Qual Life Res (2010) 19:1255–1263
DOI 10.1007/s11136-010-9690-9

Application of the health assessment questionnaire disability index to various rheumatic diseases

Maaike M. van Groen · Peter M. ten Klooster · Erik Taal · Mart A. F. J. van de Laar · Cees A. W. Glas

Accepted: 6 June 2010 / Published online: 18 June 2010

Abstract
Purpose To investigate whether the Stanford Health Assessment Questionnaire Disability Index (HAQ-DI) can serve as a generic instrument for measuring disability across different rheumatic diseases and to propose a scoring method based on item response theory (IRT) modeling to support this goal.
Methods The HAQ-DI was administered to a cross-sectional sample of patients with confirmed rheumatoid arthritis ($n = 619$), osteoarthritis ($n = 125$), or gout ($n = 102$). The results were analyzed using the generalized partial credit model as an IRT model.
Results It was found that 4 out of 8 item categories of the HAQ-DI displayed substantial differential item functioning (DIF) over the three diseases. Further, it was shown that this DIF could be modeled using an IRT model with disease-specific item parameters, which produces measures that are comparable for the three diseases.
Conclusion Although the HAQ-DI partially functioned differently in the three disease groups, the measurement regarding the disability level of the patients can be made comparable using IRT methods.

M. M. van Groen · P. M. ten Klooster (✉) · E. Taal · M. A. F. J. van de Laar · C. A. W. Glas
Institute for Behavioral Research, Faculty of Behavioral Sciences, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
e-mail: P.M.tenKlooster@utwente.nl

M. A. F. J. van de Laar
Department of Rheumatology, Medisch Spectrum Twente, PO Box 50.000, 7500 KA Enschede, The Netherlands

Keywords Rheumatoid arthritis · Osteoarthritis · Gout · Health-related quality of life · Item response theory · Differential item functioning

Abbreviations

DIF	Differential item functioning
HAQ-DI	Health Assessment Questionnaire Disability Index
IRT	Item response theory
LM	Lagrange multiplier
OA	Osteoarthritis
PF	Physical functioning scale
PsA	Psoriatic arthritis
RA	Rheumatoid arthritis
SF-36	Medical Outcomes Study 36-Item Short Form

Introduction

Besides the traditional use of physical and biochemical measures, patient-centered outcomes have become more and more important as outcome measures of interventions [1]. For example, patient-reported disability has become a standard outcome in the clinical studies of rheumatic diseases. One of the most widely used self-reported measures of physical disability is the Stanford Health Assessment Questionnaire Disability Index (HAQ-DI) [2]. Although often referred to as a disease-specific measure, it assesses physical disability in general and does not focus on specific disease-associated impairments. In fact, according to its developers, it was originally intended for use in multiple illnesses so that the impact of different disease processes could be compared [1, 3]. As a result, the scale has been used across a wide range of general and clinical populations.

Especially in the field of rheumatology, the HAQ-DI has become the measure of choice for assessing physical disability in several specific rheumatic diseases. Although physical disability is common among all musculoskeletal conditions, rheumatic diseases can vary widely in their underlying disease mechanisms, clinical manifestations, progress and severity, and composition of the populations generally affected. All of which may influence the measurement characteristics and resulting disability scores across diseases. Nonetheless, mean HAQ-DI scores are frequently used to directly compare the severity of disability across different rheumatic diseases, whether or not adjusted for some known covariates [4–9]. The purpose of the current study is to investigate whether the HAQ-DI is a generic instrument indeed, and if this proves problematic, to model response behavior on disease-specific items of the instrument in such a way that the measurement results are comparable over different groups of rheumatic patients.

The construct validity of the HAQ-DI has been previously established in numerous studies [1], mostly using classical psychometric techniques such as factor analysis. Cole et al. (2005, 2006), for instance, show that there is considerable support for a single-factor structure and for comparability of scores of patients with systemic sclerosis and patients with rheumatoid arthritis. However, some of the results of these analyses, such as the presence of correlated residuals, invite further attention. In the present article, construct validity is investigated using a unidimensional item response theory (IRT) model. The relation between IRT modeling and factor-analytic approaches will be returned to the discussion section.

