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 2491 Rheumatoid Arthritis Patients Following Leflunomide Initiation

FREDERICK WOLFE

ABSTRACT. ***Objective***. To determine whether the full Health Assessment Questionnaire (HAQ), the shortened modified HAQ (MHAQ), or the new shortened RA-HAQ, developed on the basis of Rasch item response theory (IRT), performs best in terms of distributional characteristics, detection of functional loss, and identification of change in functional status in patients with active rheumatoid arthritis (RA).

Methods. A total of 2491 clinic patients with RA with active disease from the practices of 519 US rheumatologists were assessed by questionnaire at the time leflunomide was started and at subsequent followup when there had been sufficient time for response.

Results. The HAQ scores were almost normally distributed along the 0–3 scale, but 95% of MHAQ and RA-HAQ values were clustered between 0 and 1.5. Normal or minimally abnormal scores (0 or 0.125) were noted in 6.6% of HAQ but in 21–22% of MHAQ/RA-HAQ. Mild functional loss (≤ 0.375) was found in 12.7, 39.1, and 36.1% of patients by the HAQ, MHAQ, and RA-HAQ, respectively. This indicates that the MHAQ and RA-HAQ generally fail to identify appropriately the extent of functional loss in RA. The HAQ was significantly better at detecting changes than the MHAQ or RA-HAQ, with relative efficiencies of 1.28 and 1.37 compared to the MHAQ and RA-HAQ, respectively. This results in roughly a 20–26% reduction in sample size requirements. Two additional HAQ were identified that performed better than the HAQ itself, a 20 item HAQ without the use of aids and devices and an 8 item HAQ composed of the most difficult item in each of the 8 HAQ subscale categories.

Conclusion. The HAQ is better (more efficient) than the MHAQ or RA-HAQ at detecting treatment change, and identifies the extent of functional disability better than the shortened questionnaires. The 3 questionnaires have different means, sensitivities, and distributional properties and cannot be thought of as simply different versions of the same questionnaire. The benefits of the MHAQ and RA-HAQ are that they are short and easier to score. But these benefits come at the price of loss of sensitivity and loss of sensitivity to change. The 20 item HAQ and the difficult 8 item HAQ are intriguing additional choices that are worthy of further study. (J Rheumatol 2001;28:982–9)

Key Indexing Terms:
HAQ MODIFIED HEALTH ASSESSMENT QUESTIONNAIRE RA-HAQ
HEALTH ASSESSMENT QUESTIONNAIRE RHEUMATOID ARTHRITIS
OUTCOME ASSESSMENT RASCH ANALYSIS

Functional assessment by self-report questionnaire has become virtually mandatory in randomized clinical trials (RCT) following the strong imperative of the American College of Rheumatology (ACR)[1] and European groups that such assessments should form part of the "core" of disease activity measurements[2-4]. A similar strong recommendation for the use of functional assessment questions has been made for observational studies[5], and for clinical practice as well[6].

By far, the most commonly used questionnaire is the Health Assessment Questionnaire (HAQ). The HAQ family of instruments derives from the Stanford Health Assessment Questionnaire[7]. In its original form the HAQ consists of 20 questions in 8 different categories (Table 1). Each category contains 2 to 3 questions based on a common theme:

From the National Data Bank for Rheumatic Disease–Arthritis Research Center Foundation, and University of Kansas School of Medicine, Wichita, Kansas, USA.

Supported in part by a grant from Aventis Pharmaceuticals, Inc.

F. Wolfe, MD.

Address correspondence to Dr. F. Wolfe, National Data Bank for Rheumatic Diseases, 1035 N. Emporia, Suite 230, Wichita, KS 67214. E-mail: fwolfe@southwind.net

Submitted May 1, 2000 revision accepted November 17, 2000.

dressing, standing, eating, walking, toileting, reach, grip, and instrumental activities. A score is derived for each category based on the most abnormal activity (question) in that category. In addition, the use of aids and devices to help with function is taken into consideration in the scoring. The final HAQ score is the average score of the 8 categories.

There have been 2 modifications to the HAQ in rheumatoid arthritis (RA) (Table 1). The MHAQ or modified HAQ is a subset of 8 items taken from the 8 categories. Designed by Pincus from the original 20 item HAQ[8], it has had extensive use in many rheumatic disorders. The MHAQ was conceived to address several perceived problems with the HAQ. First, the HAQ was thought to be too long: 20 questions and a list of more than 20 aids and/or devices. Second, it was perceived as complicated and time consuming to score. By contrast, the MHAQ does not consider aids or devices, has only 8 questions, is simple to score, and has the same range as the HAQ (1–3). It is widely used. Ziebland, *et al* found that the MHAQ correlated better with important clinical measures than did the HAQ[9], in support of the shortened version.

