The assessment of functional status in rheumatoid arthritis: a cross-cultural, longitudinal comparison of the HAQ and the GARS


3.1 Introduction

Instruments assessing functional ability in patients with Rheumatoid Arthritis (RA) have been studied extensively (Liang and Jette 1981, Liang 1987, Bombardier and Tugwell 1987, Pincus et al. 1989, Spector and Hochberg 1992, Liang and Katz 1992, Hawley and Wolfe 1992, Fitzpatrick 1993). In these studies, the importance of the health status instruments is stressed in that they provide feedback to the physician about the treatment programme, indicate the need for health services and monitor the course of the disease. A distinction can be made between disease-specific measures, developed to measure characteristics of one specific group of patients, and generic instruments, which are developed to measure general aspects of different diseases. The Health Assessment Questionnaire (HAQ) is a disease-specific and frequently used instrument for the evaluation of functional ability in patients with RA (Fries et al. 1980). The HAQ covers a range of 8 domains of activities of daily living and provides information about the use of devices or help from other sources in performing these activities.

A point of debate concerns the possible sex specificity of certain items that may lead to sex differences in outcome assessment. It has been found that female RA respondents appear to be physically more restricted than male patients with RA (Thompson and Pegley 1991, Da Silva and Hall 1992, Krol et al. 1995). An explanation for this result may be that the questions with respect to the activities of interest are sex biased. Inevitably, questions directed to functional capacity in stabilizing independency, include questions on self-care activities and household tasks. According to the role allocation in most societies, the performance of some of these activities will more often and with more ease be done by women and therefore more realistically be appraised by women. Due to physical strengths, other tasks will be performed with more ease by men. Consequently, these differences between men and women will be reflected in the answers on the questionnaires. However, another explanation for this
The phenomenon may just be the fact that women and men with exactly the same observed functional status may report differently, for example, because pain is differently appraised and more easily expressed (Da Silva and Hall 1992).

Five years ago, the Groningen Activity Restriction Scale (GARS) has been developed as an instrument which quantifies the degree of functional capacity in doing self-care and household activities independently (Kempen and Suurmeijer 1990, Kempen et al. 1993, Suurmeijer et al. 1994). Like the HAQ, the GARS concentrates on activities of daily living (ADL; i.e. self-care activities), but, contrary to the HAQ, it also measures problems in the performance of instrumental activities of daily living (IADL; i.e. household activities). Although both ADL and IADL items are included, the GARS is an unidimensional instrument. Factor analysis yields one (strong) general component, which can be ascribed as ‘disability’ or ‘functional capacity’. It is also a hierarchical scale, which implies firstly that the items can be ordered according to their difficulty, and secondly that two respondents with the same score meet the same problems in doing ADL or IADL activities. As opposed to the HAQ, the GARS is a generic instrument, which has been administered in several fields of research: among the elderly, among patients with Multiple Sclerosis (MS), cancer, RA, and among healthy persons. Although the GARS is a recently developed instrument in comparison with the HAQ, its psychometric properties are highly satisfactory and stable across different patient populations (Kempen and Suurmeijer 1990, Kempen et al. 1993, Suurmeijer et al. 1994).

Both instruments are similar in that they are short and can be asked or used as self-report questionnaires and completed within a few minutes. As such, both instruments are appropriate for integration in routine individual patient care. Also the focus of the two instruments is similar: assessing the degree of functional capacity in the performance of ADL, although the GARS takes also IADL into consideration.

Concerning the differences between both instruments, first of all, the HAQ has a longstanding record and has been proven and documented in many studies to be an effective predictor of disability, and sensitive in clinical trials (Liang and Katz 1992, Hawley and Wolfe 1992, Fitzpatrick 1993, Fries et al. 1980, Thompson and Pegley 1991, Leigh and Fries 1992). Therefore, it was rather remarkable to find that the internal consistency, expressed as Cronbach’s $\alpha$ or $\rho$, has never been published as far as could be traced. This is probably due to the rather complex way of assigning scores to respondents. The GARS is well documented on its psychometric properties.

Information on the use of aids or devices is not the primary focus of the GARS. Instead, the instrument is directed toward restrictions of the individual in performing certain self-care (ADL) as well as
household (IADL) activities and tasks. Finally, both the ADL items and the IADL items within the GARS also form separate scales with very good psychometric properties, which can be used independently from the total scale (GARS).

Based on cross-sectional and longitudinal data sets, derived from patients with RA from the Netherlands, France and Norway, the following research questions have been addressed: (1) how do the psychometric qualities of the HAQ and the GARS compare to each other? (2) What is the convergent validity of the HAQ and the GARS scores? (3) What are the sensitivities to change within a one-year period of both instruments? And finally (4), what can be said about sex differences in scoring on the HAQ and the GARS.

