University of York Department of Health Sciences

Measurement in Health and Disease

Exercise: Validity of the PI HAQ

Read the attached paper ‘Measuring the meaning of disability in rheumatoid arthritis: the Personal Impact Health Assessment Questionnaire (PI HAQ)’ and answer the questions. Ignore the references to Cronbach’s α, which we shall do later in the course.

1. What is ‘face validity’ and how would Study 1 establish it?
2. What is ‘content validity’ and in what sense would Study 2 establish it?
3. In the results for Study 2, what can we deduce about the distributions of the long and short value scales?
4. In Study 3, why did they use Spearman’s ranked product-moment correlation coefficient and Wilcoxon’s signed rank test to assess short term reliability? What would the Wilcoxon test tell us?
5. In the results for Study 3, the authors say that ‘Patients who gave identical value scores at entry and exit had given a range of scores (that is, had not simply ticked the maximum score to every domain each time)’. Why is this important?
6. What is the difference between ‘criterion validity’ and ‘construct validity’?
7. How does Study 4a assess ‘construct validity’?
8. In Study4a, why is it important that the value scale was independent of the level of disability, clinical status, psychological status, and personality, and that both values and change in values were independent of recent change in any variable? What aspect of validity does this address?
9. In the results for Study 4b, in what sense do the authors us the term ‘discriminant validity’?
10. How does Study 4b assess ‘criterion validity’?

EXTENDED REPORT

Measuring the meaning of disability in rheumatoid arthritis: the Personal Impact Health Assessment Questionnaire (PI HAQ)

S Hewlett, A P Smith, J R Kirwan

Ann Rheum Dis 2002;**61**:986–993

Background: Measurement of disability in rheumatoid arthritis is often used to support treatment decisions and outcome assessments, but is used without reference to the impact of disability on individual patients.
Objective: To develop and validate a scale to measure individual values for functions, which is used to weight the level of an individual patient's functional loss and thus calculate the personal impact of disability.
Methods: In four linked studies, first the phraseology for values was explored to develop a stem question for the value scale couched in terms patients understand (face validity). Then short and long versions of the value scale were compared (content validity) and tests of internal consistency and short term reliability undertaken (criterion validity). Finally, the value scale was examined for long term reliability and agreement with expected variables (criterion and construct validity), after which personal impact scores were calculated and their construct validity examined.
Results: Patients understand the concept of values, and a positively phrased stem question was developed for the value scale, for which a short version was reasonably equivalent to a long version. The value scale was reliable over one week (96% changed by <1 point) with positive interitem correlation. Reasonable six and 12 month reliability was shown (52% changed by <0.5 points), and the value scale was independent of disability and clinical, psychological, personality, and social support variables. Personal impact scores were then calculated by using the value scores to weight disability scores. Impact scores varied widely between patients of similar disability. Personal impact for disability showed convergent validity with dissatisfaction with disability, perceived increase in disability, increased disease activity, worse psychological status, low social support, and time trade off for disability. It discriminated between patients with low and high dissatisfaction with disability, life satisfaction, depression, pain, and helplessness.
Conclusion: This individualised personal impact scale should lend meaning to disability scores, improving the interpretation of clinical and research data.

See end of article for authors' affiliations

Correspondence to:
Dr S Hewlett, University of Bristol Academic Rheumatology, Bristol Royal Infirmary, Bristol BS2 8HW, UK;
Sarah.Hewlett@bristol.ac.uk

Accepted 16 April 2002

Outcome in rheumatoid arthritis (RA) used to be measured primarily by the progress of disease processes (for example, bony erosion or C reactive protein), but over the past 20 years it has included measurement of clinical variables deemed important to the patient, supported by validation of reliable measures of physical and emotional function.[1 2] However, evidence is accumulating that measuring the "facts" of disability alone may be insufficient to understand the personal effect of limited activity on the patient. Relatively modest or even poor associations have been reported between dissatisfaction with disability and level of disability,[3–5] between calculated change and patient perceived change in disability,[6 7] and between clinician and patient assessment of disability.[8 9]

The value that a person places on the ability to perform a particular activity might influence the personal impact of those physical limitations. For example a strong value (an "enduring belief that something is personally or socially preferable")[10] for showering rather than bathing might lead a person with difficulties getting into the bath to experience little impact from their limitations. It would follow that small changes in activities held in high value might have a greater personal impact than even large changes in activities of little value, which might account for some of the discrepancies reported above. It has been shown that stopping ≥10% of valued activities is a strong predictor of later depression in RA.[11] Therefore the ability to capture the impact of disability, in addition to the "facts" of disability would place disability within the context of its meaning for the individual person, allowing more accurate interpretation of data. One method of calculating impact is to weight level of disability by the value for that activity.[12]

In the revised WHO classification disability has been replaced by "activity limitation" and handicap by "participation limitation", both of which may be restricted in nature, duration, or quality.[13] Movement occurs between categories and may affect, or be affected by, contextual factors (environmental or personal). Although this model deals with consequences of health conditions such as limitations in activity and participation, there may also be emotional consequences (for example, helplessness, depression, frustration). Conceivably, the combination of the three WHO categories (limitations in body, activity, and participation) and their interaction with environmental and personal factors leads to the *overall consequence* of health conditions, or personal impact.