But the MHAQ has come under considerable criticism. Stucki, *et al* and Serrano, *et al* found it unacceptable because of a bunching of values at the lower end of the scale[10] and lack of sensitivity to change[11,12]. Tennant, *et al* examined the MHAQ using Rasch analyses[13,14] and found that it did not have desirable questionnaire characteristics. Reporting Rasch analyses, Stucki, *et al* suggested that differences in item difficulty might explain the distributional differences between the HAQ and MHAQ, and proposed that a properly designed shortened HAQ instrument might obviate these differences[12]. Tennant, *et al* then developed such a questionnaire, a 3rd HAQ called the RA-HAQ because its 8 questions were derived from the HAQ and because it was validated in an international sample of patients with RA[13,14]. The RA-HAQ differs from the MHAQ in that 3 of the 8 questions are different. The RA-HAQ has near perfect characteristics according to the Rasch item response theory model[15-17], and is scored on a 0–3 scale. The RA-HAQ has never been used clinically. We include it here because it enables us to test a series of different HAQ, including one that has near perfect theoretical characteristics, thereby investigating the hypothesis of Stucki, *et al*.

While it is desirable to use the shortest possible questionnaire, it is of interest whether a short questionnaire performs as well as a long one and whether it seems to be capturing patients' functional loss adequately. Katz, *et al* examined short and long questionnaires (including the MHAQ) in 54 patients undergoing hip surgery[18]. They concluded that none of the instruments could be shown to be more sensitive to change than another, and that much larger samples would be required to demonstrate statistically significant differences. We recently studied initial and followup questionnaires on 2491 patients with RA from the time they started leflunomide. We used this data set to examine the distribution characteristics of the 3 HAQ and to determine their relative sensitivity to change. The patients starting leflunomide had quite active RA, and received this drug shortly after its release in the US. Because the MHAQ has been criticized for having too many normal values, using a set of patients with severe RA where not many normal or mild patients would be found would appear to be a fair test of the MHAQ in a real clinical setting.

MATERIALS AND METHODS

Patients. Patients with RA in this study were from the practices of 566 US rheumatologists. Patients completed a HAQ questionnaire in their rheumatologists' offices at the time they started taking leflunomide. A followup HAQ was sent to the patients in the 6 month period that followed their initial enrollment. Since this project is continuing, only those patients who completed at least 2 questionnaires are analyzed in this report. To assess whether 2 assessments were an adequate measure of treatment effect, analyses were performed that included more than 2 HAQ assessments per patient. Results (data not shown) were very similar to those obtained using 2 assessments.

Scoring of the HAQ. The categories of the HAQ are shown in Table 1. In the HAQ[7], the most abnormal score in one of 8 categories becomes the score for that category. In addition, if an aid or device or help from others is used to perform an activity within the category then the minimum score for that category is set at 2 or "performed with difficulty or assistance." Scoring for the HAQ is: without difficulty (0), with some difficulty (1), with much difficulty or with assistance (2), unable (3). Thus the range of the HAQ is 0–24, or 0–3 when the total score is divided by 8. The MHAQ and RA-HAQ are scored by taking the average of the 8 items in Table 1. Aids and devices are not considered. The HAQ can be scored in 15 seconds[19], and the MHAQ and RA-HAQ in half that time. To address the issue of the effect of an assistive device, we calculated a 20 item HAQ based on the average of all 20 HAQ items without the use of assistive devices, and an 8 item HAQ score as in the original HAQ, but without the use of assistive devices, and an 8 item HAQ based on the most difficult items in each subscale category, as determined by Rasch analysis as shown in Table 1 (see below).

Statistical methods. The relative efficiency of the MHAQ and RA-HAQ was compared to the HAQ using the method of Liang, *et al*[20]. Sample sizes were calculated comparing change scores to 0, assuming an alpha level of 0.05 and power of 80%. Statistical differences in the ability of the HAQ, MHAQ, and RA-HAQ to detect changes were assessed by conditional logistic regression followed by determination of the Bayesian information criterion (BIC)[21,22], and were also assessed in distribution-free analyses by the sign test. The BIC is a measure of overall fit and a means to compare nested and non-nested models. The alpha level was set at 0.05, and all tests were 2 tailed. Statistical analyses were performed using Stata, version 6.0[23].

Rasch analysis is a method for obtaining objective, fundamental linear measures (qualified by standard errors and quality control fit statistics) from stochastic observation of ordered category responses[24]. The details of this method are available in a number of conceptually and mathematically simple[15] and more complex texts[16,17,25,26]. An important property of Rasch analysis is that the difficulties of the individual HAQ items, called HAQ difficulties, are easily expressed on a linear scale, allowing comparison between items as to difficulty or severity. This leads to the condition in which a total score of a patient (e.g., total HAQ score) will be a *linear* representation of the HAQ item difficulties, and as such will represent function loss. Thus the difference in a Rasch score between 0.5 and 1 is exactly the same as the difference between 2.3 and 2.8, and, assuming that 0 represents no disability or the beginning of the scale, a patient who increases his score from 0.5 to 1 on the Rasch linear ruler will be exactly twice as bad.

