Health Assessment Questionnaire modifications: is standardisation needed?

M M Zandbelt, P M J Welsing, A M van Gestel, P L C M van Riel

Abstract

***Background*—Physical disability is part of the end point measures in rheumatoid arthritis clinical trials. The Stanford Health Assessment Questionnaire Disability Index (HAQ DI) is often used for this purpose but lacks international uniformity owing to variations in the translated and adapted questionnaires and variations in its calculation. To study the consequences of these variations the previous Dutch HAQ (HAQ90) was revised, resulting in a new Dutch HAQ (HAQ99).**

***Objective*—To compare DI scores from the two versions, and to study the consequences of applying different calculation methods for the DI score.**

***Methods*—78 patients completed both the HAQ99 and the HAQ90. To compare the use of different category score calculation methods a post hoc analysis on prospectively collected data obtained in clinical trials was performed.**

***Results*—No statistically significant differences were observed between the DI scores of the HAQ90 and the HAQ99 using the alternative method (that is, without correcting for aid and devices). However, correcting for the use of aid or devices or not did result in statistically significant different DI scores. The systematic shift when using the maximum or mean item score for calculation of the category score resulted in non-comparable absolute DI scores.**

***Conclusion*—The use of HAQ DI questionnaires with different numbers of items and/or categories does not hinder international comparability, except when these variations interfere with the calculation method of the DI (as in the case of questionnaires without a section correcting for devices). For the sake of international uniformity the HAQ or any validated translation should be used and calculated in a standard way, including correcting for the use of aid and devices, and taking the maximum within each category as the category score.**

(*Ann Rheum Dis* 2001;**60**:841–845)

Department of Rheumatology, University Medical Centre St Radboud, Nijmegen, The Netherlands
M M Zandbelt
P M J Welsing
A M van Gestel
P L C M van Riel

Correspondence to:
Dr M M Zandbelt,
Department of Rheumatology, UMC St Radboud Nijmegen, PO Box 9101, 6500 HB Nijmegen, The Netherlands
M.Zandbelt@reuma.azn.nl

Accepted 9 February 2000

In the past two decades the need for standardisation of measurement procedures in rheumatoid arthritis clinical trials has been recognised. As a consequence a worldwide consensus about a “core set of end point measures in rheumatoid arthritis (RA) clinical trials” was established.[1] Physical disability, as measured by self reported questionnaires, is part of this core set.

The most frequently used worldwide questionnaire to measure physical disability is the Stanford Health Assessment Questionnaire Disability Index (HAQ DI).[2] This questionnaire has been validated and shown to be reproducible. Since its publication in 1980 the Stanford HAQ,[3] which consists of several sections with questions, for example, about functional disability, pain, drug side effects, and economic aspects, has been modified several times. The disability index section (HAQ DI), which measures functional disability is the only standardised section and has remained the same since 1982. It contains 20 questions in eight categories, and includes a section about aid from other people or the use of devices to correct, if necessary, the answers given to the 20 questions. Since the introduction of this questionnaire there have been numerous translations and modifications, resulting in a wide range of versions.[4] These versions differ in number of items, number of categories, number of items in each category, and the presence or absence of a section correcting for the use of devices. Table 1 shows these differences, which may occur even within countries.

The frequently used modified HAQ (MHAQ) DI with eight items is based on a different (transition) question than the one used in the original HAQ DI, and provides three instead of four alternatives to choose from (less difficult, equally difficult, or more difficult than before).[5 6]

In addition to the existence of different HAQ DI versions, the way they are calculated has not been consistent. According to the manual, the DI score can be calculated using either a so-called standard method or an alternative method. In the standard method the eight category scores are corrected using the section on aid and devices at the bottom of the questionnaire. Whenever aid by others or the use of devices is required to perform a certain activity, the corresponding category score (range 0–3, 0 = best, 3 = worst) is increased to 2 when that score was 0 or 1. In the alternative method no correction for aid and devices is made.

Table 1 Different versions of the Health Assessment Questionnaire (HAQ) in the Netherlands

	Categories	*Items*	*Aid/ devices*
Siegert (1984)	8	24	No
Van der Heijde (1990, HAQ90)	9	23	No*
New version (1999, HAQ99)	8	20	Yes

The original Stanford HAQ contains 20 items in eight categories and includes a section to correct for the use of aid/devices.
*No section on aid/devices but to each of the 23 items an option was added for this purpose.