3.2 Materials and methods

3.2.1 Selection of patients

In the context of the EUropean Research on Incapacitating DIseses and Social Support (EURIDISS 1990) a homogeneous sample of patients with RA was constituted. The patients all fulfilled 4 or more of 7 ARA criteria: for research purposes widely used criteria for classification of rheumatoid arthritis (Arnett et al. 1988). The presence of 4 ARA criteria was assessed by the treating rheumatologist, who also determined the year of onset of RA, defined as the year (and month) in which the patient fulfilled the fourth ARA criterion for the first time. The patients should be aged between 20 and 70 years, and should not have another chronic disease or handicap that might interfere with RA. Finally, patients with RA as severe as Steinbrocker grade IV were also excluded because the research period of four years was expected to be too much of a burden for them. Preceding the main study, an extensive pilot study was conducted in France and in the Netherlands, to test the research protocol and the instruments to be used in the main study (Doeglas et al. 1994). The final sample for T1 comprised 634 patients with RA: 290 from the Netherlands, 116 from France, and 228 patients from Norway (see table 3.1).

The mean disease duration for all countries together was 2.2 years (sd 1.3). The mean age of this group was 52 years (sd 12), and 69% were women. These values varied a little between the countries. Patients were followed for a four year period. Each year they will be medically checked and within a period of ten days they will be interviewed by a trained interviewer. The GARS and the HAQ were not administered at the same time and the same way: the HAQ was conducted in the hospital (self-report) during the medical examination while the GARS was filled out at home by the interviewer. The longitudinal results in this article are based upon two yearly measurements: T1 and T2. Due to either loss in follow-up or to missing data, complete T1 and T2 data of 565 patients were
available for comparison: 271 from the Netherlands, 104 from France, and 190 from Norway.

3.2.2 Measures

The HAQ as a self-report questionnaire, has been frequently used in research among patients with RA (Fries et al. 1980). In this study, the modified HAQ version for use among British patients was used (Kirwan and Reeback 1986, see appendix 1). The HAQ contains 20 items distributed on eight components, all reflecting ADL activities: ‘dressing and grooming’, ‘arising’, ‘eating’, ‘walking’, ‘hygiene’, ‘reaching’, ‘gripping’ and a component ‘other’. The scores of the HAQ run from: 0) without any difficulty; 1) with some difficulty; 2) with much difficulty; to 3) unable to do. The total HAQ index is obtained by the procedure proposed by Fries, and incorporates the use of devices and/or assistance from other people (Fries et al. 1980). The highest score on any of the items within one component, represents the component score. The respondent has also to indicate whether (s)he uses devices or help from other people. Any positive response regarding help or assistive devices raises the corresponding score automatically to ‘with much difficulty’. The addition of the 8 component scores divided by 8, results in the HAQ score. The total HAQ score ranges from zero to three. The higher the score, the more limitations there are in daily functioning.

Though the GARS can be used as a self-report questionnaire, in the present study the GARS was administered by the interviewer (Kempen and Suurmeijer 1990, Kempen et al. 1993, Suurmeijer et al. 1994). Contrary to the HAQ, which was conducted in the hospital, the GARS was filled out at home. The GARS has 18 items² divided across two subscales (appendix 1): an ADL subscale, assessing activities of daily living (dressing, washing oneself, etc.) and an IADL subscale, assessing instrumental activities of daily living (mainly household activities). The answer categories of the GARS run from: 1) fully independently without any difficulty; 2) fully independently but with some difficulty; 3) fully independently but with great difficulty.

Table 3.1 Demographic variables at T1

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>The Netherlands</th>
<th>Norway</th>
<th>All countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>116</td>
<td>290</td>
<td>228</td>
<td>634</td>
</tr>
<tr>
<td>Women (%)</td>
<td>70</td>
<td>64</td>
<td>74</td>
<td>69</td>
</tr>
<tr>
<td>Age (±sd)</td>
<td>53 (±11)</td>
<td>54 (±12)</td>
<td>51 (±13)</td>
<td>52 (±12)</td>
</tr>
<tr>
<td>Disease duration (±sd)</td>
<td>2.6 (±1.4)</td>
<td>1.8 (±1.2)</td>
<td>2.4 (±1.2)</td>
<td>2.2 (±1.5)</td>
</tr>
</tbody>
</table>