If it is difficulty in valued activities that represents the concept of personal impact, then the simplest approach to measuring this might be to use a population mean value to weight

Abbreviations: ADLs, activities of daily living; HAQ, Health Assessment Questionnaire; mHAQ, modified HAQ; PI HAQ, personal impact HAQ; RA, rheumatoid arthritis specific activities. However, the use of mean values assumes concordance between and within the views of population and patient groups, but the evidence questions this assumption.[14 15] It has been shown that patient, professional, and healthy control values for disability items are discordant and that patient views vary widely[12] even when using a well validated disability scale (Health Assessment Questionnaire (HAQ)).[1] These data argue the case for using individual rather than mean values for disability. A method of using individual values as weights has been developed in arthritis,[16 17] but the lengthy interview format precludes postal research. However, it has been shown that the HAQ contains 70% of the disability items important to patients with RA, and that no HAQ item is consistently rated by patients as being of no importance.[12] Therefore the HAQ, a well respected and commonly used disability scale,[18 19] would be an appropriate tool for measuring individual values for disability in RA. These values could then be used to weight the HAQ items and calculate the personal impact of disability.

AIMS

This study aimed at developing and validating a method of calculating the personal impact of disability[20] using recommended methodology.[21] The specific aims of the four studies are firstly, to explore phraseology for values in order to develop a stem question for the value scale couched in terms patients understand (face validity); secondly, to compare short and long versions of the value scale (content validity); thirdly, to examine its internal consistency and short term reliability (criterion validity); and fourthly, to examine long term reliability of the value scale and agreement with expected variables (criterion and construct validity). After validation of the value scale, individual patient's values are used to weight their disability scores on the HAQ, resulting in personal impact scores (PI HAQ). The construct validity of the PI HAQ scores is then examined.

METHODS

The study group comprised consecutive patients with confirmed RA[22] from teaching or district general hospitals and had local research ethics committee approval.

Study 1: Development of the stem question

Twenty eight inpatients and 31 rheumatology health professionals were invited to complete a questionnaire about their opinions on phraseology for values for disability (administered by interview to the patients). Professionals (doctors, physiotherapists, occupational therapists, nurses, and psychologists) were asked, "I am trying to capture the value that functions hold for patients, what sort of phrase would you use?" with three prompts (value, importance, upset) and an open option. Patients were asked, "When I talk about the idea of the importance of being able to do something, what sorts of words make sense to you—something being important, or being valuable, or do you think about it in a different way altogether? Perhaps about how much it upsets you not being able to do something? Or how much it bothers you? Or another way still?" Pilot work suggested patients did not respond well to an open question, therefore prompts were based on the professional and patient pilot results (important, valuable, means a lot, upsets me, bothers me, annoys me) plus an open option. A pilot value scale based on the HAQ was then designed.

Study 2: Short versus long versions

Forty eight outpatients were invited to complete a 20 and an eight item version of the pilot value scale twice, one week apart, in random order. The 20 item version contains the 20 HAQ activities of daily living (ADLs) while the eight domain version is based on the eight item modified HAQ (mHAQ).[3] The mHAQ uses only one ADL from each of the eight categories to represent that category. To try to broaden the questions but maintain the brevity of the value scale, some of the mHAQ eight ADLs were expanded slightly to try to capture more of the HAQ functions for that category (for example, the HAQ asks two questions about walking and stairs, the mHAQ selects walking as the single ADL, the value scale expands this to a domain question on "walking, including going up and down stairs"). The stem question asks, "How important is it to you this week to be able to do the following things?" with the responses "not at all important, a little bit important, quite important, or very important" (0–3). As the final impact calculation incorporates the HAQ score, this timescale is identical to the HAQ. For validation purposes only, value scores were summed and then calculated as a percentage of the maximum possible score (60 for the 20 ADL scale and 24 for the eight domain scale) and the percentage scores compared using Pearson's correlation coefficient. As correlation measures the strength of a relationship rather than agreement, Bland and Altman methods of assessing agreement were also used.[23] The mean percentage difference between the eight domain and 20 ADL scales is calculated. As the mean of both scale versions is the most likely approximation of the "true" measure, the differences between the percentage scores of the two versions for each patient are plotted against the mean of their short plus long percentage scores, showing how far from the "true" measure each patient's difference between short and long scales lies. As it is not known which is the true measure, the clinical significance of these differences is considered.

Study 3: Short term reliability, internal consistency

Thirty one patients were invited to complete the final, eight domain value scale at entry and one week. For validation purposes only, the eight domain scores were summed and divided by eight (in the manner of the HAQ) to obtain an average domain score. Spearman's ranked product-moment correlation coefficients and Wilcoxon's signed rank test were used to assess short term reliability, and a correlation matrix and Crohnbach's α used to examine internal consistency. For a correlation of r_s=0.5 to be significant at the 1% level, 24 patients would be required.[24]

Study 4a: Long term reliability and construct validity of the value scale

One hundred and nine patients were invited to complete the value scale and HAQ at 0, 26, and 52 weeks, together with questionnaires on perceived change in disability, dissatisfaction with disability, psychological status, and social support, plus measures of clinical status (10 cm visual analogue pain scale, early morning stiffness, and articular index).[25] Personality variables were measured at entry while time trade off (TTO) for disability was explored at week 52.

Questionnaires

Value scale: See appendix 1. For validation purposes only, the eight domain scores are summed and divided by eight.

Disability: The HAQ is used throughout.[1]

Perceived change in disability: The mHAQ[4] measuring perceived change over six months (easier, no different, more difficult, −1 to +1) with the eight ADLs expanded as for the eight domain questions on the value scale. Scores summed (range −8 to +8).

Dissatisfaction with disability: The mHAQ[4] for dissatisfaction (very satisfied to very dissatisfied, 0–3) with the same expanded domains. Scores summed and divided by eight (range 0–3).

