Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 53, No. 4, August 15, 2005, pp 536–542
DOI 10.1002/art.21325

ORIGINAL ARTICLE

Validation of Single-Factor Structure and Scoring Protocol for the Health Assessment Questionnaire-Disability Index

JASON C. COLE, SAROSH J. MOTIVALA, DINESH KHANNA, JESSICA Y. LEE, HAROLD E. PAULUS, AND MICHAEL R. IRWIN

Objective. **The extensively used Health Assessment Questionnaire Disability Index (HAQ-DI) has been well received by the research and clinical community, notably because of its measurement strengths including reliability and stability of scores over time, utility in observational studies and clinical trials, predictive relationship with morbidity and mortality in rheumatoid arthritis (RA), and its translation for use in different countries. However, HAQ-DI scoring has not been validated. The purpose of this study was to examine the structural validity of the HAQ-DI and evaluate the latent factors underlying HAQ-DI scoring.**

Methods. **This study used a cross-validation approach on a total of 278 patients with RA. Exploratory and confirmatory factor analyses were performed.**

Results. **Results yielded a single-factor HAQ-DI score, which favored the current scoring system of the HAQ-DI. Additionally, modification indices suggested improved model fit with the secondary inclusion of correlated residual scores from a motor skills subdomain.**

Conclusion. **The current study provides the first validation of the HAQ-DI scoring system as determined by its latent factor structure. In addition, the findings suggest some benefit from a secondary interpretation of the scores based on domains that measure motor skills.**

KEY WORDS. Rheumatoid arthritis; Latent analysis; Confirmatory factor analysis; HAQ-DI.

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic, systemic, inflammatory disorder of unknown etiology that primarily involves the joints. It may be remitting, but if uncontrolled, may lead to deformity and destruction of joints due to the erosion of cartilage and bone. This symmetrical disease often progresses from peripheral to more proximal joints and, in many patients, results in significant functional disability. This disability can lead to difficulties in performing simple physical activity and everyday tasks such as cleaning, cooking, and dressing. Patient outcomes have not been fully explained by laboratory or radiographic measures, and as such, disability assessment offers an important component of disease activity characterization.

The Health Assessment Questionnaire-Disability Index (HAQ-DI) published in 1980 by Fries et al (1) has been used extensively in the evaluation of disease-specific disability or quality of life (QOL) related to RA in the United States and other countries (2–4). Both observational studies (2–4) and clinical trials (5–7) have used the HAQ-DI and found its scores to be an important predictor of work disability (8), morbidity (8,9), and mortality (10). Additionally, a recent study (11) has shown that the modified HAQ (12) correlated with a latent construct of physical disability at 0.87. Latent constructs represent constructs of interest that can not be measured directly (e.g., measurement of one's preference).

Notwithstanding the widespread use of the HAQ-DI and its appropriate reliability and convergent validity, empirical support for the factor structure and scoring system of the HAQ-DI is limited. In other words, the structural va-

Supported in part by grants from the National Institutes of Health MH55253, T32-MH18399, AG18367, AT00255, AR/AG41867, AR049840, and M01 RR00827. Dr. Khanna's work was supported in part by the Arthritis and Scleroderma Foundations (Physician Scientist Development Award).

Jason C. Cole, PhD, Sarosh J. Motivala, PhD, Dinesh Khanna, MD, MS, Jessica Y. Lee, Harold E. Paulus, MD, Michael R. Irwin, MD: University of California, Los Angeles.

Address correspondence to Michael R. Irwin, MD, Cousins Center for Psychoneuroimmunology, UCLA Neuropsychiatric Institute, 300 UCLA Medical Plaza, Room 3109, Los Angeles, CA 90095-7076. E-mail: mirwin1@ucla.edu.

Submitted for publication November 9, 2004; accepted in revised form March 15, 2005.

lidity of the HAQ-DI has not been adequately assessed; structural validation is important so that one can understand how to score and interpret the HAQ-DI (13). For example, the HAQ currently uses a single total score, yet it is not known whether the interrelationship among domains that comprise the HAQ support the use of a single score or whether multiple scores should be obtained. Daleo et al (14) and Kaufman (15) have noted that the scoring system of a measure should reflect its latent structure: if a measure has 3 factors, 3 scores should be calculated and interpreted. For example, the popular Center for Epidemiologic Studies Depression Scale (CES-D) has long been viewed as having 4 factors, but only 1 score is calculated (16), yielding what Daleo et al and Kaufman would deem as an inappropriate scoring system given the factor structure. Cole et al (17) demonstrated that the CES-D has a single hierarchical factor that subsumed the 4 factors in the CES-D, and therefore provided the first empirical evidence that the CES-D should be interpreted as a single score based on the refined factor structure.