a In years

The assessment of functional status in rheumatoid arthritis
difficulty; 4) cannot do it fully independently, only with someone’s help; to 5) cannot do it at all, need complete help. To standardize the GARS and the HAQ scores, the GARS scores were adapted as follows: first, the categories 4 and 5 of the GARS were collapsed into one category. Next, to make the scores comparable, the four categories were recoded from 0 till 3, as with the HAQ. After adding up all item scores, and dividing the result by the number of items, a total scale score for the GARS is obtained which also runs from 0 till 3; the higher the score, the more activity restrictions. Apart from the other disease related components, the ARA criteria for functional status (Steinbrocker functional classes, 1949) were the first quantified criteria, based on clinical judgment. Major changes in function can be detected by this rather global procedure. Steinbrocker classes are frequently used in order to compare new instruments assessing functional status. The classes are defined as follows: I) complete functional capacity with ability to carry on usual duties without handicaps; II) functional capacity adequate to conduct normal activities despite of discomfort or limited mobility of one or more joints; III) functional capacity adequate to perform only few or none of the duties of usual occupation or self-care; IV) largely or wholly incapacitated with patient bedridden or confined to wheelchair, permitting little or no self-care. As stated before, patients with RA as severe as Steinbrocker grade IV were excluded from the study. The Ritchie Articular Index (RAI, 1968) was used to assess the level of tenderness. The administration of the RAI was conducted by the research nurse, who firmly pressed the 24 joints on which the index is based. The patient’s reaction was registered as follows: 0) no pain, 1) patient complains of pain, 2) patient complains of pain and winces, 3) patient complains of pain, winces and withdraws. The tenderness score is obtained by adding up all 24 joint scores; ranging from 0 to 72. The higher score, the more tenderness the patient experiences. Finally, the Overall Evaluation of Health (OEH) was measured by the question: ‘How would you rate your health at the moment? Would you say that it is very poor or that it is excellent or that it is somewhere is in between?’ A ruler was used as a Visual Analogue Scale (VAS) to rate the subjective response.

3.2.3 Statistical Analysis
To investigate how the GARS and HAQ items compare to each other, as well as the psychometric properties of both instruments, several steps have been undertaken. First, a delphi procedure involving the authors in two judgment rounds was used to determine the highly and moderately comparable and related items between the two instruments. Those items achieving at least 80% expert agreement at the second judgement round were selected for comparison. This
procedure yielded four comparable items that together with the
HAQ and the GARS (total) scores were compared. Secondly, the
internal consistency and the scalability coefficients were determined.
To test whether the items in the scales can be hierarchically ordered,
Mokken Scale analysis for Polychotomous items (MSP) was used
(Debets and Brouwer 1989). A (strong) hierarchical scale implies, that
respondents with the same score face the same difficulties. MSP is
based on a probabilistic model that can be used for building and
testing unidimensional scales, and can analyze items with more than
two categories (Kempen and Suurmeijer 1990, Kempen et al. 1993,
Suurmeijer et al. 1994, DeJong and Molenaar 1987, Debets and
Brouwer 1989). The model assumes the existence of an underlying
latent attribute, for example ‘functional status’. The core of Mokken
scale analysis is the scalability of the items separately as well as of
the scale as a whole. First, all items are ordered according to their
difficulty, from ‘easy’ to ‘difficult’ according to the item mean scores.
Whereas other scaling models, like Guttman’s method, proceed from
the deterministic idea that one can or cannot do a certain item,
Mokken scale analysis is based on probabilities. Respondents who
react negatively on item i with a certain degree of difficulty will have
a greater chance to react negatively as well to item j with a lower
degree of difficulty compared to respondents who react positively on
item i. The extent to which this assumption is violated is expressed in
three scalability coefficients. The scalability coefficient $H_{ij}$ concerns an
item pair i and j. If i and j are totally unrelated, $H_{ij}=0$. When
according to all respondents, item j is denoted as the easiest item
compared to item i (without any violations) than $H_{ij}=1$. The
scalability coefficient $H_i$ indicates to what extend item i fits into the
total scale, which is formed by all other items. The scalability
coefficient $H$ refers to the entire scale. A set of items form a
satisfactory hierarchical scale if all $H_i$ are positive (which is
equivalent to positive correlations per pair), while $H_i$ (and thus $H$) do
not fall below a constant c. A minimum value of c of 0.30 is
recommended, but higher values of $H_i$ H and c imply fewer
violations of the model, and thus a better hierarchy. A rule of thumb
is that $0.30 \leq H < 0.40$ is called a ‘weak’ scale, $0.40 \leq H < 0.50$ is an ‘moderately
strong’ scale, and $H \geq 0.50$ is a ‘strong’ scale (DeJong and Molenaar
1987).
Furthermore, Cronbach’s $\alpha$ is based on the assumption that in an
ideal situation all items are equally correlated to each other, whereas
in typical (Mokken) scales there has to be ‘... a substantial variation in
item difficulties, and in such cases $\alpha$ strongly underestimates $\rho$
(DeJong and Molenaar 1987). Therefore, by taking the real item
distributions into account, MSP computes $\rho$, which is a better
approximation of the (real) reliability of the scale.
Third, the variance accounted for by a one factor solution with a
Principal Component Analysis (PCA) will be compared for the HAQ and GARS. The accounted variance is a measure to what extent the underlying construct (functional status) is represented by the factor solution. Suurmeijer et al. (1994) have demonstrated that the (I)ADL items together form one major and unidimensional scale, which account for 48.4% of the variance with respect to the underlying construct. Although the ADL and IADL subscales of the GARS can also be used separately, the present study concentrates on the total GARS scale.