In IRT models, observed responses are related to a unidimensional latent trait, that is, to some underlying scale. The unidimensional latent scale of the HAQ-DI pertains to the disability level of the patients. The observed responses are explained by the persons' disability parameters and by item parameters related to the probability that a person with a certain disability parameter endorses an item. One of the common assumptions of IRT is measurement invariance, that is, the latent scale applies to all respondents from some population and items have the same measurement characteristics, that is, the same item parameters, for these respondents. A violation of these two assumptions is known as differential item functioning (DIF). An item shows DIF if the probability of responding in the different categories of the item varies across groups of patients with the same disability level [10, 11]. In other words, an item is biased if the observed item score, conditional on the latent disability level of the patients, differs between subgroups [12]. In the current study, the construct validity of the HAQ-DI is investigated by assessing DIF for patients with three different types of arthritis.

DIF is often investigated using the generalized partial credit model as an IRT model [11]. The generalized partial credit model [13] applies to polytomously scored items, such as the items of the HAQ-DI. The probability of a score in category x of item i is given by the item response curve

$$P(X_{ni} = x|\theta) = \frac{\exp\left[\sum_{j=1}^{x} \alpha_i(\theta_n - \delta_{ij})\right]}{1 + \sum_{r=1}^{m_i}\left[\exp\sum_{j=1}^{r} \alpha_i(\theta_n - \delta_{ij})\right]},$$

where θ_n is the latent disability level of patient n. In the model, m_i denotes the number of item categories. Further, δ_{ij} and α_i are item parameters. δ_{ij} is a category intersection parameter, that is, it is the point in which the probability of responding in category $j - 1$ is equal to the probability of responding category j. Finally, α_i is a discrimination parameter that indicates the extent to which the item response is related to the latent scale. This discrimination parameter is comparable to a factor loading in a factor analysis model.

If DIF is not present, this is unambiguous support for the construct validity of the instrument. If DIF is present, however, the type of DIF becomes important. As previously noted, measurement invariance pertains to the presence of the same latent variable in all subgroups and constancy of item parameters over subgroups. If only the latter assumption is violated by a limited number of items, comparability can often still be realized and construct validity may still be defendable. For example, a question regarding the number of cars in the household may be a good item for measuring the latent variable Wealth, though the metric in downtown New York and in Texas may be quite different. In IRT, such differences can be modeled by group-specific item parameters. This approach is, of course, only defendable if it can be explicitly shown that the responses to the items given in the two groups pertain to the same latent variable, that is, that it can be shown that the same IRT model holds for the entire set of response data. This approach to modeling DIF, which has a considerable tradition in educational measurement [14–17] and in consumer research [18], will also be applied in the present study.

Patients and methods

Respondents for this study were recruited during several waves of data collection in the period between 2005 and 2008 at the outpatient rheumatology clinic of the Medisch Spectrum Twente hospital in Enschede, the Netherlands. During data collection days, consecutive patients visiting

the outpatient clinic were asked to participate. As the study did not interfere with usual treatment, ethical approval was not required according to national legislation and local institutional policy.

In total, 846 patients with physician-confirmed rheumatoid arthritis (RA), osteoarthritis (OA), or gout agreed to participate. Of the included patients, 619 patients were treated for RA, 125 for OA, and 102 for gout. Table 1 gives a number of characteristics of the sample. The majority of the patients were women, but as would be expected, the gout sample consisted of only 18% women. Mean age was 62 with a standard deviation of 13.6 years. A validated Dutch version of the HAQ was used [19]. The average scores on the Medical Outcomes Study 36-Item Short Form (SF-36) health survey [20] were reasonably comparable across the three conditions. HAQ-DI scores were similar for patients with RA and OA, whereas patients with gout reported substantially less disability.

Scoring the HAQ-DI

The Health Assessment Questionnaire Disability Index (HAQ-DI) consists of 20 questions regarding the limitations patients experience in performing daily physical activities [2]. Patients are asked how difficult it is to perform an activity on a scale of 0 (without any difficulty) to 3 (unable to do). Patients are also asked whether they need assistance or aids for the activity.

The questions of the HAQ-DI are ordered into eight categories of daily living, covering Rising, Walking, Dressing and grooming, Reach, Eating, Grip, Activities, and Hygiene. The highest item score within a category is used as the score for this category, essentially reducing the HAQ-DI to an 8-item scale. If a respondent indicates the use of assistance or aids for a category and his highest item score within the category is 0 or 1, the category score is raised to the value 2. The scores on the categories are averaged to construct a single total score.