Only 1 study detailing a factor analysis of the HAQ-DI was found (18) during an exhaustive review of the published literature using PubMed, PsychInfo, and Social Science Citation Index. Although Daltroy et al (18) found that a single dominant factor comprised the factor structure of the HAQ-DI, their results were generated using only exploratory factor analysis (EFA) even though confirmatory factor analysis (CFA) is now regarded as essential after the initial development of a measure (19). The use of CFA tests the viability and stability of the underlying construct(s) being evaluated (20,21). Indeed, CFA should be used as part of the process when determining the structural validity of any previously validated measure (21), and optimally EFA and CFA can be integrated using a cross-validation strategy (20).

The goal of the current study was to examine the structural validity of the HAQ-DI using a cross-validation approach with EFA and CFA. CFA was used to compare the structure obtained through EFA with other logical structures for the HAQ-DI. The results were used to guide and clarify scoring and interpretation procedures for HAQ-DI domains.

SUBJECTS AND METHODS

Subjects. Subjects were a subset of individuals with RA participating in a longitudinal study involving the Western Consortium of Practicing Rheumatologists, which is a regional consortium of 29 rheumatology practices in the western United States and Mexico as described in previous studies (22,23). The inclusion criteria for this study included a diagnosis of RA as defined by the American College of Rheumatology (formerly the American Rheumatism Association) criteria (24) within 15 months of symptom onset, no previous disease-modifying antirheumatic drug treatment, rheumatoid factor seropositive (RF titer $\geq$1:80 or $\geq$40 IU/ml), $\geq$6 swollen joints, and $\geq$9 tender joints. Symptom onset was defined as the date when musculoskeletal symptoms began, provided that these symptoms persisted and led to the diagnosis of RA. This study was approved by the appropriate institution review boards.

The consortium rheumatologists assessed patient disease status at study entry (baseline), 6 months, 1 year, and yearly thereafter. Using standard methods, detailed physician assessment included all of the core set outcome measures required to calculate the disease activity score (DAS), including complete tender and swollen joint counts and acute phase reactant measures, as well as 0–100-mm visual analog scales for global and pain assessments. The DAS was calculated according to the published algorithm using the Ritchie index, swollen joint count of 44 joints, and Westergren erythrocyte sedimentation rate (ESR) in mm/hour (25). In addition, study visits included radiographs of the hands, wrists, and forefeet; assays for RF; and self-report measurements such as the HAQ-DI and the CES-D (26). At each scheduled physician visit, blood specimens were collected to determine C-reactive protein levels; ESR was determined, when clinically indicated, in the rheumatologist's office or local laboratory.

Measures. The HAQ-DI is a condition-specific measure of functional status or QOL (measuring activities of daily living) intended for use in arthritis (1). The original HAQ-DI was designed as a 20-item self-administered questionnaire that examined difficulties with the performance of activities of daily living on a 0–3 scale in 8 domains (dressing and grooming, arising, eating, walking, hygiene, reach, grip, and other activities). A grade 3 of difficulty is assigned to patients using assistive/adaptive devices (such as canes, walker). The HAQ-DI score is calculated by summarizing the highest score in each of the 8 domains and dividing the sum by 8, resulting in a score range of 0 (no disability) to 3 (severe disability) on an ordinal scale.

Statistical analysis. Data were entered and cross checked using SPSS version 11.5 (SPSS, Chicago, IL) by research assistants with ample data entry experience. HAQ-DI scores were obtained per the instructions of Bruce and Fries (27). A single-extraction variant of the multiple imputation procedure for missing data replacement (28) was conducted for the missing points using NORM software (29). Multiple imputation uses a regression-type approach to estimate each missing datum. Imputed values are generated taking into account responses from the same participant on other correlated variables and responses to the same domain from participants who responded similarly. Using such multiple imputation formulae, Rubin and Schenker (30) have demonstrated that single imputation yields virtually identical results to that of the more laborious multiple database process. HAQ-DI domain descriptive statistics are listed in Table 1.