In both standard and alternative methods the DI should be calculated by taking the maximum item score within each category as the category score, and then calculating the mean of the eight category scores. Others prefer taking the mean within a category, and then calculating the mean of the category scores,[7] or just calculating the mean of the 20 items. In most papers, however, the methods section does not reveal at all which HAQ DI was used and which method was applied to the DI score.

As a result, in practice, the HAQ appears to lack international uniformity owing to variations in the translated and adapted questionnaire itself as well as variations in calculation of the DI.

To obtain an instrument to study possible consequences of using different HAQ versions and different calculation methods the previous Dutch HAQ (HAQ90) was revised, resulting in a new Dutch HAQ (HAQ99). The HAQ99 is based on an accurate translation of the Stanford HAQ DI and does not have any modifications. This in contrast with both other HAQ DI versions that are used in the Netherlands, one of which is the validated HAQ90.[8] This HAQ90 differs from the Stanford HAQ DI in a number of ways: (*a*) it contains 23 questions; (*b*) it has nine categories; and (*c*) it has no section on need for aid and/or devices, but in answering the 23 questions, the third of four alternatives is "yes, but with a need for devices or aid from others". This was done because in some cases it was unclear which category should be corrected when a patient mentioned the use of a specific device. Finally, the number of items within each category differs in comparison with the Stanford HAQ DI. This is partly because when the HAQ90 was developed, questions were added especially for the Dutch situation.

This study aimed at comparing DI scores obtained from the HAQ99 with those resulting from the HAQ90, and studying the consequences of applying different calculation methods for the DI. Furthermore, patients' problems when filling in the HAQ DI were listed.

Methods

REVISION OF THE DUTCH HEALTH ASSESSMENT QUESTIONNAIRE

Two rheumatology researchers translated the Stanford HAQ DI. Both translations were then back translated from Dutch into English by one native English speaker and one English teacher, independently of each other and without consulting the other translators. The resulting four back translations, two translations, and the original Stanford HAQ DI were then compared by all participants. Any differences in the versions were discussed and a consensus was reached about the alternative that best matched the original Stanford HAQ DI. This resulted in the HAQ99, consisting of 20 questions in eight categories and including a section on aid/devices.

COMPARISON OF THE STANFORD HAQ AND DUTCH HAQ99 RESULTS

Nine bilingual patients with RA (that is, patients who watched BBC regularly) were asked to complete the HAQ99 at the outpatient clinic and the English Stanford HAQ DI at home. DI scores of both questionnaires were calculated in both alternative (without correcting for the use of aid or devices) and standard (including this correction) ways and compared using a paired *t* test.

COMPARISON OF THE HAQ90 AND HAQ99

During three weeks 92 consecutive patients with RA were asked to complete the HAQ99 directly after consulting their rheumatologist at the outpatient clinic, and to complete the HAQ90 within a week at home as well. Only the results of patients who filled in both questionnaires were analysed. DI scores were calculated according to the manual—that is, taking the maximum item score within each category as the category score. Firstly, the results without correcting for devices (that is, the alternative method) of both HAQ versions were compared using a paired *t* test. Then the standard method (that is, including correction for use of devices or aid from others) was used to calculate the DI score of the HAQ99. This score was also compared with the HAQ90 score using a paired *t* test; because the HAQ90 lacks a section on aid/devices, correcting for use of aid/devices was not possible for this version.

Furthermore, the DI scores obtained from the HAQ99 after correction and without correction for aid/devices (standard and alternative methods, respectively) were compared.

INVESTIGATION OF DIFFICULTIES ENCOUNTERED WHEN FILLING IN THE QUESTIONNAIRE

After completing the HAQ99 at the outpatient clinic, 62 consecutive patients were interviewed and asked if they had any problems when answering the items of the questionnaire. After this, any complaints or remarks were categorised in main problem fields and frequencies were calculated.