Fourth, both measures will be related to a third (independent) measure as a measure of convergent validity. For this goal Steinbrocker functional classes will be used. When the results of the psychometric tests give no indication to keep the data for the countries separate, the following analysis was performed on all data together.

Finally, the sensitivity of both measures to measure change was assessed and compared. To this end, the longitudinal (T1-T2) change in mean scores were examined for both instruments. First, the Pearson correlation was compared of (1) the change in the RAI and the change in HAQ score and (2) the change in the RAI and the change in GARS score.

Furthermore, the sensitivity to change of the GARS total scale and subscales as well as for the HAQ was expressed with the Standardized Response Mean (SRM; Katz et al. 1992). The SRM, which is an effect size, was calculated as the mean score on T1 minus the mean score on T2 divided by the standard deviation of the change. Effect sizes of .20 are generally considered as small, of .50 as moderate differences, and of .80 as large differences (Fitzpatrick 1993). The SRM of both instruments can also be compared in terms of Relative Efficiency (RE; Liang et al. 1985). The RE was computed by squaring the ratio of appropriate t statistic, for example, \((t_{GARS}/t_{HAQ})^2\), while according to Liang et al. (1985), the t statistic is calculated as: the mean change / (standard deviation of the change / √n). The HAQ was used as the ‘standard’, a choice that is arbitrary. A RE > 1 (or < 1) means that the instrument mentioned in the numerator is a more (or less) efficient tool for measuring change than the one mentioned in the denominator. Stated simply, the instrument with the highest relative efficiency should give the best balance between sample size and statistical power to detect change between assessments.

The SRM and the RE were examined for those individuals who at T2 had improved or who had deteriorated (Fitzpatrick 1993). The OMERACT committee has stated that a change score of 30% on the RAI can be conceived as relevant change (Goldsmith et al. 1993). Accordingly, two groups were formed. Another approach was to separate patients on basis of a relative change in scores of .30 on the
OEH. Thus, another two groups were constructed. For all four groups, the SRM and the RE were calculated.
With respect to sex differences in relation to HAQ and GARS scores, a paired t test was performed both on item level and on (sub)scale level to investigate whether men and women scored differently. Additionally, to test for sex differences, a regression analysis was performed of sex on the GARS and HAQ score, respectively, controlling for the RAI. Next, sex differences were illustrated based on the observations of both T1 and T2 (1130 responses). The course of RA as measured with the HAQ and the GARS will be outlined for women and men separately, by plotting the HAQ and the GARS scores against disease duration in months.

3.3 Results

3.3.1 Psychometric Properties

In table 3.2 the mean scores and standard deviations are shown for the four most similar items of the HAQ and the GARS (according the delphi procedure). The differences in scores were tested by means of a paired t test. Also the were compared and tested for significant differences.

Table 3.2 The difference in item mean scores of the four most comparable items of the GARS and the HAQ

<table>
<thead>
<tr>
<th></th>
<th>HAQ</th>
<th>GARS</th>
<th>t value</th>
<th>p</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 In/out bed</td>
<td>.49 .66</td>
<td>.42 .62</td>
<td>-2.85</td>
<td>.005</td>
<td>.61</td>
</tr>
<tr>
<td>2 Wash body</td>
<td>.60 .85</td>
<td>.77 .95</td>
<td>5.85</td>
<td>.000</td>
<td>.65</td>
</tr>
<tr>
<td>3 Toilet</td>
<td>.47 .77</td>
<td>.29 .55</td>
<td>-5.92</td>
<td>.000</td>
<td>.39</td>
</tr>
<tr>
<td>4 Shopping</td>
<td>.72 .96</td>
<td>1.45 1.23</td>
<td>16.86</td>
<td>.000</td>
<td>.53</td>
</tr>
</tbody>
</table>

a Not significant for the Dutch sample and the Norwegian sample.
b Not significant for the French sample.

Although these four items are almost identical, the scores on these items differed significantly. The largest difference was found on the item ‘run errands and shops’ versus ‘do the shopping’. The mean value for the GARS (1.45) was about twice the score of the HAQ (.72), which is significant. For the ‘toilet’ item, the HAQ score (.47) was about twice as large as the results of the GARS (.29).
Because the scale scores ranged also from zero to three, the mean values of the GARS and the HAQ could be compared too. The mean scores of the HAQ (.97; sd .72) and the IADL subscale (1.02; sd .77) were of comparable magnitude, but differed significantly (t = -2.74; p
The mean scores of the GARS (total scale: .71; sd .57) and the ADL subscale (.51; sd .49) differed more from the HAQ mean score (t = .006, t = 16.74 and 26.15, respectively). The correlation of the HAQ with the GARS, ADL and IADL (sub)scale was .84, .80 and .79, respectively. The correlations between the single items were less strong. In table 3.3 the reliability coefficients Cronbach’s α, η as well as the coefficients of scalability (H) are presented. The results for both instruments were comparable, for the three different countries separately as well as for all countries together.