Statistical analysis

The scores on the eight categories of the HAQ-DI were used as input for the statistical analysis. The item parameters and the means and variances of the latent person parameters were estimated by marginal maximum likelihood, and DIF was examined using Lagrange multiplier (LM) statistics [21]. To compute these statistics, the sample of respondents is divided into subgroups labeled $g = 1,\ldots,G$. For the present application, these are the three disease groups, that is, $G = 3$. The statistic is based on the difference between average observed scores on every item i in the subgroups, that is, $S_{ig} = \frac{1}{N_g}\sum_{n|g}^{N_g} X_{ni}$ (where the summation is over the N_g respondents in subgroup g), and their expectations $E(S_{ig})$. The differences are squared and divided by their covariance matrix (for the exact expressions, refer to Glas [15]). The hypothesis thus tested is equivalent to testing the hypothesis that the parameters of the items are equal for the subgroups. The LM statistic has an asymptotic chi-square distribution with $G - 1$ degrees of freedom. Below, the statistics will be accompanied by effect sizes $d_{ig} = \max_g \left|S_{ig} - E(S_{ig})\right|$, which show the seriousness of the model violation. Since the effect sizes d_{ig} are on a scale ranging from 0 to the maximum score m_i, effect sizes $d_{ig} < 0.10$ can be considered indicative of minor, acceptable model violation. It can be noted that this cut-off point is somewhat arbitrary, but its effectiveness can be evaluated from whether enough DIF items are detected and modeled to obtain a fitting overall model.

When items with DIF are identified, the next step is trying to model the DIF in such a way that the measures

Table 1 Sample characteristics

Characteristics	Total sample	RA sample	OA sample	Gout sample
N	846	619	125	102
Gender (%)				
Female	64	69	79	18
Male	36	31	21	82
Age (years)				
Mean (SD)	62 (13.6)	62 (14.2)	63 (11.5)	62 (12.6)
Disease duration (years)				
Mean (SD)	13 (12.8)	13 (12.4)	14 (13.8)	10 (13.3)
HAQ-DI (range 0–3)				
Mean (SD)	0.82 (0.7)	0.97 (0.7)	1.00 (0.65)	0.54 (0.67)
SF-36 PCS (range 0–100)				
Mean (SD)	40 (8.4)	39 (8.1)	38 (8.3)	43 (9.2)
SF-36 MCS (range 0–100)				
Mean (SD)	39 (7.0)	40 (6.8)	39 (7.3)	38 (7.0)

RA Rheumatoid arthritis, *OA* Osteoarthritis, *HAQ-DI* Health Assessment Questionnaire Disability Index, *SF-36* Medical Outcomes Study 36-Item Short Form (version 2), *PCS* Physical component summary, *MCS* Mental component summary

obtained in the subgroups are still comparable. To this end, DIF can be modeled by assigning these items disease-specific parameters within a generalized partial credit model that still pertains to all respondents. So it is assumed that the same construct is measured in all subgroups, but for some subgroups the item locations on the latent scale are different. In this study, this was done in an iterative procedure in which the item with the largest significant LM test was given disease-specific item parameters (for more information on this procedure, see Glas and Verhelst [16]). These iteration steps were repeated until no items were left with significant LM tests ($P < 0.01$) or when the effect size was below the set cut-off point ($d_{ig} < 0.10$). The results of these iteration steps are presented here as results of an analysis consisting of two steps to enhance clarity.

The final step in the statistical analyses was to assert that the resulting model was valid in all disease groups, that is, to assert that the same latent scale with disease-specific item parameters for some of the items was applicable in all disease groups. This was done again by computing LM statistics: one targeted at the form of the item response curves and one targeted at the assumption of local independence. The latter assumption implies that item responses are independent given a person's value of on the latent variable. If this would not be the case, other, unaccounted, variables influence response behavior and unidimensionality is violated.

Results

The results of the DIF analysis before modeling for presence of DIF are given in Table 2. Three items showed DIF according to the criteria defined earlier: Dressing and grooming, Reach, and Activities. The LM statistics of those items were significant, and their effect sizes were larger than 0.10. The item Dressing and grooming was given disease-specific item parameters first. In a second analysis, the Activities item showed DIF. The process was repeated until in the third analysis four items were given disease-specific item parameters. The resulting item parameters are shown in Table 3. It is important to note that the significant items in Table 2 (Walking, Dressing and grooming, Reach, Activities) are not completely analogous to the items in Table 3 (Walking, Dressing and grooming, Eating, Activities). The reason is that the presence of DIF items biases the item parameter estimates of all items, both the items with and without DIF. This motivates the iterative nature of the procedure where items are processed one at a time.

