ARE WE MAKING THE MOST OF THE STANFORD HEALTH ASSESSMENT QUESTIONNAIRE?

A. TENNANT, M. HILLMAN, J. FEAR, A. PICKERING* and M. A. CHAMBERLAIN
Rheumatology & Rehabilitation Research Unit, University of Leeds and * North Yorkshire Health, York

SUMMARY

For many years, the Stanford Health Assessment Questionnaire (HAQ) has provided an effective measure of disability. Recently, some debate has emerged about whether or not the HAQ is an 'ordinal' or 'interval' scale. The opportunity to test its level of measurement arose when the scale was applied in a community survey which undertook a two-stage random sample using postal questionnaires to ascertain the health care needs of those with arthritis. The HAQ data are fitted to the Rasch model which tests for the presence of certain desirable characteristics of measurement, e.g. unidimensionality. The fit of the data to the model for those self-reporting rheumatoid arthritis (RA) was adequate. The transformed HAQ score, derived from the Rasch analysis, is compared with the ordinary HAQ (raw) score. This shows that, for those with RA, incremental units of the raw score at the margins of the scale reflect an increasing level of (dis)ability compared to similar units in the centre of the scale. Thus, the traditional HAQ score (range 0-3) is an ordinal score. The findings also indicate that scoring all 20 items may lead to greater sensitivity. Questions are also raised about the construct validity for those with other types of arthritis. For osteoarthritis, the grip item does not appear to belong to the same underlying construct as the other items.

KEY WORDS: Disability outcome, Rasch, HAQ, Rheumatoid arthritis.

For many years, the Stanford Health Assessment Questionnaire (HAQ) [1, 2] has provided a simple and apparently effective measure of disability in numerous studies of arthritis care [3-5]. Recently, one review highlighted over 100 articles utilizing the scale [6]. The HAQ has four dimensions: disability, discomfort and pain, drug side-effects and dollar costs. Most use has been made of the disability dimension, which is the concern of this paper.

The HAQ disability scale has 20 items scoring 0-3 which contribute to eight subscales. The score on an individual item can be adjusted when the response is 'with some difficulty' (scores one point), such that the score can be increased to two points (with much difficulty) if there is evidence of the use of assistive devices, or help needed on any item. Within each subscale, the item of greatest value (most difficulty) gives the score for the subscale. In this way, the HAQ differs from most other disability scales which simply add the sum of item scores within a subscale. The sum of the subscale scores is then divided by eight (or less if not all are completed) to give an overall score which ranges from zero to three. It has been argued that this 'disability index' gives 'a continuous scale' [6].

There has, however, been some debate in this journal about whether or not the HAQ is an 'ordinal' scale, or whether the items combine to 'provide a relatively continuous and scalable variable at the 'interval' level' [7, 8]. A variable at the 'interval' level of measurement has intervals between numbers which are equal. Thus, an interval level variable is one where there is an arithmetical relationship between the responses [9]. For example, someone who is 40 yr old is twice as old as someone 20 yr; someone who scores 12 on a continuous measure of walking disability (which may be, for example, walking distance) is twice as disabled as someone scoring six, given that a higher score is worse. On the other hand, an ordinal variable measuring walking disability is one which ranks (dis)ability. The critical difference between the 'interval' and 'ordinal' level is that all we can say about the latter is that a score of 12 is worse than one of six. Resolving this dispute for the HAQ is important, for otherwise we are left with uncertainty about which statistics may be appropriate. The utilization of parametric statistics requires that not only should the variable be normally distributed in the underlying population, but also that the measurement should be at the interval level (see Siegel and Castellan [10] for a useful discussion on this topic).

The opportunity to examine the scaling of the HAQ, to test its level of measurement (i.e. is it an ordinal or interval scale), and review whether or not we make the most of the scale, arose when it was applied in a community survey of those with joint problems.