Table 3.3  Scalability (H) and reliability coefficients (Cronbach’s α) and η for the HAQ and the GARS

<table>
<thead>
<tr>
<th></th>
<th>HAQ</th>
<th>GARS</th>
<th>ADL</th>
<th>IADL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>ρ</td>
<td>α</td>
<td>H</td>
</tr>
<tr>
<td>France</td>
<td>.41</td>
<td>.93</td>
<td>.93</td>
<td>.50</td>
</tr>
<tr>
<td>The Netherl.</td>
<td>.40</td>
<td>.94</td>
<td>.94</td>
<td>.47</td>
</tr>
<tr>
<td>Norway</td>
<td>.42</td>
<td>.94</td>
<td>.94</td>
<td>.46</td>
</tr>
<tr>
<td>All countries</td>
<td>.40</td>
<td>.94</td>
<td>.94</td>
<td>.47</td>
</tr>
</tbody>
</table>

The GARS and the HAQ yielded high reliability coefficients (Cronbach’s α and η). Furthermore, although the scalability coefficient (H) for the GARS (.47) was better than that of the HAQ (.40), both scales appeared to be ‘moderately strong’ unidimensional and hierarchical scales (.40≤H<.50). The results of the ADL and IADL subscales were even better in this respect. Additional analysis learned that all H_i were positive, for both the HAQ and the GARS, and that the lowest value of H_i was .31 for the GARS and .32 for the HAQ. The variance accounted for by a one factor solution with PCA was comparable for the HAQ and the GARS. The first extracted factor accounted for 47.7% of the variance of the GARS (range: 45.1 to 50.6 across countries) versus 46.1% of the variance of the HAQ (range: 42.9 to 47.6 across countries).

In table 3.4 the results are summarized of the relationships between Steinbrocker functional classes and the RAI on the one hand, and both the GARS and HAQ on the other hand. Table 3.4 shows that the mean scores increased with the functional classes, although not to the same extent. Compared to the mean HAQ scores, the GARS total scores were lower for Steinbrocker class II and III, whereas for class I the mean scores were almost the same. The mean ADL and IADL scores for Steinbrocker functional classes displayed a corresponding pattern. The correlation of Steinbrocker functional classes with the GARS and the HAQ were of comparable magnitude and did not
Table 3.4 Mean scores of the HAQ and GARS for each of the three categories of Steinbrocker functional classes, together with the overall correlation and the correlation with the RAI

<table>
<thead>
<tr>
<th>Steinbrocker functional classes</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mn</td>
<td>sd</td>
<td>mn</td>
</tr>
<tr>
<td>HAQ</td>
<td>.11</td>
<td>.22</td>
<td>.88</td>
</tr>
<tr>
<td>GARS</td>
<td>.13</td>
<td>.16</td>
<td>.65</td>
</tr>
<tr>
<td>ADL</td>
<td>.06</td>
<td>.12</td>
<td>.45</td>
</tr>
<tr>
<td>IADL</td>
<td>.24</td>
<td>.32</td>
<td>.96</td>
</tr>
</tbody>
</table>

a Category IV was an exclusion criteria.

differ significantly from each other; .61 and .65, respectively. The correlations between the RAI and these instruments were slightly lower, but of comparable magnitude.

With respect to the longitudinal part of this study, first some general results. For all countries together, between T1 and T2 the HAQ score decreased with .03, which was not significant (p=.094). Also from the GARS it appeared that the patients had improved in this period. The scores on the GARS decreased significantly with .08 (p=.000). Next, the change scores of the GARS and the HAQ (δGARS and δHAQ) were related to change scores on the RAI and on the OEH (δRAI and δOEH). For δRAI, the correlations with δHAQ and δGARS were .34, and .27, respectively, and differed not significantly from each other (t=1.58; the critical region is t>2.04). For δOEH these correlations were -.11 and -.22 (significantly different; t=2.49).

In table 3.5, the longitudinal results, in terms of mean values, SRM and RE, are summarized for the patients who improved and who deteriorated according to the RAI and the OEH. For the patients who worsened according the RAI, the HAQ demonstrated a significant increase, whereas the GARS (and GARS subscales) did not. The SRM of the HAQ for this group was small (.27) but much larger than the SRM of the GARS (sub)scale(s). Also the RE of the GARS (and for the ADL and IADL) was smaller (.05-.18) than for the standard, i.e the RE of the HAQ (1.00). For the patients who improved according to the RAI, the HAQ and the GARS showed a significant change between T1 and T2. The SRM’s can be called moderate and were little higher for the GARS. The RE of the GARS was 1.65, compared to the standard (1.00). The HAQ and the GARS score of patients who worsened according to the OEH did not change significantly between
Functional ability, social support and quality of life

T1 and T2. The SRM’s for this group were all about the same and very small. The results in terms of RE were varying. Compared to the HAQ, the results of the GARS and ADL subscale were better, whereas for the IADL subscale worse. Finally, the HAQ did not demonstrate a change for the patients who had improved according the OEH, whereas the GARS did. Also the SRM’s of the GARS for this group were moderate and larger compared to the SRM of the HAQ, which resulted in a much larger RE for the GARS.