Table 1. The HAQ, MHAQ, and RA-HAQ questionnaire item sets. Items in parentheses represent HAQ category difficulty when aids and devices are included in the coding. The most difficult item in a category subscale is indicated by ±.

MHAQ	RA-HAQ	HAQ	Question	Subscale	Difficulty
				Dressing and grooming	0.84 (0.81)
•	•	•	Dress yourself		0.08 ±
		•	Shampoo your hair		0.78
				Rising	0.90 (1.10)
	•	•	Stand up from a chair		0.15 ±
•		•	Get in and out of bed		0.38
				Eating	–0.05 (0.40)
		•	Cut your meat		0.33
•	•	•	Lift a full cup or glass to mouth		1.03
		•	Open a new carton of milk		–0.94 ±
				Walking	0.62 (0.73)
•		•	Walk outdoors on flat ground		0.71
	•	•	Climb 5 steps		–0.28 ±
				Hygiene	–0.99 (–0.82)
•	•	•	Wash and dry entire body		0.83
		•	Take a bath		–1.94 ±
		•	Get on and off the toilet		1.16
				Reach	–0.47 (–0.44)
		•	Reach and get down a 5 lb object		–1.30 ±
•	•	•	Bend down and pick up clothing		0.25
				Grip	0.09 (–1.10)
	•	•	Open car doors		0.47
		•	Open jars (previously opened)		–0.57 ±
•		•	Turn taps on and off		0.78
				Activities	–0.93 (–0.68)
	•	•	Run errands and shop		– 0.19
•		•	Get in and out of car		0.02
		•	Do chores		–1.74 ±

This is in contrast to the "raw" HAQ score that is not linear in its original metric. Rasch analysis was performed using Winsteps version 3.05[27]. This analysis was used to determine the difficulty (measure) of each HAQ item on a linear scale. Negative (–) measures indicate more difficult items and positive measures less difficult items (Table 1). The linear scales mean that the measures represent difficulty on a linear scale, and therefore can be compared to each other. Rasch analysis reports several other statistics. The patient separation statistic indicates how clearly the test can "see" between patients. It is increased by increasing the number of items and indicates the utility of the test as a measuring device. Item separation indicates how clearly the sample "shows" the differences between HAQ items. It is increased by the number of persons, and indicates the clarity with which the test articulates a construct[17]. The INFIT and OUTFIT statistics are measures of how well a HAQ item fits to the Rasch model. The mean square INFIT and OUTFIT statistic is expected to be 1. Values above 1 indicate the presence of noise or "fuzziness." As applied to HAQ items, values > 1.2 indicate substantial and unacceptable noise. Noisy items can provide more noise than information, and may not contribute useful information to an overall score[17].

RESULTS

Demographic and clinical characteristics. The demographic and clinical characteristics of study patients are shown in Table 2a and 2b. HAQ and MHAQ scores are high. The median HAQ and MHAQ score was 0.62 and 1.38, respectively. Therefore patients in this series were considerably more severe than the clinic patients seen by Stucki, *et al*, as might be expected since they had all failed a disease modifying antirheumatic drug (DMARD) and were starting a new DMARD[10].

The distribution of the various HAQ. Figure 1 displays the distributional characteristics of the 3 HAQ clinical scales. There are a number of important points. The MHAQ and RA-HAQ are quite similar, but both differ substantially from the HAQ. The HAQ is almost normally distributed, but the shortened HAQ have far fewer "severe" observations and many more "mild" and normal observations. These

Table 2A. Demographic and clinical characteristics of 2491 patients with RA at the start of leflunomide therapy.

Variable	Mean OR%	SD
Age (years)	58.01	12.57
Sex (% male)	22.4	
Disease duration (years)	12.37	10.31
High school graduate	87.4	
VAS pain scale (0–10)	4.71	2.68
VAS global severity scale (0–10)	3.98	2.45

VAS: visual analog scale.

Table 2B. Metric properties of the various HAQ.

HAQ	Mean	SD	Patient Separation	Item Separation	Percentage of Patients with Score of 0
HAQ score (0–3)	1.30	0.67	2.37	22.69	4.00
MHAQ score (0–3)	0.64	0.49	1.96	9.64	12.84
RA-HAQ score (0–3)	0.71	0.53	2.06	12.05	12.26
HAQ — no assistive devices	1.11	0.68	2.40	19.46	6.48
HAQ — difficult 8 items	1.02	0.63	2.53	20.04	5.51
HAQ — 20 items	0.82	0.55	3.32	22.39	5.00

differences have clinical implications. Minimal functional loss (0 or 0.125) is found in 6.6% of patients scored by the HAQ, but 22.3 and 21.0% scored by the MHAQ and RA-HAQ, respectively. Similarly, it had been the author's clinical opinion that mild functional loss is represented by a HAQ score of 0.375 or less. Using that criterion, 12.7% of patients scored by the HAQ, 39.1% scored by the MHAQ, and 36.1% scored by the RA-HAQ would have mild functional loss while the rest would have more severe loss. The percentage of patients with 0 scores is shown in Table 2b. From data in Figure 1 and Table 3, it would follow that the means of the 3 measures would differ as well. As shown in Table 1, the respective mean and SD for the HAQ, MHAQ, and RA-HAQ is 1.30 (0.67), 0.64 (0.49), and 0.71 (0.53).