Because the relationship between many health-outcome variables is typically nonnormal (17,31), adjustments need to be made to control for nonnormality for any latent analysis using maximum likelihood estimation (MLE). Bootstrapping was used during model estimation to control for multivariate nonnormality (32,33). The process of

Table 1. Health Assessment Questionnaire domain correlations and descriptive statistics*

	1	2	3	4	5	6	7	8
1. Dressing and grooming	—	0.67	0.63	0.58	0.64	0.61	0.67	0.66
2. Arising		—	0.55	0.65	0.62	0.58	0.62	0.65
3. Eating			—	0.57	0.53	0.73	0.70	0.69
4. Walking				—	0.63	0.65	0.63	0.70
5. Hygiene					—	0.59	0.55	0.66
6. Reaching						—	0.69	0.72
7. Grip							—	0.73
8. Activity								—
Mean ± SD score†	1.00 ± 0.76	0.98 ± 0.76	1.07 ± 0.94	0.89 ± 0.83	1.17 ± 1.01	1.28 ± 0.99	1.05 ± 0.85	1.22 ± 0.88

* Correlations provided for descriptive purposes and were not analyzed for significance. All data were based upon multiple imputation data replacement database.
† Score range for all domains was 0–3. Mean ± SD score for all 278 participants was 1.17 ± 0.70.

bootstrapping takes multiple random subsamples from the current sample to smooth over any inaccuracies in the estimates of model fit due to nonnormality.

Exploratory factor analysis. One randomly divided subsample (n = 134) of the total sample was analyzed with EFA using SPSS version 12.0 (SPSS). Principal components analysis was used to determine the number of factors to retain for the EFA, per the recommendations of Preacher and MacCallum (19). In doing so, we examined the scree plot (a plot of eigenvalues, or the strength of a factor, to the number of factors – when the plot line becomes flat, factors to the right are considered useless) along with the Kaiser-Guttman criterion (eigenvalues >1.0 should be kept; see reference 19). Subsequently, EFA was carried out using MLE extraction factor analysis with direct oblimin rotation (a type of oblique rotation), as suggested by Preacher and MacCallum (19). The EFA analyses generate factor loadings, which are measures of how strongly the observed variables in the HAQ-DI are associated with its latent factor(s). Factor loadings for each domain were compared with criteria established by Comrey and Lee (34): values ≥0.71 signify excellent loadings, 0.63–0.70 are very good, 0.55–0.62 are good, 0.45–0.54 are fair, 0.32–0.44 are deemed poor, and any values <0.32 are discarded.

Confirmatory factor analysis. Once the EFA was completed, a CFA was undertaken in the second subsample (n = 144) to test the stability and replicability of the latent model produced by the EFA (Figure 1). Therein, the rectangular blocks represent HAQ-DI domains with circles to their left that represent each domain's residual (i.e., anything not measured by the relationship between the HAQ-DI domain and the latent variable). The circular figure to the right of the domains represents the overall HAQ-DI latent variable of disease impact.

CFA was performed using the AMOS statistical software package (35). MLE extraction was used to estimate the CFA model. The purpose of the CFA was to determine whether the EFA-derived model provided sufficient goodness-of-fit with the data in the second subsample, thus providing evidence for the stability of the model (e.g., how closely the model's purported covariance matrix fits with the actual covariance matrix of the subsample). Schumacker and Lomax (13) suggest that it is best to review multiple measures of model-data fit to examine the model from various perspectives. Therefore, in the current study, 4 fit indexes were used: Goodness-of-Fit (GFI), Adjusted Goodness-of-Fit (AGFI), Comparative Fit Index (CFI), and Root Mean Squared Error of Approximation (RMSEA). GFI and AGFI were evaluated with a minimum criterion of 0.90 (36), and CFI should be no less than 0.95 (37). RMSEA yields both a score and a 90% confidence interval; good fit would be indicated when the scores at the lower bound are ≤0.06 (38). GFI and AGFI are used to estimate strengths of association; GFI measures the association between the model and data, whereas AGFI adjusts GFI by taking into account the degrees of freedom (df) in a model (GFI can be inflated by high df). CFI and RMSEA provide estimates of Type I and Type II error, respectively. CFI is a measure of Type I error in that it specifies the amount of difference between the examined model and the independence model (i.e., a standard comparison model that asserts none of the components in the model are related), with higher scores indicating larger differences; RMSEA is complimentary to CFI because it is a measure of Type II error, determining the difference between the examined model and the saturated model (i.e., another standard model that asserts each of the components in the model are related to all other components in the model), with lower scores indicating