### 3.3.2 Sex differences

In table 3.6, the common items of the GARS and HAQ are shown (item 1 to 4, see table 3.2), together with the other items of the HAQ and the GARS on which women and men scored significantly different. It appeared that much more items of the HAQ (11 items; item 3 to 13) were sensitive for sex differences compared to the GARS (4 items; item 4 to 7). Secondly, on all the concerning GARS and HAQ items, women obtained higher scores. Furthermore, although items 1 to 4 were comparable to each other according the delphi procedure, the t values (for sex differences) of the toilet item differed significantly from each other.

Additionally, multiple regression analyses were performed with, successively, the HAQ and the GARS score as the dependent variables. As a first step, the RAI was entered into the equation. The R square change was .362 for the HAQ (p < .000), and .318 for the
Table 3.6  t Test of sex differences on item level

<table>
<thead>
<tr>
<th>Item</th>
<th>Women</th>
<th>Men</th>
<th>t value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 In/out bed</td>
<td>1.05</td>
<td>.72</td>
<td>-4.40</td>
<td>.000</td>
</tr>
<tr>
<td>2 Wash body</td>
<td>.49</td>
<td>.66</td>
<td>.49</td>
<td>.66</td>
</tr>
<tr>
<td>3 Toilet</td>
<td>.52</td>
<td>.80</td>
<td>.37</td>
<td>.70</td>
</tr>
<tr>
<td>4 Shopping</td>
<td>.83</td>
<td>1.00</td>
<td>.47</td>
<td>.81</td>
</tr>
<tr>
<td>5 Cut meat</td>
<td>.76</td>
<td>.92</td>
<td>.40</td>
<td>.68</td>
</tr>
<tr>
<td>6 Lift cup</td>
<td>.29</td>
<td>.53</td>
<td>.14</td>
<td>.40</td>
</tr>
<tr>
<td>7 Open milk carton</td>
<td>.95</td>
<td>.92</td>
<td>.62</td>
<td>.84</td>
</tr>
<tr>
<td>8 Take bath</td>
<td>1.11</td>
<td>1.14</td>
<td>.81</td>
<td>1.05</td>
</tr>
<tr>
<td>9 Get down 5 lb</td>
<td>.94</td>
<td>1.03</td>
<td>.65</td>
<td>.91</td>
</tr>
<tr>
<td>10 Open cardoor</td>
<td>.50</td>
<td>.75</td>
<td>.22</td>
<td>.46</td>
</tr>
<tr>
<td>11 Open jars</td>
<td>1.02</td>
<td>.96</td>
<td>.68</td>
<td>.90</td>
</tr>
<tr>
<td>12 Turn tap</td>
<td>.91</td>
<td>.90</td>
<td>.55</td>
<td>.79</td>
</tr>
<tr>
<td>13 Housekeeping</td>
<td>1.20</td>
<td>1.04</td>
<td>.78</td>
<td>.92</td>
</tr>
<tr>
<td>GARS</td>
<td>.74</td>
<td>.56</td>
<td>.64</td>
<td>.59</td>
</tr>
<tr>
<td>ADL</td>
<td>.52</td>
<td>.49</td>
<td>.48</td>
<td>.50</td>
</tr>
<tr>
<td>IADL</td>
<td>1.07</td>
<td>.75</td>
<td>.90</td>
<td>.80</td>
</tr>
<tr>
<td>1 In/out bed</td>
<td>.41</td>
<td>.60</td>
<td>.44</td>
<td>.66</td>
</tr>
<tr>
<td>2 Wash body</td>
<td>.77</td>
<td>.93</td>
<td>.78</td>
<td>1.01</td>
</tr>
<tr>
<td>3 Toilet</td>
<td>.29</td>
<td>.56</td>
<td>.30</td>
<td>.53</td>
</tr>
<tr>
<td>4 Shopping</td>
<td>1.62</td>
<td>1.23</td>
<td>1.07</td>
<td>1.14</td>
</tr>
<tr>
<td>5 Feet care</td>
<td>1.17</td>
<td>1.19</td>
<td>.84</td>
<td>1.07</td>
</tr>
<tr>
<td>6 Heavy housekeeping</td>
<td>1.90</td>
<td>1.12</td>
<td>1.35</td>
<td>1.16</td>
</tr>
<tr>
<td>7 Make beds</td>
<td>1.38</td>
<td>1.15</td>
<td>1.18</td>
<td>1.17</td>
</tr>
</tbody>
</table>

GARS (p < .000). As a second step, sex was entered into the model; the R square changed .011 (p = .001) and .000 (p = .658) for the HAQ and the GARS, respectively.