To clarify the interpretation of the results in the table, the item probabilities for OA and gout are illustrated in Fig. 1. Consider the item Dressing and grooming. The discrimination indices (under the heading α_i in Table 3) show that the item has the highest loading on the latent dimension for the patients with gout and the lowest for the patients with OA. Further, in Table 3 and Fig. 1, it can be seen that the category intersection parameters δ_{ij} are higher for the patients with gout than for the patients with OA. This means that the expected score on Dressing and

Table 2 Outcomes of tests for DIF

HAQ-DI categories	LM	*P*	Abs. DIF
Rising	3.33	0.19	0.03
Walking	25.02	0.00	0.09
Dressing and grooming	71.20	0.00	0.17
Reach	37.63	0.00	0.15
Eating	6.68	0.04	0.06
Grip	2.66	0.26	0.03
Activities	51.52	0.00	0.15
Hygiene	7.29	0.03	0.06

DIF Differential item functioning, *HAQ-DI* Health Assessment Questionnaire Disability Index, *LM* Outcome Lagrange multiplier test, *Abs. DIF* Amount of absolute DIF d_{ig}

Degrees of freedom = 2

Table 3 Item parameters after modeling DIF

HAQ-DI categories	Item parameters			
	αi	δ_{i1}	δ_{i2}	δ_{i3}
Rising	3.758	−0.089	3.777	4.660
Walking				
RA	3.253	0.429	3.568	6.534
OA	3.691	−0.515	3.878	6.875
Gout	3.987	0.196	3.844	9.377
Dressing and grooming				
RA	3.285	−0.428	2.124	3.905
OA	2.532	0.101	2.885	3.666
Gout	3.850	3.066	4.832	4.994
Reach	2.671	−0.004	2.306	3.499
Eating				
RA	2.629	−0.883	1.982	1.861
OA	3.077	−0.299	2.828	2.585
Gout	2.464	−0.079	2.786	1.673
Grip	3.824	−0.149	3.176	4.455
Activities				
RA	2.915	−1.234	2.159	3.205
OA	2.431	−0.388	2.070	4.226
Gout	3.124	1.713	3.829	4.172
Hygiene	3.768	−1.557	2.089	3.242

DIF Differential item functioning, *HAQ-DI* Health Assessment Questionnaire Disability Index, α_i Discrimination parameter, δ_{i1}, δ_{i2}, and δ_{i3} Category intersection parameters, *RA* Rheumatoid arthritis, *OA* Osteoarthritis

Log likelihood = −5941.921

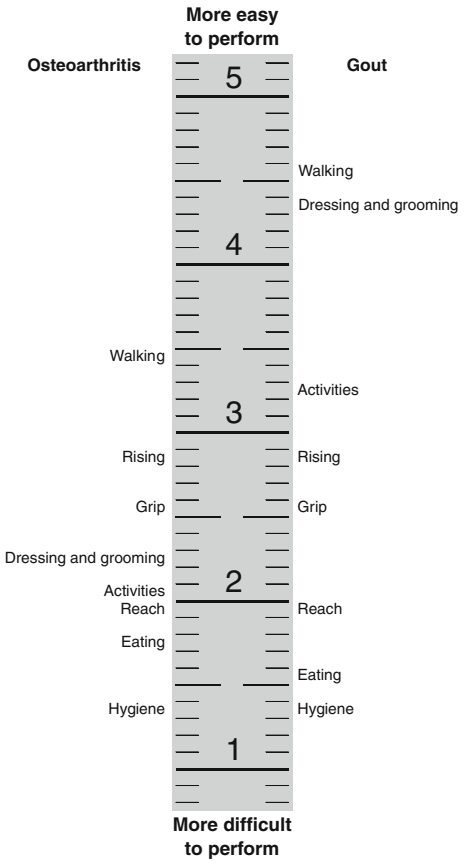

Fig. 1 An illustration of the item difficulty locations (average of the 3 category intersection parameters) of the Health Assessment Questionnaire Disability Index on the IRT latent scale in patients with osteoarthritis and gout

grooming given a certain disability level is higher for the patients with OA than for the patients with gout. That is, patients with OA endorse this item more and the item Dressing and grooming is more difficult for them than for the patients with gout.