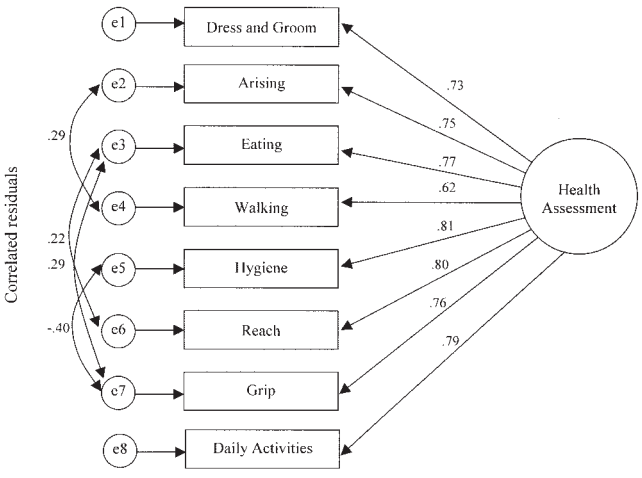

Figure 1. Health Assessment Questionnaire single-factor model from the confirmatory factor analysis.

greater differences. Ideally, the examined model should be markedly different from the independence model and the saturated model. If each of these 4 fit indices meet or surpass these thresholds, then the model can be considered satisfactory.

Model refinement. Often a model's fit indices may come close to reaching the abovementioned thresholds, but not close enough to be considered satisfactory. In such a case, minor adjustments to the relationships in the model can be made and the model can then be retested. The determination of which adjustments to make can be guided by using modification indices, which provide an estimate of the improvement in model fit that will occur by adding a given relationship, including direct paths and correlations (13). A standard approach of using a modification index of at least 10.0 was used; paths with a modification index <10 were considered to be too weak to provide substantive benefit. Modification of the model after an initial analysis will only be conducted if the modification meets statistical criteria and fits with the theoretical understanding of the HAQ-DI (13). When modifications are added to a model, the model will be rerun and interpreted with the new fit indices (39).

RESULTS

A total of 315 participants were admitted into the study, of which 27 participants (12%) were missing responses on 3 or more domains on the HAQ-DI. These subjects were removed from the database to allow for proper use of missing data replacement techniques (40). No more than 20% of missing data for any domain were found after removal of the 27 participants. According to Schafer and Graham (28), data should be missing at random to use missing data replacement appropriately. The presence of random or nonrandom missing data can be ascertained by examining the patterns of missing data to ensure that no one pattern (or patterns) is particularly likely over other patterns of missing data. A review of the current data found no such patterns, suggesting that these data are accurately described as missing at random.

The final sample comprised 278 participants with a mean ± SD age of 51 ± 13 years, a mean disease duration of 8.7 ± 10 months, and a mean HAQ-DI score of 1.17 ± 0.70. This sample size provided extensive power for the planned analyses (38). The descriptive statistics for the HAQ-DI domain scores are provided in Table 1, including correlations among the HAQ-DI domains as well as the mean ± SD and range of scores for each domain. Correlations among all of the domains were large (according to the criteria from Cohen [41]), ranging from the mid 0.50s to the low 0.70s. Each of the 8 HAQ-DI domain scores range from 0 to 3, with the means and SDs near 1.0 for most scales. Furthermore, HAQ-DI total scores ranged from 0 to 3 with a mean ± SD of 1.17 ± 0.70.

To provide an exploratory analysis of the HAQ-DI latent structure, an EFA was run on a randomly assigned sample. Results of the EFA are displayed in Table 2, where each domain is given a loading and a communality value. The loading refers to the correlation of a domain with the obtained latent factor, and the communality is the shared variance between the domain and the factor (i.e., the square of the loading). In other words, a high loading and communality mean that the domain has a strong relationship with the latent factor. Additionally, Table 2 shows that 68.4% of the variance was accounted for by a single factor, suggesting that this single dominant factor alone comprised the latent structure of the HAQ-DI. The single-factor structure was favored over the next most-viable model, a 2-factor structure, because examination of the eigenvalues showed a sharp decrease from 5.47 (68.38% variance explained) for the single factor to 0.61 (7.62% variance explained) for the 2-factor model. Additionally, inspection of the scree plot revealed that the scree was obtained at 2 factors (indicating a single-factor structure). All 8 HAQ domains had excellent loadings (ranging from 0.74 to 0.87).