METHODS

Design and setting

A two-stage random sample using postal questionnaires was commissioned by North Yorkshire Health in the UK to ascertain the health care needs of those with arthritis, particularly the likely demand for hip and knee arthroplasty. The population register of the North Yorkshire Family Health Services Authority, which is co-terminous with the North Yorkshire District Health Authority, was used as a sampling frame.
The initial questionnaire (Phase 1) was posted to 18 827 people, ~1 in 11 of the population ≥55 yr (~210 000) at the beginning of June 1993. Non-responders were sent further copies up to a maximum of two over the next 12 weeks, leading to a response rate of 87%. Those answering positively on items of interest were sent a more detailed questionnaire (Phase 2) which sought information on a variety of topics, such as health status, disability, dependency, and use of hospital and community services. It included various standard instruments, including the Medical Outcomes Study Short-Form 36 (SF36) [11] and the HAQ [2]. One thousand questionnaires were returned (an 82% response rate) at this stage. Full details of the methodology are given elsewhere [12]. This paper focuses on 62 people who self-reported rheumatoid arthritis (RA) and 442 who reported 'other arthritis', predominantly osteoarthritis (OA).

**Rasch analysis**

We assessed the HAQ using what has become known as the Rasch model [13]. The Rasch model formalizes desirable characteristics of measurement, e.g. unidimensionality. Unidimensionality is important because when we add together items from more than one dimension, lack of change in the score may result, for example, from an increase in score on items belonging to one dimension and a decrease on items belonging to another. More commonly, the items from one dimension may change a little, other items from another dimension much more, but the overall effect is diluted by those items with little change. In other words, we cannot tell from the overall score what is happening. This is why many modern measures of health status have scores for different dimensions, e.g. pain and function, but avoid a single overall score.

The Rasch analysis then seeks to combine the two (person ability and item difficulty) by their difference. This difference must govern the probability of what is supposed to happen when a person, of a given ability, uses that ability against a given task [17]. Results of the Rasch transformation are reported in 'logits' which is the distance along the line of the variable which increases the odds of observing the event (i.e. taking a 'step' on an item) by a factor of 2.718. The relationship between person ability and item difficulty can be best understood by the fact that a person with a logit score of 2.0 will have a 0.5 probability of 'passing' an item (or step on an item) with a difficulty level of 2.0 logits.

It is necessary to determine how well data fit the Rasch model. The model constructs a strong frame of reference against which the particular properties of the data are contrasted [18]. Various error estimates and fit statistics are provided for this purpose, including those testing unidimensionality [19] and a hierarchical structure [20]. This latter requirement is not the rigid requirement imposed by the Guttman model [21], but a probabilistic approach which requires that the greater the ability of an individual, the less likely they will have difficulty with any given item.

**RESULTS**

**Fit of data to the Rasch model**

The fit of the HAQ data to the model, for those self-reporting RA, is shown in Table I. The mean square information-weighted fit statistic (INFIT) is between −0.7 and +1.3, a range considered to represent an adequate fit of the data to the model. This fit statistic can also be standardized such that it takes the approximate form of a $t$-distribution. In this
manner, the data are also shown to fit the model (i.e. between -2.0 and +2.0) and thus indicate that the eight subscale items contribute to a single underlying construct.

The hierarchical nature of the scale, expressed by item separation, is somewhat restricted at 2.82. This meets the basic requirement that a scale should identify at least two strata, but suggests that in the HAQ the underlying scale construct of disability is limited in its range. In this community sample, examination of item separation indicates that the most ‘difficult’ item, reaching, fails to capture the upper level of disability experienced by some respondents. Likewise the ‘easiest’ item, eating (i.e. the item which most people could manage), expresses a level of disability above that experienced by some respondents. Thus the scale, in this population, has both floor and ceiling effects.

The fit of the data from 442 respondents with OA was found to be less than adequate, with strong indications that grip (mean square INFIT 1.56) does not contribute to the same underlying construct. In other words, this item had a highly variable response pattern, compared to other items, which would accommodate with clinical experience where grip is not usually affected with lower limb OA. In contrast, the item activities (mean square INFIT 0.67) is shown to have a more muted response pattern compared to other items, and this is usually referred to as dependency. While the conceptual model used as a basis for developing the HAQ comprises of both lower- and upper-limb subdimensions [6], no attempt has been made to score these dimensions separately.