Anxiety and depression: Hospital Anxiety and Depression Scale (range 0–21), in which a score ≥11 means a probable case.[26]

Table 1 Demographic data for patients at entry. Mean (SD)

Variable	Study 1 (n=25)	Study 2 (n=45)	Study 3 (n=24)	Study 4 (n=93)	Comment
Age	58.6 (16.6)	60.2 (11.8)	52 (16.2)	60 (10.8)	
Male: female	8:17	9:36	7:17	33:60	
Disease duration	16.7 (10.8)	13.9 (9.5)	14.2 (8.4)	12.1 (10.5)	
Disability (HAQ, 0–3)	2.16 (0.36)	1.77 (0.61)	1.83 (0.88)	1.42 (0.71)	
Pain (0–10)				4.17 (2.46)	
Early morning stiffness (min)				56 (65)	
Tender joints (0–28)				5.5 (4.9)	
Swollen joints (0–28)				6.5 (4.5)	Possible/probable cases 38.8%
Anxiety (0–21)				6.91 (4.16)	Possible/probable cases 20.4%
Depression (0–21)				4.85 (3.30)	
Helplessness (5–30)				16.38 (5.16)	Normal population mean 24.4
Satisfaction with life (5–35)*				21.60 (7.16)	Normal population mean 32.9
Social support					
Overall (0–40)*				31.88 (7.04)	
Appraisal (0–10)*				8.13 (2.30)	
Belonging (0–10)*				8.35 (2.30)	
Self esteem (0–10)*				6.77 (2.16)	
Tangible (0–10)*				8.82 (1.58)	Normal population mean 21
Optimism (0–32)*				19.2 (4.38)	Normal population mean 9
Neuroticism (0–24)				10.52 (4.92)	

*Reverse scored, low score is poor.

Helplessness: Five item subscale of the Arthritis Helplessness Index (range 5–30), in which ⩾20 means high helplessness.[27]

Life satisfaction: Satisfaction with Life Scale (range 5–35), where a low score means low satisfaction.[28]

Social support: Interpersonal Support Evaluation List, comprising subscales for appraisal, belonging, self esteem, and tangible support, (range 0–10 each or 0–40 overall), where a low score means low support.[29]

Optimism/pessimism: Life Orientation Test (range 0–32), where a low score means low optimism.[30]

Negative personality: Neuroticism scale of the Eysenck Personality Inventory (range 0–23), where a high score means high neuroticism.[31]

Time trade off for disability: This question asked the patient to consider only the physical difficulties relating to their arthritis. The patient's completed HAQ form was discussed with them and they were asked to consider an imaginary question in which they could trade years of healthy life to be rid of their physical difficulties immediately. A standard example was offered of how many years they would lose if they traded half their remaining life (assumed to age 85) and the number of years traded was calculated as a percentage of years to age 85.

Study 4b: Construct and criterion validity of PI HAQ scores

The value scale is not designed to be used alone, but only as a weighting tool. Therefore after validation of the value scale, the PI HAQ scores were calculated. Each of the eight HAQ category disability scores (0–3) is weighted by its corresponding domain value score (0–3)—for example, disability for hygiene weighted by value for hygiene. The resulting eight value weighted scores are summed and divided by eight to yield the PI HAQ score, ranging from 0 to 9 (no personal impact to great personal impact). Analysis in studies 4a and 4b was by Spearman's ranked product-moment correlation coefficients, and the mean correlation coefficients over the three visits were calculated using Fisher's Z transformations for correlation coefficients.[32] Differences between low and high scoring patient groups were examined using Mann-Whitney U tests. In the absence of published data on values for function in RA, no power calculation was possible, but 100 subjects would be likely to produce a good spread of disease and psychological states.

RESULTS

Table 1 presents the demographic data for all four studies.

Study 1: Development of the stem question

Twenty five patients and 25 professionals participated. Patients made 64 comments (during the interviews or by selecting or suggesting phrases), of which 72% reflected negative phrases for values, whereas 69% of the 29 comments or suggestions made by professionals were positively phrased (table 2). From these data two value scales were designed, with either a positively phrased stem question (importance of activities) or a negatively phrased one (upset over activities). Both versions were subjected to the same four validation studies,[20] but validation of the "importance" version was stronger, therefore for this and other reasons described in the "Discussion", the "upset" version was discontinued (in the interests of space, only the "importance" scale validation is presented here). Short and long versions of the importance value scales, based on the HAQ and mHAQ were then piloted.

Table 2 Phraseology of the value of disabilities (study 1), n=25 in each group

Phraseology	Patients	Professionals	Total
Offered in questionnaire			
Valuable	1	1	2
Important/means a lot	17	19	36
Upsets/bothers	17	6	23
Annoys	16	N/A	16
Free comments or phrases			
Frustrates	8	1	9
Cross	1		1
Agitated	1		1
Drives me mad	1		1
Makes me swear	1		1
Cussed nuisance	1		1
Affects quality of life		1	1
How do you cope?		1	1
Summary			
Positive (No (%))	18 (28)	20 (69)	38 (41)
Negative (No (%))	46 (72)	7 (24)	53 (57)
Neutral (No (%))	0	2 (7)	2 (2)