Sensitivity to change: relative efficiency and sample size requirements. The 6 month change in HAQ scores (mean and SD) as a function of treatment for the 3 HAQ was 0.053 (0.43), 0.042 (0.38), and 0.041 (0.39) for the HAQ, MHAQ, and RA-HAQ, respectively. From this the relative efficiency of the HAQ compared to the shortened HAQ can be calculated. Compared to the MHAQ, the HAQ relative efficiency is 1.28, and compared to the RA-HAQ it is 1.37. We can then apply the study data to the sample size requirements for finding a difference as great as observed in this report. Assuming an alpha of 0.05 and a power of 80%, the sample size required is 515, 644, and 698 for the HAQ, MHAQ, and RA-HAQ, respectively.

Conditional logistic regression was then used to estimate

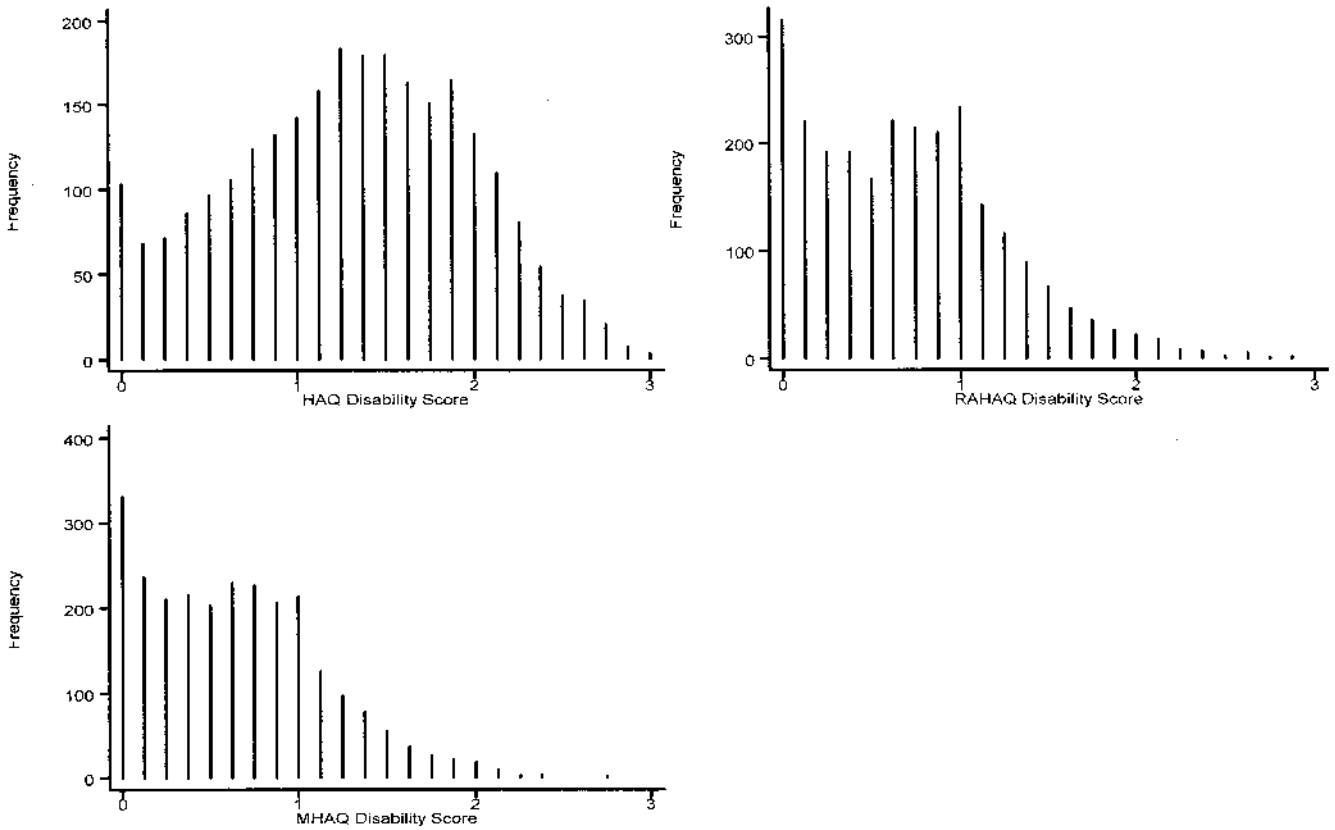

Figure 1. Distributional graphs of HAQ, MHAQ, and RA-HAQ for 2491 clinic patients with RA at the start of leflunomide therapy. HAQ scores are roughly normally distributed. The MHAQ and RA-HAQ scores cluster between 1 and 1.5 and many values are normal or almost normal.