The next question addressed is whether the scale presented in Table 3 actually fits the data. This was investigated using two LM statistics [21], one targeted at the form of the item response curves and one targeted at the assumption of local independence. The first statistic is defined analogous to the statistic for DIF, only this time the subgroups are total-score level groups within the disease groups. The observed total score is the sum score of the responses on all items except the item targeted. Glas [21] demonstrated that this statistic pertains to the hypothesis that the response probabilities as a function of the latent disability parameters are as predicted by the model. Within the three disease groups, three total-score level groups were formed in such a way that the numbers of respondents in each group were approximately the same. The ranges of the scores in the total-score level groups are given at the bottom of the table. The results for the patients with RA are

Table 4 Outcomes of tests for model fit in score level groups for patients with RA

HAQ-DI categories	Total-score level groups								
			Level 1		Level 2		Level 3		*d*
	LM	*P*	Obs.	Exp.	Obs.	Obs.	Obs.	Exp.	
Rising	2.77	0.25	0.21	0.22	0.81	0.77	1.56	1.51	0.03
Walking	7.26	0.03	0.13	0.16	0.68	0.62	1.20	1.22	0.04
Dressing and grooming	0.22	0.89	0.32	0.32	1.03	1.02	1.84	1.84	0.01
Reach	7.05	0.03	0.25	0.27	0.76	0.83	1.55	1.58	0.04
Eating	1.53	0.47	0.47	0.47	1.22	1.19	2.12	2.12	0.01
Grip	4.63	0.10	0.20	0.23	0.88	0.83	1.68	1.67	0.03
Activities	4.70	0.10	0.53	0.51	1.05	1.12	1.90	1.86	0.05
Hygiene	0.70	0.71	0.50	0.50	1.25	1.27	2.21	2.19	0.01

RA Rheumatoid arthritis, *HAQ-DI* Health Assessment Questionnaire Disability Index, *LM* outcome Lagrange multiplier test, *Obs.* Observed scores, *Exp*. Expected scores by the model, The observed total score is the sum score of the responses on all items, *d* effect size

Level 1: total scores 0–4, Level 2: total scores 5–8, Level 3: total scores 9–23

shown in Table 4. The results for the two other diseases were analogous. Note that none of the outcomes of the LM tests were below the significance level of 1%. The last column gives the effect sizes d_{ig}. The highest effect size was 0.05, which was well below the set criterion of 0.10. The overall conclusion is that the model fitted very well, and the hypothesis that the same latent scale pertained to the three diseases was not rejected.

The second test pertained to local independence. The test is also sensitive to violations of unidimensionality. The test targets the dependence between responses on pairs of items. In the present case, responses to consecutive items were evaluated, but this choice is not essential. The test statistic is based on the evaluation of the average scores on some item given the scores on some other item. With this alternative definition of score groups, the test statistic is defined analogous to other LM statistics. The results for the patients with RA are displayed in Table 5. As with the tests for DIF and the form of the item response curves, the results for the two other diseases were analogous and none of the outcomes of the LM tests were below the significance level of 1%. The columns labeled '0' to '3' give the observed and expected average scores on some item *i* given that the score on item *i-1* was '0' to '3', respectively. That is, the average score on the item $i = 2$, i.e., Walking, for patients scoring '0' on item $i - 1$, i.e., Rising, was 0.18. The associated expected score was 0.23. The last column gives the effect sizes d_{ig}. The highest effect size was 0.10, which just attained the criterion of 0.10. So again, the predictions by the model were quite acceptable.

Table 5 Outcomes of tests for local independence for patients with RA: score level given the score level on the previous item

Category	LM	*P*	Score on previous category								
			0		1		2		3		*d*
			Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	
2	8.03	0.05	0.18	0.23	0.80	0.76	1.28	1.26	1.61	1.70	0.09
3	3.65	0.30	0.53	0.51	1.28	1.30	1.91	1.97	2.83	2.79	0.06
4	2.25	0.52	0.37	0.34	0.81	0.83	1.35	1.40	1.99	2.04	0.05
5	2.50	0.48	0.61	0.64	1.38	1.34	1.95	1.95	2.57	2.49	0.06
6	7.69	0.05	0.22	0.26	0.86	0.80	1.27	1.25	1.78	1.84	0.06
7	0.34	0.95	0.57	0.57	1.18	1.19	1.82	1.82	2.40	2.30	0.10
8	1.01	0.80	0.50	0.47	1.17	1.18	1.84	1.85	2.45	2.46	0.03

RA Rheumatoid arthritis, The categories are numbered in the order as they appear in Table 4, *LM* outcome Lagrange multiplier test, *Obs.* Observed scores, *Exp.* Expected scores by the model, The observed total score is the sum score of the responses on all items, *d* effect size

The last question addressed concerns the impact of the DIF for inferences concerning differences between the three diseases on the latent scale. As mentioned previously, the item parameters and the means and variances of the latent person parameters were estimated by marginal maximum likelihood. The obtained mean values of disability for each disease are presented in Fig. 2, together with 99% confidence intervals. The mean for the patients with gout was set equal to zero to identify the latent scale. Note that average disability of the respondents for each disease decreases after the introduction of disease-specific item parameters. Patients with OA had the highest average disability level in all analyses. Patients with gout had the lowest disability. From the confidence intervals, it can be inferred that conclusions from statistical tests would not change. However, after modeling DIF, absolute score differences clearly decreased.