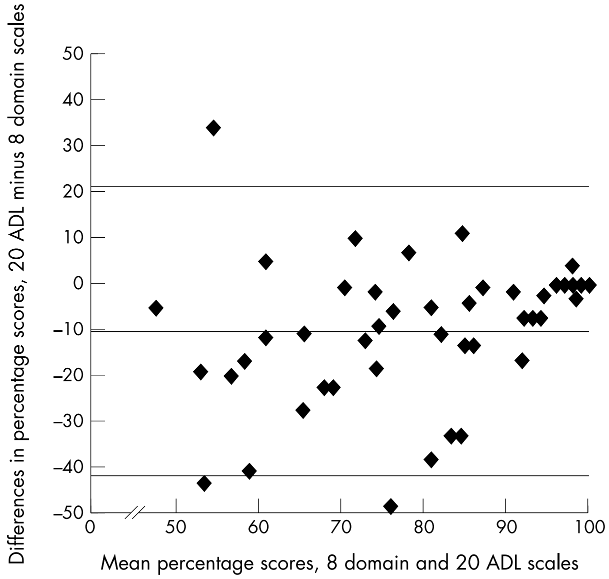

Figure 1 Percentage difference of eight domain and 20 ADL scales plotted against mean percentage scores of both scales (n=45).

Study 2: Short versus long versions

Forty five patients participated in study 2. Scores for both short and long value scales were high with a mean of 20.04 for the eight domain scale (SD 3.64, range 9–24 of a possible 0–24) and a mean of 43.91 for the 20 ADL scale (SD 11.22, range 19–60 of a possible 0–60). Nine patients (20%) gave identical value scores for long and short scales, with an overall correlation of r_s=0.584 (p<0.01). The mean of both scales was assumed to be the best approximation of the "true" measure,[23] and the differences in the percentage scores of the two versions were plotted against the mean of the two percentage scores added together (eight domain plus 20 ADL) (fig 1). The mean difference between the two scales was 10.3% (SEM 2.349, CI 5.7% to 14.9%), with long scales yielding lower values. As it is not known which scale is the "true" measure, only that there is a mean 10% difference, the clinical significance of the difference should be considered. In patients scoring 45 out of 60 on the long value scale this difference would give short scale scores not of 18, but of 19.4–21.6 out of 24. It would seem unlikely that this is sufficient to be clinically important, therefore only the short eight domain scale is further validated. A small one to one interview study of comprehension[19] in nine patients showed only one minor misunderstanding by one patient and so the final version contains the addition of the word "yourself" to the stem question to aid clarity (appendix 1).

Study 3: Internal consistency

Twenty four patients participated in study 3. Value scores were relatively high with a mean of 2.52 (range 1.5–3 of a possible 0–3). Of the eight domains, only hygiene was always valued at the maximum. All domains correlated positively with each other (p<0.05 to p<0.001), although correlations for the hygiene domain were lower and did not always reach statistical significance. Crohnbach's α was 0.895.

Study 3: Short term reliability

No change was made in value scale scores over one week by nine (38%) patients, a further 12 (50%) changed by <0.5 points in either direction and overall, 23/24 (96%) patients changed by <1 point. Patients who gave identical value scores at entry and exit had given a range of scores (that is, had not simply ticked the maximum score to every domain each time). The most stable domain was hygiene (21/24 (88%)

Table 3 Correlation between value scale, current variables, and change in variables (study 4), n=93

	Correlation between:		
Variable	Value scale and variables (over 3 visits)*	Value scale and change in variables (0–6, 6–12 months)*	Change in value scale and change in variables (0–6, 6–12 months)*
Age	−0.129		
Disease duration	0.024		
Disability			
Disability	−0.180	0.025	−0.015
Perceived change	0.043		
Dissatisfaction	0.005	0.054	0.152
Disease activity			
Pain	−0.098	0.014	0.015
Swollen joints	0.054	0.107	0.006
Tender joints	−0.007	0.017	0.093
EMS	−0.068	−0.075	−0.054
Psychological status			
Anxiety	0.007	0.036	0.025
Depression	0.106	−0.063	−0.114
Helplessness	−0.024	0.062	0.111
Life satisfaction†	−0.090	−0.032	0.028
Social support			
Overall†	−0.112	−0.028	−0.017
Appraisal†	−0.116	−0.048	−0.075
Belonging†	−0.127	−0.069	0.016
Self esteem†	−0.090	0.046	−0.039
Tangible support†	−0.015	−0.060	0.048
Single visit correlation			
Personality			
Optimism†	−0.129		
Neuroticism	0.035		

All non-significant
*Z transformation to calculate correlation over several visits; †reverse scored, low score indicates worse status

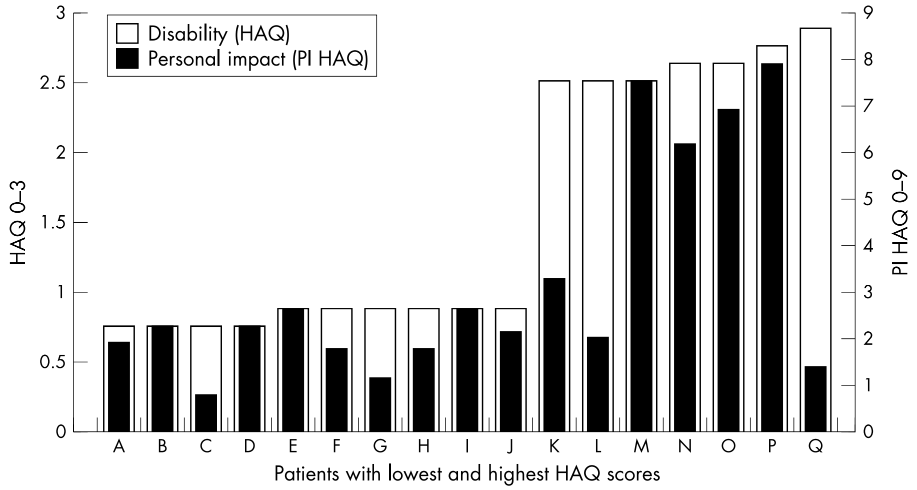

Figure 2 Disability and personal impact for patients with low and high disability (study 4).

unchanged). Entry and exit scores correlated at r_s=0.63 (p<0.001), with no significant change in scores over one week (Wilcoxon signed rank test).