Table 3. Distribution and cumulative distribution of HAQ, MHAQ, and RA-HAQ scores at first assessment.

Score	HAQ, %	HAQ Cumulative, %	MHAQ, %	MHAQ Cumulative, %	RA-HAQ, %	RA-HAQ Cumulative, %
0.000	4.00	4.00	13.02	13.02	12.38	12.38
0.125	2.64	6.63	9.28	22.31	8.62	21.00
0.250	2.75	9.39	8.26	30.57	7.52	28.53
0.375	3.34	12.72	8.50	39.06	7.52	36.05
0.500	3.72	16.45	7.99	47.05	6.54	42.59
0.625	4.07	20.52	9.01	56.06	8.66	51.25
0.750	4.77	25.29	8.89	64.95	8.42	59.68
0.875	5.12	30.41	8.10	73.05	8.23	67.91
1.000	5.51	35.92	8.38	81.43	9.17	77.08
1.125	6.13	42.05	4.92	86.35	5.60	82.68
1.250	7.10	49.15	3.78	90.13	4.55	87.23
1.375	6.94	56.09	3.03	93.15	3.49	90.71
1.500	6.94	63.03	2.12	95.28	2.59	93.30
1.625	6.32	69.36	1.46	96.73	1.80	95.10
1.750	5.82	75.17	1.02	97.76	1.37	96.47
1.875	6.36	81.54	0.87	98.62	1.02	97.49
2.000	5.12	86.66	0.71	99.33	0.86	98.35
2.125	4.23	90.88	0.35	99.69	0.71	99.06
2.250	3.10	93.99	0.12	99.80	0.31	99.37
2.375	2.09	96.08	0.16	99.96	0.27	99.65
2.500	1.44	97.52	0.00	99.96	0.08	99.73
2.625	1.32	98.84	0.00	99.96	0.20	99.92
2.750	0.78	99.61	0.04	100.00	0.04	99.96
2.875	0.27	99.88	0.00	100.00	0.04	100.00
3.000	0.12	100.00	0.00	100.00	0.00	100.00

the extent of the differences over 6 months for each HAQ test, and the differences between the tests were described using the Bayesian information criterion (BIC)[21,22]. The BIC values for HAQ versus MHAQ, HAQ versus RA-HAQ, and MHAQ versus RA-HAQ were 10.93, 12.46, and 1.55. These results indicate "very strong support" for the HAQ versus MHAQ and RA-HAQ, and weak support for the superiority of MHAQ compared to RA-HAQ[21,22]. The various HAQ were also compared using distribution free statistics. The 2 sided sign test showed that the median 6 month differences for the HAQ exceeded the median differences for the MHAQ ($p = 0.003$) and the RA-HAQ ($p = 0.0004$). The MHAQ and RA-HAQ did not differ ($p = 0.476$).

Rasch analysis and reasons for the differences between the various HAQ. Table 1 presents the Rasch item difficulties for the HAQ items. Each HAQ subscale category is composed of items of substantially different difficulties. For example, for hygiene the difference in difficulty between "wash and dry body" (0.83) and "take a bath" (–1.94) is 2.77. The average HAQ difficulty for this item is –0.99 without aids and devices and –0.82 with aids and devices. But both MHAQ and RA-HAQ select the "wash and dry body" for their item in the category. Indeed, for almost all of the categories the various HAQ have selected items with substantial differences in difficulty. Overall, the MHAQ and RA-HAQ are "easier" tests than the full HAQ.

There were generally small differences between the HAQ with and without aids and devices, with one striking exception, the grip category. Here, the HAQ grip difficulty with aids or devices was 0.09, but with assistive devices it was –1.10. This difference is explained almost entirely by the very common usage of devices that help with opening of jars. More than 75% (75.5%) of patient observations indicated the use of a gripping or jar opening aid. The next highest percentage was for a reaching aid (36.7%).

The various HAQ instruments were also examined by other Rasch analysis statistics (Table 2b). The person and item separation statistics for the HAQ, MHAQ, and RA-HAQ were: HAQ 2.37, 22.69; MHAQ 1.96, 9.64; RA-HAQ 2.06, 12.05. As noted in Materials and Methods, person separation indicates how clearly the test can "see" between "patients." It indicates the utility of the test as a measuring device. Item separation indicates how clearly the sample "shows" the differences between items. It indicates the clarity with which the test articulates a construct. These data indicate, as confirmed by statistical and efficiency data noted above, that the HAQ distinguishes patients better than the other HAQ instruments. Using the Winsteps fitting criteria, we found non-fitting items in all the HAQ. The HAQ had one non-fitting item, "take a tub bath," but the non-fit was large, 1.57 and 1.51 for the INFIT and OUTFIT, respectively. The MHAQ had 2 slightly non-fitting items,

"turn taps on and off" and "lift a full cup or glass to the mouth." The INFIT and OUTFIT statistics for these items ranged between 1.20 and 1.29. In this sample, we found the RA-HAQ to have one slightly mis-fitting item, "lift a full cup or glass to the mouth," with INFIT and OUTFIT statistics of 1.26 and 1.30, respectively.