Discussion

An item response theory (IRT)-based method is presented that can be used to make HAQ-DI disability scores better comparable across different rheumatic diseases, and the results of the application of this method suggest that the HAQ-DI can function as a generic instrument.

By now, there is extensive literature on the evaluation of construct validity using factor analyses and IRT analyses

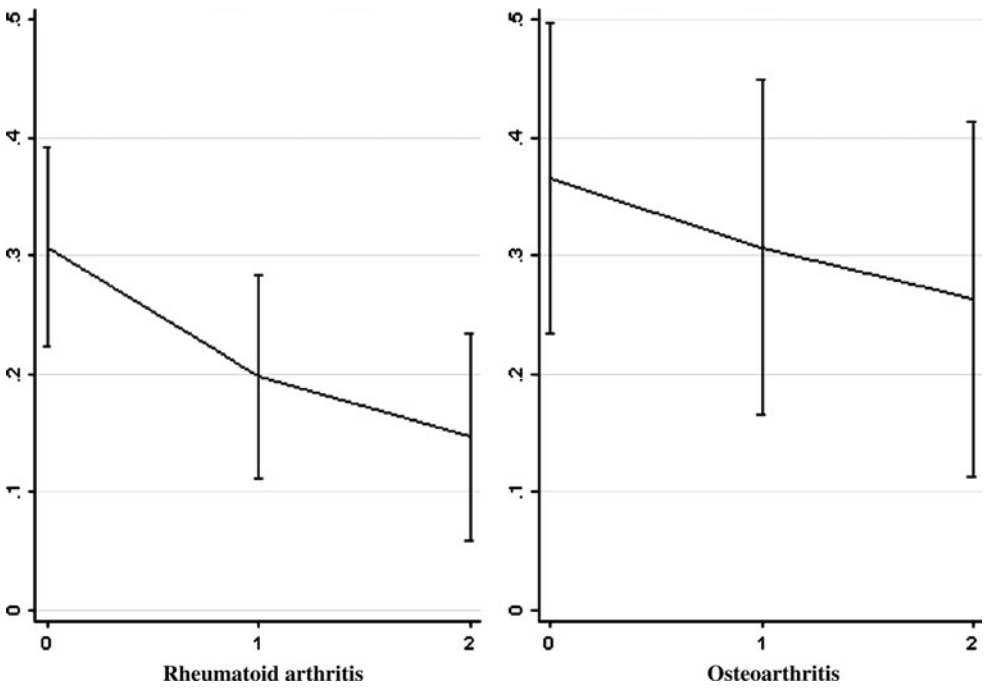

Fig. 2 Means of IRT disability estimates (*y*-axis) in rheumatoid arthritis (*left panel*) and osteoarthritis (*right panel*) in three analyses (*x*-axis) labeled 0, 1, and 2. The mean for gout was set equal to zero to identify the latent scale. Analysis 0 was the initial analysis. In analyses 1 and 2, 2 and 4 items with disease-specific item parameters were introduced, respectively

[22]. It is important to note that these two classes of models are closely related. In fact, Takane and de Leeuw have shown that under quite general assumptions, these two models are equivalent [23]. Only the traditions of statistical inference are different: factor analysis is usually based on a covariance matrix, while IRT analysis is based on the complete response patterns. This motivates the term "full-information factor analysis" used for multidimensional IRT by Bock, Gibbons, and Muraki [24]. Both approaches have their advantages and disadvantages. One of the advantages of the IRT approach is that it uses more information in the data and, therefore, assumptions such as the form of the item response curves and local independence can be investigated. However, the results obtained using both approaches are closely related. In that sense, the correlated residuals reported by Cole [25, 26] can be interpreted as an indication for lack of local independence and multidimensionality, which can be further investigated in detail using IRT-based techniques. Although both factor analysis and IRT can be used to assess the construct validity of the HAQ-DI, it is important to note that construct validity is not so much a property of an instrument, but a property of inferences made using the instrument [27]. In the present study, it was shown that when a number of disease-specific item parameters are used and the HAQ-DI is scored using θ-estimates, these θ-estimates relate to the same unidimensional scale. Therefore, these scores can support the construct validity of the HAQ-DI for inferences across diseases.