Given these limitations, the hierarchical structure of the scale for those with OA is indicated by a separation value of 11.3. This suggests that the ‘spaces’ between items are also contributing to strata (i.e. the spread of eight items plus the spread of some of the spaces between), and this arises because of the larger sample size for those with OA giving rise to very small standard errors. Elsewhere, because of this problem, strata separation has been defined as within 0.15 logit values [22-23] and in this way four strata can be identified.

The HAQ scale: its level of measurement and summation properties

Over two-thirds (69%) of the 62 respondents with RA were female with mean age 66.6 yr (s.D. 7.9) and their mean HAQ score was 1.85 (s.d. 0.70). The 442 respondents with OA had a mean age of 68.6 yr (s.d. 9.5) and mean HAQ score of 1.39 (s.d. 0.67).

Comparison of the logit measure derived from the Rasch analysis and the raw score enables an examination of the level of measurement (i.e. ordinal or interval) of the raw score. Table II shows that for those with RA, incremental units of the raw score at the margins of the scale reflect an increasing level of (dis)ability compared to similar units in the centre of the scale. Equality of logit distance is maintained only within the range 0.75-2.25. Outside these scores, each HAQ point represents a widening (dis)ability level (at least 50% greater than the interval within the central band). For example, the distance between a score of 2.75 and 2.875 represents an increase in disability 3.4 times that of the apparently equivalent distance between 1.0 and 1.125.

DISCUSSION

The use of ordinal measures for outcome is widespread in rheumatology. While such measures are perfectly acceptable as ordinal scales, the use of parametric statistics in these circumstances is debatable. The data from the North Yorkshire survey confirm that the traditional score of the HAQ is at the ordinal level and should be transformed or subjected to appropriate non-parametric statistics. This would be essential when patients are entered into (or conclude) a clinical trial with scores which are outside the central linear band (0.75-2.25) of the scale. Adjustments for baseline differences between experimental and control groups which fail to take into account the increasing magnitude of disability outside of this central band may lead to incorrect conclusions.

It might be the case that the unusual way in which the HAQ is scored (i.e. one item determining the subscale score) contributes to the lack of equal intervals in its overall score. The eight subscales are derived from 20 items. Some subscales (e.g. walking) have two items, others (e.g. eating) have three items. We therefore looked at the calibration of all items (the item level) to test whether or not they had equivalence of difficulty.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>Fit of data to the Rasch model for those with RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item (subscale)</td>
<td>Logit measure</td>
</tr>
<tr>
<td>Eating</td>
<td>1.23</td>
</tr>
<tr>
<td>Walking</td>
<td>0.56</td>
</tr>
<tr>
<td>Rising</td>
<td>0.43</td>
</tr>
<tr>
<td>Hygiene</td>
<td>-0.02</td>
</tr>
<tr>
<td>Activities</td>
<td>-0.35</td>
</tr>
<tr>
<td>Dressing</td>
<td>-0.40</td>
</tr>
<tr>
<td>Grip</td>
<td>-0.60</td>
</tr>
<tr>
<td>Reach</td>
<td>-0.86</td>
</tr>
</tbody>
</table>

*Estimated.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>Calibration of HAQ scores against Rasch measure for those with RA</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAQ score</td>
<td>Rasch measure (logit)</td>
</tr>
<tr>
<td>0</td>
<td>-4.9*</td>
</tr>
<tr>
<td>0.125</td>
<td>-4.12</td>
</tr>
<tr>
<td>0.25</td>
<td>-3.34</td>
</tr>
<tr>
<td>0.375</td>
<td>-2.84</td>
</tr>
<tr>
<td>0.5</td>
<td>-2.45</td>
</tr>
<tr>
<td>0.625</td>
<td>-2.12</td>
</tr>
<tr>
<td>0.75</td>
<td>-1.83</td>
</tr>
<tr>
<td>0.825</td>
<td>-1.55</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.29</td>
</tr>
<tr>
<td>1.125</td>
<td>-1.03</td>
</tr>
<tr>
<td>1.25</td>
<td>-0.77</td>
</tr>
<tr>
<td>1.375</td>
<td>-0.50</td>
</tr>
<tr>
<td>1.5</td>
<td>-0.22</td>
</tr>
</tbody>
</table>
within each subscale. For example, rating impossible on an easy item should not carry the same weight as impossible on a difficult item. If, however, two such disparate items were within the same subscale, then vital information could be lost.