Rescoring the HAQ: using all 20 items. When all the 20 HAQ items were examined together (Table 2b) we found the overall separation to be 3.32 for patients and 22.39 for HAQ items. There were 2 non-fitting items, "opening jars" with fit statistics of 1.17 and 1.24, and "being able to take a tub bath" at 1.85 and 1.84, for INFIT and OUTFIT, respectively. Eight of the individual items duplicated difficulty measures already present in other items. The mean HAQ score for the 20 item HAQ was 0.82.

A modified HAQ using the most difficult items: the Difficult HAQ (DHAQ). When the HAQ was scored in the usual way, using the most abnormal score of the category but not making use of assistive devices, the separation was 2.40/19.46. The mean was 1.11, and the percentage with 0 scores was 6.48%. Removal of assistive devices left only one of the 8 categories with misfit, "hygiene," with INFIT/OUTFIT statistics of 1.52/1.51.

We also performed Rasch analyses on the most difficult item in each of the 8 categories. These items are shown in Table 1 and marked with a ± symbol. The patient and item separation was 2.53 and 20.04. There were 2 mis-fitting items, "opening jars" and "taking tub baths," with INFIT and OUTFIT mis-fitting from 1.21 to 1.64. The mean HAQ score for these difficult 8 items was 1.02.

As shown in Figure 2, the HAQ with the greatest spread and greatest mean was the HAQ, followed by the HAQ without assistive devices, the "difficult" 8 item HAQ, the 20 item HAQ, the RA-HAQ, and the MHAQ. Therefore the 3 HAQ that select for the hardest items, the HAQ, HAQ without assistive devices, and difficult 8 item HAQ, have the highest HAQ score, while the 2 items that select for easier items have the lowest scores, MHAQ and RA-HAQ.

DISCUSSION

As shown by the data of this study, the HAQ is more efficient and more sensitive to change than the MHAQ or RA-HAQ. Therefore, fewer patients would be required to demonstrate change when the HAQ is used compared to the shortened versions. We confirmed this by parametric and nonparametric statistical analyses, effect size and efficiency statistics, and Rasch analyses separation statistics. Our results confirm those of Stucki, *et al*[10]. It seems possible that a different set of 8 HAQ items would perform better than the 8 used by the MHAQ or RA-HAQ, and we found some evidence for this in the "difficult" 8 item HAQ. Although the HAQ makes use of 20 items, the calculated score is based only on scores of the 8 categories. Among the 3 standard 8 item HAQ, the HAQ performed best, and there was little to choose between the MHAQ and RA-HAQ in terms of performance. When we examined the 20 items of the HAQ as a single scale, we found far better patient separation and slightly better item separation. This suggests that scoring all the items of the HAQ (without aids and devices) might be a good strategy for clinical trials, provided that conventional scoring with aids and devices is also used so that comparison with the large base of public data can be maintained. When we formed a HAQ using the most difficult items in the 8 HAQ category subscales, the patient separation statistics were improved compared to the HAQ, but were not as good as the separation statistic for the 20 item HAQ.

For a disability instrument, the HAQ has the controversial and unusual property of increasing the score when an

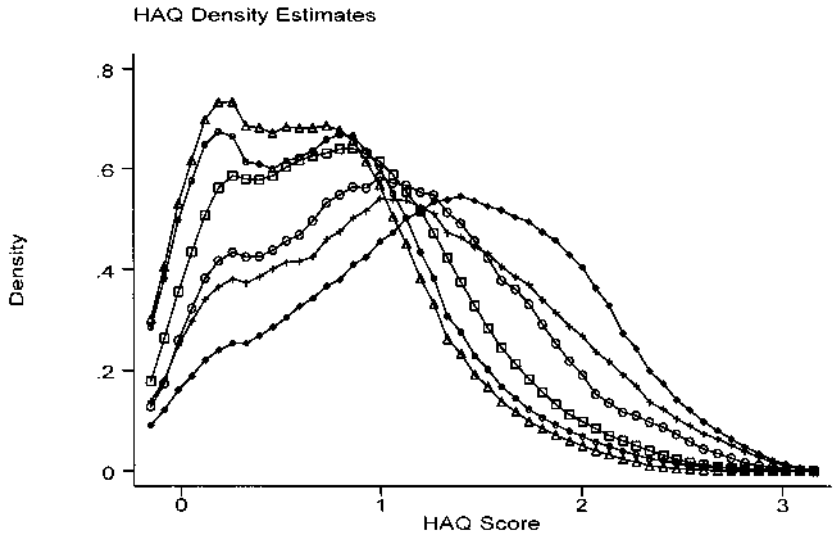

Figure 2. Kernel density estimates from right to left for HAQ (◆), HAQ without assistive devices (+ sign), "difficult" 8 item HAQ (○), 20 item HAQ (□), RA-HAQ (○) and MHAQ (Δ).