Study 4

One hundred and two patients agreed to participate and 93 completed study 4 (two could not complete an HAQ unaided, two died, five withdrew). At entry the range of values for disability was broad (1.125–3), with 30% of patients giving maximum value scores and the remaining 70% having a normal distribution (1.125–2.875). There were no significant differences in any clinical or psychological variable between high scorers and those who valued functions less (Mann-Whitney U test). Patients had a wide range of functional levels (table 1), but actual change in disability was minimal, increasing by a mean of 0.02 of an HAQ score over six months and 0.08 over 12 months (range −1.125 to +1.625). Perceived change in disability over the previous six months was consistently that disability was increasing (55% of patients at 0–6 months, 57% at 6–12 months) with few perceiving an improvement (12%, 13%). Dissatisfaction with disability ranged from very satisfied (0–0.25, 17/93 (18%)) to very dissatisfied (2.875–3, 4/93 (4%)).

Study 4a: Long term reliability of the value scale

No change in value scores was made by 33% of patients over six months (31% over 12 months), while a further 48 (52%) changed by <0.5 of a score in either direction at six and 12 months. The most stable domain was again hygiene (84% and 85% unchanged over six and 12 months). Correlation between

Table 4 Correlation between PI HAQ scores and variables over three visits (study 4), n=93

	PI HAQ at:			
Variable	0 Months	6 Months	12 Months	Mean correlation†
Disability				
Disability	0.887***	0.843***	0.895***	0.877***
Perceived change		0.388***	0.451***	0.420***
Dissatisfaction	0.427***	0.560***	0.714***	0.579***
Disease activity				
Pain	0.354***	0.460***	0.594***	0.475***
Swollen joints	0.253**	0.206*	0.242**	0.234**
Tender joints	0.295**	0.369***	0.319***	0.328***
Early morning stiffness	0.399***	0.365***	0.316**	0.360***
Psychological status				
Anxiety	0.247*	0.371***	0.337***	0.319**
Depression	0.335***	0.426***	0.507***	0.425***
Helplessness	0.404***	0.565***	0.527***	0.502***
Satisfaction with life‡	–0.439**	–0.286**	–0.363***	–0.360***
Social support				
Overall social support‡	–0.215*	–0.180	–0.285**	–0.230*
Appraisal of support‡	–0.217*	–0.108	–0.193	–0.170
Belonging‡	–0.165	0.047	–0.183	–0.100
Self esteem‡	–0.253**	–0.337***	–0.331***	–0.310**
Tangible support‡	–0.063	–0.010	–0.204*	–0.090
Personality				
Optimism‡	–0.157			
Neuroticism	0.199			
Age	0.126			
Disease duration	0.230*			
Time trade off (n=67)			0.269*	

*p<0.05; **p<0.01; ***p<0.001; †Z transformation to calculate correlation over several visits; ‡reverse scored, low score indicates worse status.

Table 5 Divergent validity for different disease groups for PI HAQ scores (study 4 entry data), n=93

Low and high scoring groups	n	PI HAQ median
Disability dissatisfaction		
Low 0–1.5	73	2.88
High 1.625–3	20	4.50*
Pain		
Low 0–5	59	2.63
High 5.1–10	34	3.75*
Depression		
Low 0–7	74	3.00
High 8–21	19	4.50**
Helplessness		
Low 0–19	66	2.75
High 20–30	27	4.38***
Satisfaction with life†		
Low 0–20	35	4.50
High 21–35	58	2.69***
Social support overall†		
Low 0–30	28	4.44
High 31–40	65	3.00*
Tangible social support†		
Low 0–7	19	3.63
High 8–10	74	3.13

*p<0.05; **p<0.01; ***p<0.001; †reverse scored, low score indicates worse status

the value scales was consistent with that seen over one week (r_s=0.636 over 0–6 months and r_s=0.610 over 0–12 months, p<0.001).

Study 4a: Construct validity of value scale

The value scale measuring personal values for disability was shown to be not only independent of the level of disability but also independent of clinical status, psychological status, and personality (table 3). In addition, both values and change in values were independent of recent change in any variable (table 3). As the value for disability scale appeared both reliable and independent of confounding variables, it was then used to weight the HAQ scores in order to calculate the personal impact of disability. The construct and criterion validity of the PI HAQ scores was then examined.

Study 4b: Construct validity of PI HAQ scores

PI HAQ scores were normally distributed (0–8 out of 0–9). There were clear differences in impact between patients with similar levels of disability, and similar levels of impact between patients of widely differing disability (fig 2).

The personal impact of disability was associated with dissatisfaction with that disability, perceived increase in disability, more active disease, worse psychological status, and reduced life satisfaction (table 4). As expected, impact was strongly related to level of disability (mean r_s=0.877, p<0.001), but disability is a constituent of the impact score. Discriminant validity was tested by dividing patients into two groups (low and high scorers) for each of those variables where one would expect the impact of disability to show differences. Appropriate and significant differences were found for the impact of disability between those with low and high levels of dissatisfaction with disability, pain, depression, life satisfaction, helplessness, and social support (p<0.05 Mann-Whitney U test) (table 5), with median scores for greater impact being associated with worse status.