For those with RA, the data at the item level are also shown to have adequate fit to the model [mean square INFIT 0.7–1.3 for all items except shampoo hair (1.36) and dressing (0.62), with a standardized INFIT range of (−2.6 [dressing] to 1.8 [shampoo hair])]. With standardization, which places ~95% of the distribution within 2 S.D., we would expect one item out of 20 to be misfitting.

The hierarchical structure of the scale is confirmed by a separation value of 5.25. This level of separation shows greater discrimination along the underlying disability construct than found using the eight subscale items. Although this can be expected with an increased number of items, the result of using the 20 items is to remove the ceiling effect that was evident at the subscale level. The item lifting a full cup or glass to the mouth adequately represents the upper level of disability for this sample. In other words, in this particular sample, those who have difficulty with or find this task impossible have the severest disability.

How can the subscale level (eight item) fail to utilize the obvious strength of this item in defining the upper boundary of disability? Calibration shows that items belonging to each subscale vary considerably in their degree of difficulty, in all but the dressing subscale. In the eating subscale, opening a milk carton is some 3.07 logits harder than lifting a full cup to the mouth. Item calibration for those with OA followed a similar pattern to those with RA, with the same two items from the eating subscale having a logit difference of 2.61. Thus, the score on the milk carton item will almost always overshadow any difficulty experienced on the cup item. As such it is likely to be the milk carton item which determines the subscale score (cutting up food lies between the two). However, reporting ‘impossibility’ on either item determines a score of three points on the eating subscale, despite the obvious differences in item difficulty. This raises questions about the appropriateness of allowing the score for the subscale to be derived from items of different ‘weight’, and shows how the value of the single item lifting a full cup to the mouth can be lost using the current scoring system.

These findings suggest that for those with RA the 20 items comprise a useful scale and that the current method of scoring, using subscale scores derived from the within-scale most difficult item, is a less than optimum approach. Vital information is being lost and true changes in the level of disability are masked. We would suggest that a supplementary score be reported based on 60 points (e.g. 48/60). Furthermore, the 60-point scale exhibits a broader central band at the interval level, equivalent to a 0.5–2.5 range on the standard HAQ. Thus, there would be less chance that standard methods of adjusting for baseline differences will be compromised.

Given the implications for interpretation of clinical trial data where the HAQ is used as a primary outcome measure, replication of these findings is urgently required. Where interval level measures are needed, transformation of the standard HAQ (eight subscale) should be considered, certainly where patients fall outside the central band of the scale. Rasch models offer one way of transforming ordinal scores into linear measures. However, there are different approaches to operationalizing the Rasch models, based upon using the unconditional maximum likelihood function (MLF), more commonly used in North America and Australasia [24, 25], and the conditional MLF, more common in Europe [26, 27]. The difference between these two approaches is that, in the former, item difficulty and person ability are estimated by an iterative procedure at the same time (which is the approach we have used), whereas in the latter, item difficulty is estimated conditional on person ability—the raw score. It is important to be aware of these two traditions for the fit statistics are quite different, and there is currently some debate as to the efficacy of these approaches and the accuracy of their fit statistics [28, 29]. However, this debate is primarily related to scales which have ≤10 items. As such, it would be preferable to assess the HAQ using both approaches.

Finally, further investigation is needed to ascertain the construct validity of the scale for different types of arthritis, and indeed for other diseases as the instrument was designed for use in all illnesses, not just rheumatic diseases [6]. The current study suggests that unidimensionality may be compromised when used in populations with predominately lower-limb involvement associated with OA.

**ACKNOWLEDGEMENTS**

With thanks to Dr William P. Fisher Jr, LSU Medical Centre, New Orleans, USA, for his helpful comments on earlier drafts of this paper.

**REFERENCES**