatigue in rheumatic disease. J Rheumatol 1996;23:1407–17.
34. Meenan RF, Gertman PM, Mason JH, Dunaif R. The Arthritis Impact Measurement Scales: further investigations of a health status measure. Arthritis Rheum 1982;25:1048–53.
35. Meenan RF. The AIMS approach to health status measurement: conceptual background and measurement properties. J Rheumatol 1982;9:785–8.
36. Stucki G, Liang MH, Stucki S, Bruhlmann P, Michel BA. A self-administered Rheumatoid Arthritis Disease Activity Index (RADAI) for epidemiologic research: psychometric properties and correlation with parameters of disease activity. Arthritis Rheum 1995;38:795–8.
37. Fransen J, Hauselmann H, Michel BA, Caravatti M, Stucki G. Responsiveness of the self-assessed Rheumatoid Arthritis Disease Activity Index to a flare of disease activity. Arthritis Rheum 2001;44:53–60.
38. Fransen J, Langenegger T, Michel BA, Stucki G. Feasibility and validity of the RADAI, a self-administered rheumatoid arthritis disease activity index. Rheumatology (Oxford) 2000;39:321–7.
39. Lerner DL, Rogers WH, Chang H. Technical report: scoring the Work Limitations Questionnaire (WLQ) scales and the WLQ Index for estimating work productivity loss. Boston: The Health Institute, Division of Clinical Care Research; 2003.
40. Mchorney CA, Ware JE Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994;32:40–66.
41. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
42. Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol. York (UK): Publications Unit, Centre for Health Economics, University of York; 1996.
43. Nord E. EuroQol: health-related quality of life measurement. Valuations of health states by the general public in Norway. Health Policy 1991;18:25–36.
44. EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.
45. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21: 271–92.
46. Van der Heijde DM, van't Hof MA, van Riel PL, van Leeuwen MA, van Rijswijk MH, van de Putte LB. Validity of single variables and composite indices for measuring disease activity in rheumatoid arthritis. Ann Rheum Dis 1992;51:177–81.
47. Linacre JM, Wright B. Winsteps (3.31). Chicago: Mesa Press; 2000.
48. Andrich D, Sheridan B, Lyne A, Luo G. RUMM 2010 (Rasch unidimensional measurement models) (3.3). Duncraig (Australia): 2001.
49. Stata Corporation. Stata statistical software: release 8.1. College Station (TX): Stata Corporation; 2003.
50. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135–60.
51. Verbrugge LM. The iceberg of disability. In: Stahl SM, editor. The legacy of longevity: health and health care in later life. Newbury Park (CA): Sage Publications; 1990. p. 55–75.
52. Verbrugge LM, Jette AM. The disablement process. Soc Sci Med 1994;38:1–14.
53. Nagi SZ. Disability concepts revisited: implications for prevention. In: Tarlov AR, editor. Disability in America: toward a national agenda for prevention. Washington (DC): National Academy Press; 1991. p. 309–26.
54. World Health Organization. International classification of functioning, disability and health (ICF). Geneva: World Health Organization; 2001.