assistive device is used. The data from this study elucidate the role of assistive devices. As shown in Figure 2 and Table 2b, when the most difficult 8 items are used, the score is 1.02 and the curve is the third from the right. The HAQ without assistive devices has a score of 1.11 and its curve is the second from the right in Figure 2. The HAQ without assistive devices differs from the 8 item difficult HAQ by allowing questions in other categories to count if the most difficult items are not scored highest. These curves indicate that only a small proportion of the differences between the HAQ without assistive devices (curve second from right) and the HAQ with assistive devices (curve at the right) is attributable to the assistive devices. These 2 HAQ differ by only 0.19 in their mean scores. In addition, the separation statistics are similar. The main differences, then, between these 2 HAQ is the small difference in mean scores.

Beyond the difference in efficiency and sensitivity to change among the HAQ, MHAQ, and RA-HAQ, there are striking differences in the distributional characteristics of the full and shortened instruments (Figures 1 and 2). The MHAQ and RA-HAQ instruments miss a great deal of functional impairment. For that reason there are many more normal or only minimally impaired patients when the shortened instruments are used. It is probably true, then, that these instruments fail to identify functional loss, particularly at the lower end of the scale, and this observation is supported by the item difficulties in the Rasch analyses. Thus the data of this study confirm and reinforce the observations of Stucki, *et al* regarding floor effect[10], but this time in a sample of patients with severe and relatively unresponsive RA.

We also confirm that HAQ and MHAQ scores are not comparable[10-12], nor can simple algorithms be used to convert one score to the other. As shown in Figure 3, a nonlinear equation can be written to describe the average relationship between the 2 scales, but predicting individual HAQ scores from individual MHAQ scores is not possible within acceptable degrees of reliability because of the wide forecast intervals. Therefore the 2 questionnaires are different: they have different scaling and measurement properties, and must be thought of as separate questionnaires, not just as short or longer versions of the same questionnaire. The differences between the HAQ and MHAQ should be extended to all the HAQ used in this study. Each has different metric properties, and the instruments cannot be mixed.

The RA-HAQ did not work as well in this sample as it did in a sample of RA patients with milder disease[13]. It clearly failed to identify functional loss as well as the HAQ. This observation underscores a number of limitations: the difficulty in making shortened instruments work as well as longer, more sensitive instruments, and the need to have larger item banks from which instruments can be developed. The HAQ, with scoring of the full 20 items, performed better than any of the 8 category HAQ. In addition, the difficult 8 item HAQ performed better than the HAQ itself. The duplicated difficulty measures of the 20 item HAQ, and to a lesser extent of the difficult 8 item HAQ, however, raise some concern, for it is possible under some circumstances for patients' HAQ scores to improve even though overall functional ability has not changed when a number of items share the same difficulty level. Additional analyses of the 20