CONSEQUENCES OF DIFFERENT CALCULATION METHODS

The consequences of different calculation methods were studied by a post hoc analysis on prospectively collected data of completed HAQ90s of two different patient groups. The groups included patients with refractory RA in a clinical trial with an intravenous anti-tumour necrosis factor α drug (group A, n=39), and patients with early RA participating in another clinical trial in which sulfasalazine and methotrexate were evaluated (group B, n=103). In each group the DI score was calculated using the maximum score within a category as the category score (that is, according to the Stanford HAQ manual, maxHAQ), as well as using the mean score within a category as the category score (meanHAQ). In both calculation methods the DI score was obtained by taking the mean of the eight category scores and using the alternative method—that is,

Table 2 Group mean disability index (DI) scores in the Stanford HAQ and HAQ99 using two different calculation methods

HAQ version (n=9)	*Stanford HAQ (standard)*	*Stanford HAQ (alternative)*	*HAQ99 (standard)*	*HAQ99 (alternative)*
DI score (SD)	1.46 (0.81)	1.21 (0.85)	1.47 (0.77)*	1.19 (0.75)†

*Stanford HAQ standard versus HAQ99 standard, not significant.
†Stanford HAQ alternative versus HAQ99 alternative, not significant.

Table 3 Group mean disability index (DI) scores in the HAQ90 and HAQ99 using two different calculation methods

HAQ version (n=92)	*HAQ90**	*HAQ99 (standard)*	*HAQ99 (alternative)*
DI score (SD)	0.96 (0.79)	1.14 (0.69)†	0.92 (0.65)‡

*The HAQ90 version can only be calculated in the alternative method (that is, without correcting for devices).
†HAQ90 versus HAQ99 standard, p<0.001.
‡HAQ90 versus HAQ99 alternative, not significant.

without correction for the use of aid or devices. DI scores were calculated at two different sequential time points in group A (0 and 20 weeks) and in group B (0 and 24 weeks). Differences between the DI scores measured at the two time points were statistically analysed using paired *t* tests. Furthermore, in both groups the absolute changes of DI scores over the 20 week period, calculated using the maxHAQ calculation method, were compared with the absolute DI score changes calculated with the meanHAQ method using Wilcoxon signed rank tests. The same was done for relative changes of DI scores.

Results

COMPARISON OF THE STANFORD HAQ AND DUTCH HAQ99 RESULTS

All nine patients who were asked filled in both the HAQ99 and the English Stanford HAQ DI. Of these patients (mean age 47, seven female, two male) all were rheumatoid factor (RF) positive and seven (78%) had an erosive arthritis. The mean erythrocyte sedimentation rate (ESR) of this group was 13 mm/1st h. There were no significant differences between the two versions in both ways of calculation (standard as well as alternative, p=0.80 and p=0.88, respectively). Table 2 summarises these findings.

Table 4 Problems patients encounter when completing the HAQ (n=62)

Problem field		*No (%)*
I	No problems at all	24 (39)
II	Variation from day to day in disability	12 (19)
III	To choose between SOME disability and MUCH disability	9 (15)
IV	Difficulties about what to fill in when using a special device	5 (8)
V	Two questions in one	5 (8)
VI	Activities which the subject normally doesn't perform at all	2 (3)
VII	Other, not classified comments	5 (8)

Table 5 Trend in group median disability index (DI) scores using the mean or maximum item score as the category score in two clinical trial groups

	No	*Week 0*	*Week 20**	*p*	*pΔ*	*pΔrel*
Group A	39					
MaxHAQ (P25–75)†		2.11 (1.33–2.56)	1.67 (1.06–2.28)	0.002		
MeanHAQ (P25–75)		1.59 (1.05–2.12)	1.19 (0.77–1.69)	0.001	0.942	0.232
Group B	103					
MaxHAQ (P25–75)		0.78 (0.42–1.58)	0.22 (0.00–0.78)	<0.001		
MeanHAQ (P25–75)		0.49 (0.18–1.15)	0.10 (0.00–0.42)	<0.001	<0.001	0.068

*Group B: 24 weeks.
†P25–75 = interquartile range.
pΔ = absolute DI changes weeks 0–20 maxHAQ versus meanHAQ.
pΔrel = relative DI changes weeks 0–20 maxHAQ versus meanHAQ.