Study 4b: Criterion validity of PI HAQ scores

There is currently no validated "gold standard" measure for the impact of disability in arthritis, therefore the best available comparator (a utility measure, TTO) was used. The TTO question was completed by 67 (72%) patients as two patients were unable to understand the concept and it was not administered to 24 (one aged >80 years; one no disability; two declined; four depressed/bereaved; 16 completed final questionnaires by post). Those who were not asked the TTO question were not significantly different from those who were, with the exception of greater satisfaction with disability (median 1.0 *v* 1.125, p<0.05) and greater helplessness (Arthritis Helplessness Index median 19.5 *v* 16, p<0.01, Mann-Whitney U test).

Willingness to trade was bimodal, with 58% of patients declining to trade while those prepared to trade showed a normal distribution (1–70% of remaining years). TTO to be rid of disability was not associated with the level of that disability, nor with the number of years available to trade (that is, age) but was associated with the personal impact of disability (r_s=0.333, p<0.01). There was a significant difference between traders and non-traders for PI HAQ scores (median 4.44 *v* 3, p<0.05, Mann Whitney U test), supporting discriminant validity for the PI HAQ.

DISCUSSION

In developing a measure of the personal impact of disability the aim was to create a questionnaire using a language for values with which patients could identify. Patients understood the concept of values and used both positive and negative phrases (study 1). Although patient preference was for the negatively phrased "upset" version, it was decided to use the more positive "importance" version to capture values. This was decided not only because the "importance" version had a stronger validation (the "upset" version was abnormally distributed, associated with personality, and did not correlate with the gold standard measure) but also for conceptual reasons. Conceptually, importance and upset are not necessarily opposite ends of the same continuum and may measure different things, while "upset" may itself reflect an emotional impact of disability, rather than the value of activities. A questionnaire asking patients with newly diagnosed RA how upset they would be not to be able to perform basic activities would be clinically inappropriate, especially if administered postally. Finally, as patients indicated a strong desire to be asked about the emotions surrounding disability, that deserves the development and validation of a specific emotional tool in its own right. Positive phraseology was clearly understood by patients, was preferred by almost 30%, and subsequent validation showed it to be an appropriate terminology.

Questionnaire fatigue, which affects study recruitment and retention, may be reduced by shortening questionnaires if validation is acceptable (study 2). Comparison of short and long versions over one week meant that clinical status was likely to be stable but recall limited, although it is noted that in this group, lower value scores were underrepresented (for example, compared with study 4). The short scale items were worded slightly differently from the long scale items and the number of subjects relatively small, which may explain the mean 10% difference between long and short scales. As it is unknown which scale represents the true values and as the clinical significance of the difference is doubtful, the shorter version was chosen.

Little is known about the stability of values for disability but by using stable outpatients to test reliability over one week, confounding variables such as changing clinical status were minimised (study 3). In the validation of the HAQ, 93% of patients made <1 point change in their HAQ scores over 0–12 days,[1] which compares with 96% in the value scale over seven days. No significant changes were found (Wilcoxon signed rank test), therefore short term stability appears to be similar to the HAQ. Interitem correlation showed that all domains correlated positively with each other, indicating that they are likely to be measuring associated but not identical concepts. Slightly lower correlations for the hygiene domain were seen, as they were during the HAQ validation,[1] which may be because hygiene is generally given high value scores (88% gave a score of 3), giving little variation to calculate a correlation.

The value scale showed reasonable reliability over six and 12 months (study 4a) and although these levels of reliability might not be adequate for an objective or permanent measure (for example, erosions on *x* ray), they may still indicate adequate reliability in an attitudinal scale, which is, by nature, likely to fluctuate. The longer term reliability of disability values or the scales to measure them is not known, although it has been suggested that health values may change daily,[33] and are reordered in the face of change,[10 34] a concept supported by data showing differences in health values between people with disabilities and non-disabled controls.[12 35] If it is the case that values are reordered then it would be unreasonable to expect better reliability in a value scale than that presented here. In the absence of published work on the stability of values it is uncertain whether this is a feature of the stability of values or of the reliability of the scale and until other validated methods of measuring values are available, it is difficult to quantify the reliability data.

The measurement of values for disability has the potential to be confounded by other personal variables. However, the value scale is independent of all the disability, clinical, psychological, and personality variables measured (study 4a). The personal impact scores calculated from that value scale have good construct validity (study 4b), showing convergent validity with perceived change in disability, dissatisfaction, worse disease and psychological status, reduced life satisfaction, and social support. It also discriminates between groups who score high or low in these variables, where a difference in the personal impact of disability might be expected. TTO for disability is associated with the personal impact of disability but not with having greater disability or more years left to trade. It might be argued that the relationship between PI HAQ and TTO should be stronger, but management of RA includes enabling patients to cope with their disease and several subjects said that they were currently coping well with their problems and it was not necessary to be rid of disability at such a cost (data not systematically collected). This suggestion has been raised by others who found TTO a feasible approach in RA.[36] In addition, although efforts were made to persuade the patient to consider disability alone for the TTO question, this may conceptually have been difficult. Utility measures such as TTO involve hypothetical choices which do not reflect real life decisions, and these issues will be discussed in more detail elsewhere.