IRT methods offer a sophisticated and robust means to test the generic nature of an instrument by examining whether the underlying latent scale is the same for different groups of individuals. This can be evaluated by examining whether the questionnaire contains items with differential item functioning (DIF), i.e., items where the probability of scoring in the various response categories differs between subgroups of patients after controlling for the general disability level as estimated by the IRT model. Although IRT-based approaches to DIF detection have been increasingly used in health outcomes assessment, research addressing the measurement equivalence of disability scales across different (rheumatic) diseases is still scarce. Only one recent study was found that examined DIF for the HAQ-DI and the 10-item physical functioning scale (PF) of the SF-36 between patients with RA and psoriatic arthritis (PsA) using Rasch analysis [28]. This study found evidence of marked DIF for three HAQ-DI items, similar to our study, and relatively minor DIF for the SF-36 PF scale. The authors concluded that the SF-36 PF scale is a better instrument than the HAQ-DI for comparing disability from PsA with disability from other diseases. However, the study did not evaluate the impact of DIF on individual items for inferences about total HAQ-DI score differences between the diseases or provide guidelines on how to deal with this DIF. Therefore, the objective of this study was twofold: first to investigate whether the HAQ-DI functions as a generic measure of disability across different rheumatic diseases by evaluating DIF and second, if not, to illustrate the use of IRT methods to model DIF so that disability scores can be made comparable across diseases. For this purpose, we used data from three common rheumatic diseases with known differences in disease characteristics: rheumatoid arthritis (RA), osteoarthritis (OA), and gout.

As would be expected, the majority of the patients with RA and OA were women, whereas patients with gout patients were predominantly men. Mean SF-36 physical and mental component scores were well below the average of 50 in the general population, suggesting that all three diseases have a substantial impact on general health status. Whereas disability scores between RA and OA were very similar, mean HAQ-DI scores were clearly lower for patients with gout and in close correspondence to a recently reported mean HAQ-DI score of 0.59 in a cross-sectional gout sample [29].

However, half of the HAQ-DI items displayed substantial DIF between the three diseases, possibly biasing total score differences between the diseases. After modeling these items by assigning them disease-specific parameters, statistical conclusions regarding disability differences across the 3 conditions did not change. Patients with OA and RA still displayed higher disability scores than patients with gout. However, absolute differences between the diseases were attenuated. HAQ-DI scores based on disease-specific item parameters fitted the data very well and resulted in an underlying latent scale that applied to all three diseases.

An important concern, however, is that only four items served as anchors across the three diseases, and these items appear to be on the "more difficult" end of the scale. To minimize the standard errors of differences between disability estimates in the different disease groups, anchoring should be preferably done in all sections of a scale. Often, this cannot be achieved, but it should be kept in mind that the precision of the method deteriorates with the number of anchor items and their position on the scale. The authors do not recommend using the method when the anchor is very small (e.g., less than 4 items or less than 50% of the items).

It is also important to emphasize here that the present study focused only on cross-sectional samples of patients with OA, RA, and gout as an example for evaluating the generic nature of the HAQ-DI. It is very well possible that these or other items of the HAQ-DI may show DIF, possibly to a different extent, between other rheumatic conditions, non-rheumatic conditions, or general population samples. Accordingly, researchers using the HAQ-DI to compare disability between different subgroups are

encouraged to examine DIF before comparing total HAQ-DI scores. The present study provides an example of how IRT methods can be used to evaluate DIF and, if necessary, how to model this DIF to obtain more accurate disability estimates.

Furthermore, all analyses presented in this study are based on so-called standard scores of the HAQ-DI, which take into account the use of aids and devices or assistance from another person [1, 3]. Although this scoring method is most frequently used and recommended [30], some clinical investigations have used an alternative scoring without this correction. Secondary analysis using the alternative scoring method in this study showed that the IRT results obtained with and without correction were very similar.

In summary, the results of this study showed that 4 out of the 8 disability items displayed substantial DIF across the 3 diseases, indicating that the HAQ-DI may not fully function as a generic instrument for the assessment of disability across different rheumatic diseases unless DIF is modeled and adjustments to the scoring method are made.