Besides on item level, the total HAQ score also differed significantly between men and women. Female scores were about .26 higher compared to male scores, which is 33%. For the GARS, women only got higher scores on the IADL subscale, although to a lesser extent. This can also be noticed in figure 3.1a and 1b.

Based on 1130 observations, the course of RA for men and women was assessed by the GARS and the HAQ as well as by the ADL and IADL subscales. In figure 3.1a the course is depicted for the GARS and the HAQ, and in figure 3.1b for the ADL and IADL subscales. There were a few clear differences between the male and the female scores. Three general patterns can be distinguished in figure 3.1a and 1b: (1) women obtained higher scores than men, (2) the use of the
HAQ resulted in higher disability scores compared to the GARS, and finally (3) the difference in scores between women and men was less on the GARS, than on the HAQ.

Looking a little closer, it appears that women demonstrated a small decrease in scores up to the second year, after which there was a steady and monotonous increase in scores from year to year. For men, however, the course of RA starts with a rather strong decrease in scores up to the third year, followed by a solid increase of scores. When the disease duration increased, men obtained even higher scores on the GARS than women.
On the IADL subscale, women scored significantly higher compared to men (see figure 3.1b). With respect to the ADL subscale this pattern was not found. Men demonstrated a deterioration in functional disability after two and a half years of disease for both ADL and IADL activities.

3.4 Conclusions

The present study has indicated that the psychometric properties of both instruments are excellent and that the correlation between the total scales is high (.84). The reliability coefficients (Cronbach’s $\alpha$ as well as $\rho$) of both instruments are high: all about .93, and a little lower for the ADL and IADL subscales (between .85 and .91). Mokken scale analysis for polychotomous items (MSP) showed that the scalability coefficient (H) for the GARS is slightly better; especially for the ADL and IADL subscales of the GARS. Furthermore, a one factor solution accounted for about 47% of the variance of both instruments, indicating that the underlying construct of both instruments is well represented by only one factor. The same results for the GARS were described elsewhere (Suurmeijer et al. 1994).

However, it was demonstrated that both instruments are not completely comparable. It is true that both instruments have a common focus and some of the items are rather similar whereas other items are included in one instrument and not in the other, and vice versa. Are the HAQ and the GARS, perhaps interchangeable, because of the similarities? We think that the answer cannot be affirmative, right away.

Some very comparable items yielded significant differences, which can be caused by the fact that the items themselves are not completely (literally) the same. Also, the setting (place and time) was different in which the GARS and the HAQ were conducted. For practical reasons, the HAQ was filled out in the hospital, while the GARS was administered at the patients’ home within a period of ten days after the medical examination. Although, health can change within a few days, it is not likely that this time lag has lead to important differences in the patients’ assessment of their own functional status, and, therefore, to differences in HAQ and GARS scores. Possibly more important is the fact that the HAQ was filled out at the hospital, and the GARS at home. As compared to the relaxed domestic situation, the hospital situation may induce feelings of strain and make patients more alert to disease related phenomena, e.g. joint tenderness, and thus may lead to a response in terms of ‘public accounts’ instead of ‘private accounts’ (Cornwell 1984). Furthermore, the GARS may show better precision since it was filled out by interview. However, it is difficult to assess whether or not
these different circumstances will affect the response, and if so, to what extent and in which direction.

Another difference between the two instruments, which has not been discussed before, is the use of devices and/or the assistance of other people by the patients. It has been stated that disability scores were mainly based on the current disease activity, whereas the use of devices had a strong relation with disease duration indicating a certain loss of independence in the course of the disease (Van der Heide et al. 1993). As to the additional information of using devices or assistance from other people as provided by the HAQ, one may wonder whether this information makes the instrument more valuable than an instrument which does not gather this extra information.

As a result of the included measurement of devices, the level of the scores of the HAQ may increase. The use of a certain device leads immediately to an increase of the score (’2’) on the corresponding domain of the HAQ. Consequently, this could be responsible for the differences in scoring pattern of the GARS and the HAQ. The use of devices could mean that these patients have more pain and more joint complaints, which, furthermore, could explain why Steinbrocker classes II and III go together with higher mean scores of the HAQ, when compared to the mean scores of the GARS, and also why the correlations of 1.) Steinbrocker functional classes and 2.) the RAI with the HAQ is stronger when compared with these correlations with the GARS.

With respect to the longitudinal results, the GARS and the HAQ were less consonant. Unexpectedly, both instruments showed a small mean improvement in scores between T1 and T2 for all countries together, although for the HAQ this change is not significant. This improvement might be due to adjustment to medication, or to a first recovery after the onset of disease. After about two to three years both instruments showed a steady decrease in functional ability. The cross-sectional correlation of the GARS and the HAQ with disease duration was not significant, and about zero (-.03 and -.01, respectively). The correlation of the T1-T2 change scores of the GARS and the HAQ with $\delta$RAI were analogous, whereas the correlation of $\delta$OEH was stronger related to $\delta$GARS, which indicates that the GARS is more closely related to the evaluation of the individuals, with respect to their feelings of health.