The PI HAQ scores calculated in the final study show that patients with similar disability levels have different levels of personal impact arising from that disability. Thus an individual patient's level of disability and the impact of that disability are clearly different entities. For example, two patients scoring 0.75 and 0.875 on the HAQ had PI HAQ scores that differ threefold (0.75 and 2.625, fig 2). This difference appears more marked at higher disability levels where patients with disability scores of 2.5 to 2.875 have impact scores ranging from 1.25 to 8. In some patients high disability can be seen to have less personal impact than in other patients with low disability (HAQ 2.875 with PI HAQ 1.375 versus HAQ 0.875 with PI HAQ 2). These differences in impact are based on individual patient opinion without the imposition of external health professional or general population values, allowing us to avoid making assumptions about the meaning of disability scores. However, for a better interpretation of the PI HAQ scores it must be appreciated that they will be related to disability levels. Thus with a difficulty (HAQ) level of 1.5, even if all functions are highly valued (at 3), then the maximum impact score can only be 4.5. Greater experience and use of the PI HAQ will improve awareness of how scores should be interpreted.

Professionals, relying on "objective" disability measures alone cannot evaluate the meaning of disability levels for patients. In clinical use the PI HAQ may identify the hidden unexpected low or high impact of disability, which might influence clinical decisions to give different priority to problems than might have been given using disability data alone. Currently, treatments are deemed successful if they change disability scores, but this assumes that this change makes a difference that matters to the patient, and data presented here show that patients with similar disability levels have different opinions on the impact of that disability. For clinical trials to reflect the impact of treatment on patients rather than change in disability, impact measures could also be reported. However, to rely solely on patient self reported impact would be restrictive as patients may not be aware of the hidden consequences of disability (for example, difficulty rising from a chair may be of minor importance, but the potential consequence is progression to an inability to use the toilet independently). Therefore it is not suggested that measuring the personal impact of disability should replace disability measurement, rather that the PI HAQ should complement measures of disability and aid interpretation.

The PI HAQ measures impact of disability, but does not deal with the impact of disability in relation to the impact of other symptoms, illnesses, or life events—that is, what slice of the cake does disability represent? In addition, it would be helpful to assess the concept of this impact scale against a quality of life measure and the standard global visual analogue scale of patient opinion. These areas should be explored in future longitudinal studies by evaluating the performance of the PI HAQ against such measures[37 38] to see whether the loss of valued activities (impact) predicts psychological distress.[11] The present observational study was of patients with relatively little change in disability over one year. To show sensitivity to change the PI HAQ will shortly be tested in patients undergoing a dynamic intervention to alter functional status, such as a major joint replacement, physiotherapy, or a change in drug treatment. If the PI HAQ is shown to be sensitive to change when disability changes, then further work is possible, such as determining whether an intervention which is not directed at changing underlying impairment alters the personal impact of disability. For example, occupational therapy considers personal and environmental issues about dealing with disability, while patient education aims at altering self efficacy, behaviour, and coping. Both these interventions might alter the impact of disability. The concept of measuring the personal impact of disease could also be explored by using values to weight the severity of other symptoms (for example, fatigue) or in other arthritides (for example, psoriatic arthritis).

Disability is an essential patient centred outcome measure in RA, and the HAQ[1] is probably the best validated unidimensional scale to date. Measurement of disability alone requires interpretation by a health professional of the meaning of that disability, an interpretation which is not necessarily accurate. Simultaneous measurement of the impact of disability using the PI HAQ might better enhance our understanding of both the effects of RA and the efficacy of treatments.

ACKNOWLEDGEMENTS

We thank statisticians Dr Barnaby Reeves, Rosemary Greenwood, and Dr Peter Blair for their advice and support in this analysis, the patients for participating, and the Arthritis Research Campaign for financial support.

APPENDIX 1: PI HAQ

Importance of abilities

These questions ask about how *important* it is to you to be able to do different things yourself. For example, you might feel it is not important that you do the gardening yourself—it could be done by someone else. On the other hand, you might feel it is important to do the gardening yourself, even though it could be done by someone else.

How *important* is it to you this *week* to be able to do the following things *yourself*?

1 Carry out the tasks involved in dressing and grooming, including tying shoelaces, doing buttons, and shampooing your hair?
2 Carry out the sort of tasks that involve getting up (for example, from a chair or bed)?
3 Carry out the tasks involved in preparing and eating food?
4 Walk, including flat ground and stairs?
5 Carry out the tasks involved in personal hygiene, including using the bath and toilet?
6 Carry out the sort of tasks that involve reaching up and bending down?
7 Carry out the sorts of tasks that involve gripping things (for example, turning taps)?
8 Carry out general activities, such as light gardening, shopping, housework?

Each is scored as "*Not at all important*", "*A little bit important*", "*Quite important*", or "*Very important.*

Authors' affiliations

S Hewlett, J R Kirwan, University of Bristol Academic Rheumatology, Bristol Royal Infirmary, UK
A P Smith School of Psychology, Cardiff University, UK