Table 2. Factor matrix for the one-factor solution*

HAQ domain	Loading	h^2
Dressing and grooming	0.78[A]	0.62
Arising	0.76[A]	0.58
Eating	0.80[A]	0.63
Walking	0.79[A]	0.62
Hygiene	0.74[A]	0.55
Reaching	0.82[A]	0.68
Grip	0.83[A]	0.68
Activity	0.87[A]	0.76
Percent of total variance	—	68.38

* HAQ = Health Assessment Questionnaire; h^2 = communality for maximum likelihood estimation extraction; A = excellent loading.

Based on the single-factor solution of the EFA, a CFA was run solely on this single-factor model using the other random half of the sample. In this CFA, the single-factor model was close but did not meet adequate fit criteria (GFI = 0.89, AGFI = 0.80, CFI = 0.93, RMSEA = 0.13). Whereas CFI was nearly acceptable, RMSEA was not. These results indicate that the latent structure was missing some significant relationships and that minor adjustments in the model were needed. Thus, to find unmodeled paths that have both statistical and theoretical importance to the HAQ-DI model (13), modification indices were inspected. A modification index is a statistic that displays how much model fit will be improved by adding a new path to the model. In most models, paths can be added as unidirectional (i.e., regression paths) or bidirectional (i.e., correlational). Because the current model contained only a single factor, additional paths could only be added as correlations, specifically correlated residuals (42).

Model refinement: motor skills subdomain. Residuals refer to the variance that is not accounted for by the relationship of a particular domain to its latent variable. For example, the residual of the domain Grip is all of the variance not otherwise accounted for by the path between Grip and Health Assessment, or 1−0.76 for standardized values (Figure 1). This residual value is influenced by multiple other sources of variance, such as method variance, shared content beyond the primary factor, and mea-

Table 3. Fit statistics for all structural models*

Model	GFI	AGFI	CFI	RMSEA	RMSEA 90% CI
Single-factor (χ^2 = 20.62; 16 df)†	0.97	0.92	0.99	0.04	0.00–0.09

* GFI = Goodness-of-Fit; AGFI = Adjusted Goodness-of-Fit; CFI = Conformed Fit Index; RMSEA = Root Mean Square Error of Approximation; 90% CI = 90% confidence intervals; df = degrees of freedom.
† $P > 0.05$.

surement error (42). Hence, a correlation between 2 residuals occurs when aspects of these residual terms are strongly related, although correlations between residuals are not generally assumed to arise from correlated measurement error, as this should be random (43). The first examination of fit indices revealed relatively high scores between the residuals for Reach and Eating (modification index = 11.88), Grip and Eating (modification index = 14.63), and Arising and Walking (modification index = 14.61), resulting in correlations of r = 0.22, 0.29, and 0.29, respectively. All of these correlated residuals appear to have a content relationship in that each focuses on motor skills. Thus, after determining the HAQ score and disease-specific QOL, these data suggest that the additional impact on motor skills can be assessed by examining pairs of scores on Arising and Walking, then Grip and Eating, and finally Hygiene and Eating.

A second examination of modification indices was undertaken, after these 3 additions of the correlated residuals of the motor skills subdomain were added to the model. This second round of modification indices indicated that one more addition should be made by correlating the residuals between Grip and Hygiene (modification index = 12.20). However, it should be noted that the correlation between residuals for Grip and Hygiene was negative (−0.40). Whereas the other correlated residuals have a more logical interpretation, interpreting negatively correlated residuals between Grip and Hygiene is more elusive and should be examined further with other measures of manual dexterity and hygiene. Moreover, the Grip-Hygiene residual correlation was only appreciable once the previous correlated residuals were added, suggesting that the residuals of Grip and Hygiene have a complicated relationship to the first 3 combined correlated residuals.