References

1. Bruce, B., & Fries, J. F. (2003). The Stanford health assessment questionnaire: A review of its history, issues, progress, and documentation. *Journal of Rheumatology, 30*(1), 167–178.
2. Fries, J. F., Spitz, P., Kraines, R. G., & Holman, H. R. (1980). Measurement of patient outcome in arthritis. *Arthritis and Rheumatism, 23*(2), 137–145.
3. Bruce, B., & Fries, J. F. (2005). The health assessment questionnaire (HAQ). *Clinical and Experimental Rheumatology, 23*(5 Suppl 39), S14–S18.
4. Husted, J. A., Gladman, D. D., Farewell, V. T., & Cook, R. J. (2001). Health-related quality of life of patients with psoriatic arthritis: a comparison with patients with rheumatoid arthritis. *Arthritis and Rheumatism, 45*(2), 151–158.
5. Johnson, S. R., Glaman, D. D., Schentag, C. T., & Lee, P. (2006). Quality of life and functional status in systemic sclerosis compared to other rheumatic diseases. *Journal of Rheumatology, 33*(6), 1117–1122.
6. Lindqvist, U. R., Alenius, G. M., Husmark, T., Theander, E., Holmstrom, G., & Larsson, P. T. (2008). The Swedish early psoriatic arthritis register–2-year followup: a comparison with early rheumatoid arthritis. *Journal of Rheumatology, 35*(4), 668–673.
7. Martinez, J. E., Ferraz, M. B., Sato, E. I., & Atra, E. (1995). Fibromyalgia versus rheumatoid arthritis: A longitudinal comparison of the quality of life. *Journal of Rheumatology, 22*(2), 270–274.
8. Slatkowsky-Christensen, B., Mowinckel, P., Loge, J. H., & Kvien, T. K. (2007). Health-related quality of life in women with symptomatic hand osteoarthritis: a comparison with rheumatoid arthritis patients, healthy controls, and normative data. *Arthritis & Rheumatism (Arthritis Care & Research), 57*(8), 1404–1409.
9. Sokoll, K. B., & Helliwell, P. S. (2001). Comparison of disability and quality of life in rheumatoid and psoriatic arthritis. *Journal of Rheumatology, 28*(8), 1842–1846.
10. Camilli, G., & Sheppard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
11. Holland, P. W., & Wainer, H. (1994). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
12. Chang, H. H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 59*(3), 391–404.
13. Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.
14. Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), 305–322.
15. Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8*(3), 647–667.
16. Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–96). New York: Springer.
17. Grisay, A., de Jong, J. H., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249–266.
18. de Jong, M. G., Steenkamp, J. B. E. M., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of consumer research, 34*(2), 260–278.
19. ten Klooster, P. M., Taal, E., & van de Laar, M. A. (2008). Rasch analysis of the Dutch health assessment questionnaire disability index and the health assessment questionnaire II in patients with rheumatoid arthritis. *Arthritis & Rheumatism (Arthritis Care & Research), 59*(12), 1721–1728.
20. Ware, J. E., Kosinski, M., & Dewey, J. E. (2000). *How to score version 2 of the SF-36 health survey*. Lincoln, RI: Quality Metric Incorporated.
21. Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273–294.
22. Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.
23. Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discredited variables. *Psychometrika, 52*(3), 393–408.
24. Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-Information item factor analysis. *Applied Psychological Measurement, 12*(3), 261–280.
25. Cole, J. C., Khanna, D., Clements, P. J., Seibold, J. R., Tashkin, D. P., Paulus, H. E., et al. (2006). Single-factor scoring validation for the health assessment questionnaire-disability index (HAQ-DI) in patients with systemic sclerosis and comparison with early rheumatoid arthritis patients. *Quality of Life Research, 15*(8), 1383–1394.
26. Cole, J. C., Motivala, S. J., Khanna, D., Lee, J. Y., Paulus, H. E., & Irwin, M. R. (2005). Validation of single-factor structure and scoring protocol for the health assessment questionnaire-disability index. *Arthritis & Rheumatism (Arthritis Care & Research), 53*(4), 536–542.
27. Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing.

28. Taylor, W. J., & McPherson, K. M. (2007). Using Rasch analysis to compare the psychometric properties of the short form 36 physical function score and the health assessment questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. *Arthritis & Rheumatism (Arthritis Care & Research), 57*(5), 723–729.
29. Alvarez-Hernandez, E., Pelaez-Ballestas, I., Vazquez-Mellado, J., Teran-Estrada, L., Bernard-Medina, A. G., Espinoza, J., et al. (2008). Validation of the health assessment questionnaire disability index in patients with gout. *Arthritis & Rheumatism (Arthritis Care & Research), 59*(5), 665–669.
30. Zandbelt, M. M., Welsing, P. M., van Gestel, A. M., & van Riel, P. L. (2001). Health assessment questionnaire modifications: is standardisation needed? *Annals of the Rheumatic Diseases, 60*(9), 841–845.