COMPARISON OF THE HAQ90 AND HAQ99

Both the HAQ90 and the HAQ99 were completed by 78/92 (85%) patients. Of these patients (mean age 61.5, 48 female, 30 male), 58 (74%) were RF positive and 35 (45%) had an erosive arthritis. The mean ESR for this group was 17 mm/1st h. When the alternative method was used (that is, without correcting for devices) the mean DI scores were almost equal: 0.96 (HAQ90) and 0.92 (HAQ99) (p=0.36). The HAQ99 using the standard method (that is, corrected for the use of devices or aid by other people) had a mean DI score of 1.14, which was significantly higher than the DI score of the HAQ90 (p<0.001) using the alternative method (table 3). Also, within the HAQ99 there was a clear difference between the two DI scores (using standard and alternative methods) (table 3).

INVESTIGATION OF ENCOUNTERED DIFFICULTIES WHEN FILLING IN THE QUESTIONNAIRE

Table 4 presents the six main problem fields reported by patients when completing the questionnaire.

CONSEQUENCES OF DIFFERENT CALCULATION METHODS

Table 5 summarises the results obtained in the post hoc analysis on prospectively collected data of completed HAQ90s, comparing the two different calculation methods for the DI score. In both patient groups a significant decrease of DI scores at the final time point was found for both calculation methods. Furthermore, in the patients with refractory RA (group A) absolute changes of DI scores over the 20 week period, calculated by using the maxHAQ as compared with the meanHAQ calculation method, did not differ significantly (p=0.94). In the patients with early RA (group B), however, absolute DI score changes, measured by the two different calculation methods, were significantly different (p<0.01). For the relative DI score changes no significant differences were seen in either patient group (p=0.23 and p=0.07, respectively).

Furthermore, of particular interest was the observation that in the patients with refractory RA (group A) the change of DI scores calculated using the mean HAQ instead of the maxHAQ calculation method was about equal to the effect of the therapeutic intervention (table 5).

Discussion

Since 1982 the HAQ has been widely used to measure functional disability in patients with RA. Patient-assessed function, which is often measured with the HAQ, is part of the American College of Rheumatology (ACR) response criteria set and is frequently used in clinical trials. Partly because of the international character of clinical trials, a large number of translations and local adaptations of the HAQ have been developed and validated. This has led to the use of many different versions of the questionnaire even within countries. Also, within our country several different versions of the HAQ DI, which have different numbers of questions and different ways of correcting for use of devices,[8 9] are being used. To improve uniformity and to obtain an instrument to study possible consequences of these differences we revised the validated HAQ90,[8] producing the HAQ99. In this study it was shown that the exact number of questions (20 or 23) and/or categories (eight or nine) in the HAQ DI used has no major influence on the DI score when calculated according to the manual. For international uniformity, however, and to facilitate comparison of results, it is better to strive for maximum similarity to the original Stanford HAQ DI of all local versions. In contrast with the number of questions and/or categories, correcting for the use of devices or not (that is, applying the standard or alternative method) leads to significantly different absolute DI scores. This is of particular interest because case report forms used in clinical trials often do not contain this section, probably to facilitate data entry. Thus to enable exchange of international results, it should be clear which method (standard or alternative) is used. It would be even more useful to reach an international consensus about whether the section on aid and devices should be included or not.

The commonly used MHAQ differs apart from a different number of questions (eight items by taking one of each category) from the Stanford HAQ DI in asking a different kind of (transition) question and offering three instead of four alternative responses to this question. Therefore it was not investigated in this study. Next to patient satisfaction the MHAQ measures transitions in functional disability from visit to visit in clinical trials instead of absolute scores of functional disability at any time point. The rationale behind this modification is understandable and the MHAQ has been reported to be a better instrument than the original Stanford HAQ DI for measuring changes in disability during clinical trials.[6] But one should realise that the MHAQ cannot be used as part of the ACR response criteria.[10] In addition, theoretically, by reducing the number of questions, the MHAQ may fail to detect clinically relevant changes in activities of daily living in patients with relatively little impairment. When these major differences are taken into account the MHAQ does not contribute to international uniformity.