No more sufficiently large correlated residuals were indicated, and therefore the model was rerun to test the new fit indices. The modified CFA model generated satisfactory fit statistics for all model fit criteria (Table 3). Figure 1 shows the final factor structure of the HAQ-DI, including the standardized factor loadings for the HAQ-DI latent variable on each of the HAQ-DI domains, as well as the level of standardized correlation between the domains. The fit for this model provides substantial evidence for the use of a single total score on the HAQ-DI.

DISCUSSION

The current study was the first to assess the latent structure of the HAQ-DI with rigorous methodologic tactics. Although prior EFA had been performed on the HAQ domains (18), those findings did not test the adequacy of how well their results fit their data given the limitations of the EFA. Beyond providing a confirmatory analysis of the HAQ's latent structure, this study also presented latent analysis in a 2-step cross validation. Because factor analysis is a sample-dependent technique, the validity of a factor structure must be tested on an independent sample for one to have confidence in the results.

The latent cross validation with EFA and CFA provides much support for the current scoring system of the HAQ-DI. Knowledge and validation of the latent structure of a measure is inextricably tied to the knowledge and validation of a scoring system for a measure. Hence, these data provide important and necessary validation for the way in which the HAQ-DI is scored. Although the results do not suggest that a new scoring for the HAQ-DI is required, new clinical data are provided to support secondary interpretations of the HAQ-DI based on residual correlations within a motor skills subdomain. However, caution in the interpretation of these pairs of scores within this model is needed. The total-score interpretation of the HAQ-DI is psychometrically the most appropriate interpretation of domain scores; secondary interpretations based on correlated residuals must only be done as embellishment and not as a replacement to the HAQ-DI total score. Second, the correlations between the residuals are moderate at best, and offer only a bit of useful information beyond the HAQ-DI total score (albeit, enough to mandate inclusion in the HAQ-DI model). Third, further validation of the correlated residuals should be undertaken before regular secondary interpretation of these factors is conducted. Such validation would require a new study that specifically tests the correlation interpretation, often necessitating the inclusion of additional variables in the model from measures of similar and dissimilar content (44).

A possible limitation to the current study is that the items for each HAQ-DI domain differ from person to person. This is a necessary and expected aspect of the HAQ-DI and all related psychometric evaluations of the HAQ-DI, because HAQ-DI scoring criteria require one to use the score of the highest item to create the score for the HAQ-DI domains. The influence of this aspect of the HAQ-DI should also be validated, and could be done within a hierarchical structural model. Unfortunately, this validation would require an immense and diverse sample that is rarely available in the study of RA.

Two key areas can be addressed in future research: the viability of the HAQ in other frequently assessed populations and determination of further scoring system information. The current study examined the HAQ with a sample exclusively with persons diagnosed with RA. However, the HAQ also is used frequently to determine the disease-

specific QOL in other populations, such as those with osteoarthritis, systemic lupus erythematosus, and other musculoskeletal conditions. At this time, there is no empirical evidence to suggest that the data obtained herein are necessarily generalizable to these other disease conditions (45). Byrne (46) has recommended that prior to the examination of similarity in CFA results for a measure across various subgroups, one should first determine the latent structure of the test on a single and appropriate sample. The current study supplies such information. Hereafter, it would be beneficial for other research to both affirm the latent structure of the HAQ for other disease populations and measure the consistency between those groups and an RA group (46).

A second method to further examine the HAQ scoring system is to use item response theory (IRT) (47). IRT weights each item so that items that indicate the strongest impact on health assessment receive stronger weights. A common IRT model used for this analysis is the Rasch model, which only examines the disease severity of each item in estimating the overall score for an individual (48). Like many IRT models, the Rasch model requires unidimensionality (i.e., a single-factor model). CFA is often used as a tool for determining the unidimensionality assumption in IRT (17) and the current study provides evidence that a single-factor model is appropriate, therefore suggesting that a Rasch model may work for the HAQ-DI. However, the Rasch model also asserts that the residuals of each item should be uncorrelated (49). Therefore, careful examination of the unidimensionality assumption is necessary to examine the HAQ-DI with IRT, including considering alternative IRT models (50).