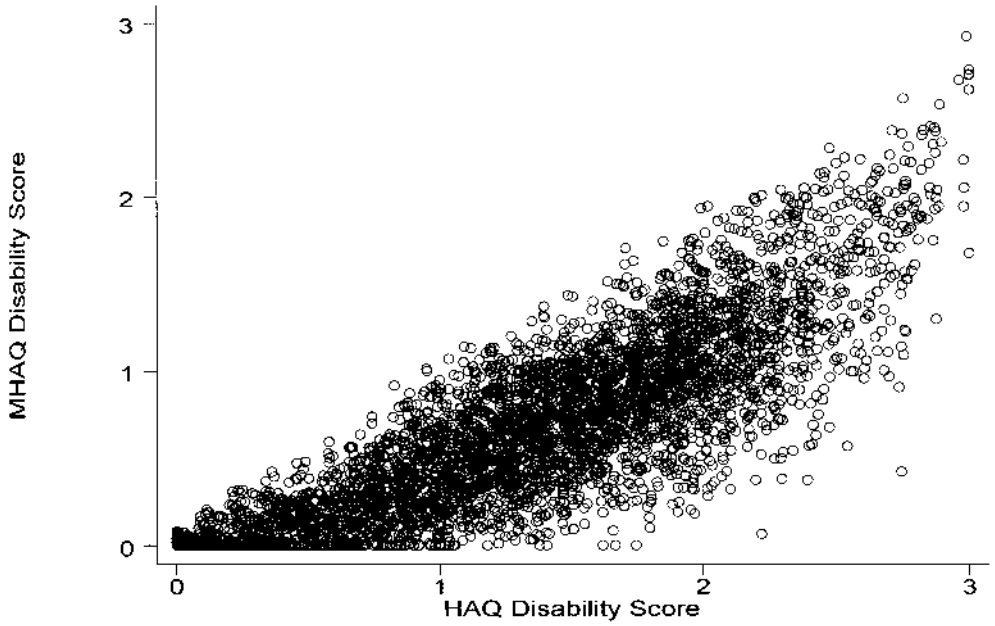

Figure 3. MHAQ versus HAQ. The questionnaires yield considerably different results when applied to the same patients. The MHAQ and HAQ are separate questionnaires and should not be considered simply as shorter or longer variants of one another. A small amount of spherical random noise is added to each point before graphing in order to display overlapping values.

item and difficult 8 item HAQ are required before firm recommendations can be made.

In the end, there are tradeoffs. The MHAQ is shorter and marginally easier to score. But this comes at the prices of less sensitivity to functional loss and less sensitivity to change. Where questionnaire length is not an important consideration, the HAQ is a better choice, but where space and time are important the MHAQ is able to capture functional loss, but with some loss of sensitivity to change and the ability to detect mild functional loss. The difficult 8 item HAQ is an intriguing additional choice that is worthy of further study.

ACKNOWLEDGMENT

The author thanks Professor Ben D. Wright of the University of Chicago for his instruction and helpful advice.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.
2. Vangestel AM, Anderson JJ, van Riel PLCM, et al. ACR and EULAR improvement criteria have comparable validity in rheumatoid arthritis trials. J Rheumatol 1999;26:705-11.
3. Boers M, Tugwell P, Felson DT, et al. WHO and ILAR core endpoints for symptom modifying antirheumatic drugs in RA clinical trials. J Rheumatol 1994;21 Suppl 41:86-9.
4. Scott DL, Panayi GS, van Riel PLCM, et al. Disease activity in rheumatoid arthritis — Preliminary Report of the Consensus Study Group of the European Workshop for Rheumatology Research. Clin Exp Rheumatol 1992;10:521-5.
5. Wolfe F, Lassere M, van der Heijde D, et al. Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999;26:484-9.
6. Wolfe F, Pincus T. Current comment — Listening to the patient — A practical guide to self-report questionnaires in clinical care. Arthritis Rheum 1999;42:1797-808.
7. Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137-45.
8. 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.
9. Ziebland S, Fitzpatrick R, Jenkinson C, Mowat A. Comparison of two approaches to measuring change in health status in rheumatoid arthritis: the Health Assessment Questionnaire (HAQ) and modified HAQ. Ann Rheum Dis 1992;51:1202-5.
10. 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.
11. Serrano MAB, Fabregat JB, Garzon JO. Should the MHAQ ever be used? Ann Rheum Dis 1996;55:271.
12. Stucki G, Stucki S, Bruhlmann P, Michel BA. Should the MHAQ ever be used? [reply to letter]. Ann Rheum Dis 1996;55:271-2.
13. Tennant A, Ryser L, Stucki G, Wolfe F. An 8-item international version of the HAQ: The Inter-HAQ [abstract]. Arthritis Rheum 1999;42 Suppl: S74.
14. Tennant A, Ryser L, Stucki G, Wolfe F. Applying item response theory to measurement issues: Deriving a short version of the Stanford Health Assessment Questionnaire (HAQ) — the RA-HAQ. (Submitted)
15. McNamara T. Measuring second language performance. London: Longman; 1996.
16. Andrich D. Rasch models for measurement. In: Quantitative applications in the social sciences. Newbury Park: Sage Publications; 1988.
17. Wright BD, Masters GN. Rating scale analysis: Rasch measurement. Chicago: Mesa; 1982.
18. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992;30:917-25.
19. Wolfe F, Kleinheksel SM, Cathey MA, Hawley DJ, Spitz PW, Fries JF. The clinical value of the Stanford Health Assessment Questionnaire Functional Disability Index in patients with rheumatoid arthritis. J Rheumatol 1988;15:1480-8.
20. Liang MH, Larson MG, Cullen KE, Schwartz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-7.
21. Raftery AE. Bayesian model selection in social research. In: Marsden PV, editor. Sociological methodology. Oxford: Basil Blackwell; 2000:111-63.
22. Long JS. Regression models for categorical and limited dependent variables. Thousand Oaks: Sage Publications; 1997.
23. Stata Corporation. Stata statistical software: Release 6.0. College Station, TX: Stata Corporation; 1999.
24. A user's guide to BIGSTEPS: Rasch model computer program. Chicago: Mesa Press; 1997.
25. Fischer GH, Molenaar IW. Rasch models: foundations, recent developments, and applications. New York: Springer-Verlag; 1995.
26. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.
27. Winsteps Version 3.05. Chicago: Mesa Press; 2000.