The effect sizes SRM and RE of both instruments have yielded a somewhat uncertain result. It was interesting to find out that, in general, the SRM’s of the GARS and the HAQ for the groups of patients who improved (on the RAI or the OEH) were much better, as compared to the SRM’s for the patients who worsened between T1 and T2. There were two exceptions on this rule. Compared to the GARS, the HAQ achieved better results (in terms of sensitivity to
change) for patients who worsened according the RAI, while for patients who improved according the OEH, the results of the HAQ were clearly worse. Furthermore, the present study has demonstrated that for patients who improved on the OEH and the RAI, the RE is in favour of the GARS. Especially with respect to the OEH, much better results were achieved by the GARS (RE: 7-15 compared to the standard (1.0), see table 3.5). On the other hand, the HAQ achieved better results for patients who worsened according the RAI between T1 and T2 (RE: .06-.18 vs the standard).

The way in which scores are assigned to the HAQ could have contributed to its low(er) sensitivity to change. Adding up all items and dividing this score by 20, instead of adding up the domain scores dividing by eight, might perhaps have heightened the sensitivity to change of the HAQ. When not all items are used, the HAQ score depends on the domain score, which can be set by different items on different times. In the latter case, the items with the highest score within each of the domains determines the ‘domain score’. And, although the scores on the items may (and will) vary in time, the domain score may still remain the same.

To explore the convergent validity, both instruments are related to Steinbrocker functional classes. No important differences can be demonstrated for the two instruments.

More important results were found with respect to sex differences, which is the final question of this study. The HAQ yields significantly higher scores (.26 on scale level) for women, which also became apparent on many of the separate items. In normal and RA populations, investigations have shown a lower pain threshold in women, which could explain the differences between men and women scores on the HAQ (Da Silva and Hall 1992). However, the GARS contains less items that yield significant results between men and women, and therefore, it seems suitable to subscribe the sex differences encountered in this study to a bias of the HAQ. In scrutinizing the items of the HAQ, it became clear that many items deal with tasks in which physical strength plays an important role (get down a 5lb object from just above your head, open car doors, milk carton, jars etc.). Because men are assigned with more physical strength than women in general, they can perform the tasks that are mentioned in the HAQ more easily. Therefore, even in a nonrestricted population, the HAQ probably will yield sex differences. The GARS, on the other hand, contains only a few items on which women and men score in a different way. As a result, it can be concluded that the GARS is less sensitive for sex differences. This conclusion is supported by the additional regression analysis that was performed, of sex on the GARS and HAQ score, respectively, while controlling for the RAI. Sex had a significant contribution in the model for the HAQ, and not in the model for the GARS.
Another explanation for sex differences on specific items can be embedded in the fact that women and men have culturally different role expectations. For example, women may perhaps pay more (often) attention to the appearance of their feet and toenails compared to men. And, therefore, they might be more often confronted with restrictions of the body, which can result in a higher degree of difficulty as measured by the GARS or HAQ. Furthermore, these culturally determined expectations may lead to different role experiences, leading to more realistic appraisals of the ability to perform certain tasks. Examples of tasks with different culturally determined expectations, and thus experiences, for men and women are housekeeping, making the beds, washing and ironing, and perhaps also shopping and cutting meat. Thus, the reason for the differences in female and male scores on the HAQ could be due to the fact that the items of the HAQ tap more role differences compared to the items of the GARS.

However, the differences between men and women on the GARS and the HAQ may refer to real differences between the sexes in the course of the disease. The courses for women and men, as depicted in figure 1, demonstrate real different patterns for the two sexes, and not just a difference that can be expressed by a constant. The differences between men and women are of such magnitude that the use of the HAQ only is acceptable when sex differences are taken into account. Finally, if the HAQ scores are elevated by the incorporated measurement of the use of devices, and if the HAQ is more sensitive for sex differences, and, moreover, if women express more easily pain in the treatment situation, then this should find expression on the RAI, when these scores on the RAI decrease. This supposition was confirmed by the regression analysis of sex and the RAI on the HAQ (significant) and on the GARS (not significant). Furthermore, a greater SRM and RE were found for the HAQ when the scores on the RAI decreased. While this was either not the case when the scores on the RAI improved, or when the change scores on the OEH were used. Perhaps, this idea is also supported by the fact that there were almost no differences in the RE of the GARS and the HAQ when the scores on the OEH decreased, while, by an increase on the OEH, the GARS achieved much better results in terms of RE and SRM.

\(^2\) Initially, the items of the GARS were formulated in Dutch. Next, the items were translated from Dutch into English by a native English speaker who also mastered the Dutch language. Then, the items were re-translated from English to Dutch, this time by a native Dutch speaker, who also mastered the English language. Different outcomes were discussed. The English version of the GARS was sent to all EURIDISS participants, who followed the same translation procedure.