REFERENCES

1 **Fries JF**, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
2 **Meenan RF**, Gertman PM, Mason JM. Measuring health status in arthritis: the Arthritis Impact Measurement Scale. Arthritis Rheum 1980;23:146–53.
3 **Pincus T**, Summey JA, Soraci SA, 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.
4 **Giorgino KB**, Blalock SJ, DeVellis RF, DeVellis BM, Keefe FJ, Jordan JM. Appraisal of and coping with arthritis related problems in household activities, leisure activities and pain management. Arthritis Care Res 1994;7:20–8.
5 **Hewlett S**, Young P, Kirwan JR. Dissatisfaction, disability and rheumatoid arthritis. Arthritis Care Res 1995;8:4–9.
6 **Fischer D**, Lorig K, Laurent D, Holman H. Patient assessment of clinical change is a reliable and sensitive measure and is not unduly biased by baseline patient expectations [abstract]. Arthritis Rheum 1995;38:S178.
7 **Hewlett S**, Kirwan JRK. Discrepancies between actual and perceived change in function in rheumatoid arthritis are not a function of memory [abstract]. Br J Rheumatol 1998;37(suppl 1):177.
8 **Berkanovic E**, Hurwicz ML, Lachenbruch PA. Concordant and discrepant views of patients' physical functioning. Arthritis Care Res 1995;8:94–101.
9 **Kwoh CK**, O'Connor GT, Regan-Smith MG, Olmstead EM, Brown LA, Burnett JB, *et al*. Concordance between clinician and patient assessment of physical and mental health status. J Rheumatol 1992;19:1031–7.
10 **Rokeach M**, ed. *Beliefs, attitudes and values*. London: Jossey-Bass, 1972:chapter V.
11 **Katz PP**, Yelin EH. The development of depressive symptoms among women with rheumatoid arthritis. Arthritis Care Res 1995;38:49–56.
12 **Hewlett S**, Smith AP, Kirwan JR. Values for function in rheumatoid arthritis: patients, professionals and public. Ann Rheum Dis 2001;60:928–33.
13 **World Health Organization**. International classification of functioning, disability and health. http://www.who.int/classification/icf/intros. 2002
14 **Rothwell PM**, McDowell Z, Wong CK, Dorman PJ. Doctors and patients don't agree: cross-sectional study of patients' and doctors' perceptions and assessments of disability in multiple sclerosis. BMJ 1997;314:1580–3.
15 **Lubeck DP**, Yelin EH. A question of value: measuring the impact of chronic disease. The Millbank Quarterly 1988;66:444–64.
16 **Bell MJ**, Bombardier C, Tugwell P. Measurement of functional status, quality of life and utility in rheumatoid arthritis. Arthritis Rheum 1990;33:591–601.
17 **Buchbinder R**, Bombardier C, Yeung M, Tugwell P. Which outcome measures should be used in rheumatoid arthritis clinical trials? Arthritis Rheum 1995;38:1568–80.
18 **Ramey DR**, Raynauld JP, Fries JF. The Health Assessment Questionnaire 1992: status and review. Arthritis Care Res 1992;5:119–29.
19 **Kirwan JR**, Reeback JS. Stanford Health Assessment Questionnaire modified to assess disability in British patients with rheumatoid arthritis. Br J Rheumatol 1986;25:206–9.
20 **Hewlett S**. Values, disability and personal impact in rheumatoid arthritis. Bristol: University of Bristol, 2000. (PhD thesis.)
21 **Tugwell P**, Bombardier C. A methodologic framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982;9:758–62.
22 **Arnett FC**, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, *et al*. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
23 **Bland JM**, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
24 **Goldstone LA**, ed. *Understanding medical statistics*. London: Heinemann, 1983:appendix E.
25 **Fuchs HA**, Brooks RH, Callahan LF, Pincus T. A simplified twenty-eight joint quantitative articular index in rheumatoid arthritis. Arthritis Rheum 1989;32:531–7.
26 **Zigmond AS**, Snaith RP. The Hospital Anxiety and Depression Scale. Acta Psychiatr Scand 1983;67:361–70.
27 **Stein MJ**, Wallston KA, Nicassio PM. Factor structure of the Arthritis Helplessness Index. J Rheumatol 1988;15:427–32.
28 **Diener E**, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985;49:71–5.
29 **Cohen S**. Measuring the functional components of social support. In: Sarason IG, Sarason BR, eds. *Social support: theory, research and applications*. Boston: Martinus Nijhoff, 1985.
30 **Scheier MF**, Carver CS. Optimism, coping and health: assessment and implications of generalised outcome expectancies. Health Psychol 1985;4:219–47.
31 **Eysenck HJ**, Eysenck SBG, eds. *Manual of the Eysenck Personality Inventory*. London: Hodder and Stoughton, 1964.
32 **Moroney MJ**, ed. *Facts from figures*. Middlesex: Penguin, 1956:312–15.
33 **Yelin E**, Lubeck D, Holman H, Epstein W. The impact of rheumatoid arthritis and osteoarthritis: the activities of patients with rheumatoid arthritis and osteoarthritis compared to controls. J Rheumatol 1987;14:710–17.
34 **Keany KCMH**, Glueckauf RL. Disability and value change: an overview and re-analysis of acceptance of loss theory. Rehabil Psychol 1993;38:199–210.
35 **Stensman R**. Severely mobility-disabled people assess the quality of their lives. Scand J Rehabil Med 1985;17:87–99.
36 **Tijhuis GJ**, Jansen SJT, Stiggelbout AM, Zwinderman AH, Hazes JMW, Vliet Vlieland TPM. Value of the time trade off method for measuring utilities in patients with rheumatoid arthritis. Ann Rheum Dis 2000;59:892–7.
37 **Carr AJ**. A patient-centred approach to evaluation and treatment in rheumatoid arthritis: the development of a clinical tool to measure patient-perceived handicap. Br J Rheumatol 1996;35:921–32.
38 **De Jong Z**, Van Der Heijde D, McKenna SP, Whalley D. The reliability and construct validity of the RAQoL: a rheumatoid arthritis-specific quality of life instrument. Br J Rheumatol 1997;36:878–83.