Another aspect which lacks international uniformity is the way in which the eight category scores are calculated. According to the manual the DI score should be calculated by taking the mean of the category scores, with the category score being the maximum score of the questions in that category (here called maxHAQ). However, a method in which the mean score within a category is taken as the category score is also being used (sometimes referred to as the alternative HAQ method[7] (here named the meanHAQ). In practice, even a third alternative is sometimes used by taking the mean of all the questions together as the DI. In the post hoc analysis on prospectively collected data the chosen method (maxHAQ/meanHAQ) for calculation of the category scores had no affect on monitoring the clinical course, but as was expected resulted in different absolute DI scores. In our dataset the effect of applying the maxHAQ compared with the meanHAQ calculation method was equal to the effect of the therapeutic intervention in the patients with refractory RA (table 5). Thus calculating the category score in one way rather than another does not affect monitoring the course of functional disability within a group, but because one method leads to higher absolute DI scores than the other method this undermines the possibilities of worldwide between-group comparison of measured DI scores; this was also noted by Ramey *et al*.[4]

Theoretically, differences in ACR20 response rates could occur when the HAQ is used as a response criterion, because different category score calculation methods (mean versus maximum) and DI score calculation methods (standard versus alternative) have been shown to lead to different absolute DI scores. In this study, however, in contrast with the observed absolute DI score changes, the relative DI score changes measured using the max HAQ and meanHAQ calculation method did not differ significantly. As the ACR20 response rate is a relative change parameter (20% improvement from baseline score) in this study ACR20 responses measured as patients' assessed function did not differ significantly between the calculation methods. However, further investigations are needed to exclude possible consequences of applying different DI calculation methods for ACR response rates.

From the patients' point of view some general problems appear to remain when using the HAQ. Although an investigation of the difficulties encountered by patients showed that about 40% of them had no difficulties at all, a few problems were noted frequently by the remaining patients. One in five patients has difficulty in choosing the correct answer because of day to day variations in disability. In the interviews most of them indicated that they choose a "mean answer" rather than a "bad answer" on bad days. In fact to choose between "some difficulty" and "much difficulty" was noted as a problem on its own in 15% of the patients. About 8% of the patients have difficulties when filling in the section on aid and devices, and an almost equal number of patients mention that there is an item with two questions in one. Because most of these problems can be solved by a clear, uniform instruction to the patient, the instructions given are of special importance to get accurate data from the HAQ. The Stanford Institute also provides a manual for the investigator. We recommend this manual to establish a standardised patient instruction method.

Taking all this together the use of HAQ DI questionnaires with different number of items and/or categories does not appear to hinder international comparability, except when these variations interfere with the calculation method of the DI (as in the case of questionnaires without a section correcting for devices). For the sake of international uniformity we suggest that the HAQ or any validated translation should be used and calculated in a standard way, including correcting for the use of aid and devices and taking the maximum within each category as the category score.

We thank A Haagmans, an english teacher, and P Donnelly, a native english speaker, for their help in back translating the translated HAQ. We also thank A den Broeder and C Haagsma for providing data of clinical trials.

1 Boers M, Tugwell P, Felson DT, van Riel PL, Kirwan JR, Edmonds JP, *et al.* World Health Organization and International League of Associations for Rheumatology core endpoints for symptom modifying antirheumatic drugs in rheumatoid arthritis clinical trials. J Rheumatol Suppl 1994;41:86–9.

2 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.

3 Fries JF, Spitz PW, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23: 137–45.

4 Ramey DR, Raynauld JP, Fries JF. The health assessment questionnaire 1992: status and review. Arthritis Care Res 1992;5:119– 29.

5 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.

6 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.

7 Tomlin GS, Holm MB, Rogers JC, Kwoh CK. Comparison of standard and alternative health assessment questionnaire scoring procedures for documenting functional outcomes in patients with rheumatoid arthritis. J Rheumatol 1996; 23:1524–30.

8 van der Heijde DM, van Riel PL, van de Putte LB. Sensitivity of a Dutch Health Assessment Questionnaire in a trial comparing hydroxychloroquine vs. sulphasalazine. Scand J Rheumatol 1990;19:407–12.

9 Siegert CE, Vleming LJ, Vandenbroucke JP, Cats A. Measurement of disability in Dutch rheumatoid arthritis patients. Clin Rheumatol 1984;3:305–9.

10 Felson DT, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, *et al.* American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum1995;38:727–35.