In summary, this study provides psychometric evidence of the structural validity of the HAQ-DI within an RA population. Considering the widespread use of the HAQ-DI, it is important to demonstrate the psychometric stability and validity of the measure. By integrating EFA with a subsequent CFA, the current study demonstrates the validity of using the HAQ-DI total score as an estimate of disability in RA. Of course, as with all studies of validity, no one study can summarily prove the validity of a measure, as this must be done through a program of research. In the future, it would be of interest to determine whether the structural validity of the HAQ-DI extends to other rheumatic diseases.

REFERENCES

1. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
2. Kumar A, Malaviya AN, Pandhi A, Singh R. Validation of an Indian version of the Health Assessment Questionnaire in patients with rheumatoid arthritis. Rheumatology (Oxford) 2002;41:1457–9.
3. El Meidany YM, el Gaafary MM, Ahmed I. Cross-cultural adaptation and validation of an Arabic Health Assessment Questionnaire for use in rheumatoid arthritis patients. Joint Bone Spine 2003;70:195–202.
4. El-Miedany Y, Youssef S, el-Gaafary M, Ahmed I. Evaluating changes in health status: sensitivity to change of the modified Arabic Health Assessment Questionnaire in patients with rheumatoid arthritis. Joint Bone Spine 2003;70:509–14.
5. Bathon JM, Martin RW, Fleischmann RM, Tesser JR, Schiff MH, Keystone EC, et al. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis [published erratum appears in N Engl J Med 2001;344:240 and N Engl J Med 2001;344:76]. N Engl J Med 2000;343:1586–93.
6. Lipsky P, van der Heijde D, St. Clair W, Smolen J, Furst D, Kalden J, et al. 102-wk clinical & radiologic results from the ATTRACT trial: a 2 year, randomized, controlled, phase 3 trial of infliximab (Remicade®) in pts with active RA despite MTX [abstract]. Arthritis Rheum 2000;43:S269.
7. Weinblatt ME, Keystone EC, Furst DE, Moreland LW, Weisman MH, Birbara CA, et al. Adalimumab, a fully human anti–tumor necrosis factor α monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial. Arthritis Rheum 2003;48:35–45.
8. Wolfe F, Hawley DJ. The longterm outcomes of rheumatoid arthritis: work disability: a prospective 18 year study of 823 patients. J Rheumatol 1998;25:2108–17.
9. Wolfe F. The determination and measurement of functional disability in rheumatoid arthritis. Arthritis Res 2002;4 Suppl 2:S11–5.
10. Wolfe F, Michaud K, Gefeller O, Choi HK. Predicting mortality in patients with rheumatoid arthritis. Arthritis Rheum 2003;48:1530–42.
11. Escalante A, del Rincon I, Cornell JE. Latent variable approach to the measurement of physical disability in rheumatoid arthritis. Arthritis Rheum 2004;51:399–407.
12. Pincus T, Summey JA, Soraci SA Jr, Wallston KA, Hummon NP. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346–53.
13. Schumacker RE, Lomax RG. A beginner's guide to structural equation modeling. Mahwah (NJ): Lawrence Erlbaum; 1996.
14. Daleo DV, Lopez BR, Cole JC, Kaufman AS, Kaufman NL, Newcomer BL, et al. K-ABC simultaneous processing, DAS nonverbal reasoning, and Horn's expanded fluid-crystallized theory. Psychol Rep 1999;84:563–74.
15. Kaufman AS. Intelligent testing with the WISC-III. New York: Wiley; 1994.
16. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 1977; 1:384–401.
17. Cole JC, Rabin AS, Smith TL, Kaufman AS. Development and validation of a Rasch-derived CES-D short form. Psychol Assess 2004;16:360–72.
18. Daltroy LH, Phillips CB, Eaton HM, Larson MG, Partridge AJ, Logigian M, et al. Objectively measuring physical ability in elderly persons: the Physical Capacity Evaluation. Am J Public Health 1995;85:558–60.
19. Preacher KJ, MacCallum RC. Repairing Tom Swift's electric factor analysis machine. Underst Stat 2003;2:13–43.
20. Cole JC, Oliver TM, McLeod JS, Ouchi BO. Cross validating the latent structure of Accuplacer: a factor analytic approach. Res Schools 2003;10:63–70.
21. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assess 1995;7:286–99.
22. Paulus HE, Oh M, Sharp JT, Gold RH, Wong WK, Park GS, et al, and the Western Consortium of Praticing Rheumatologists. Correlation of single time-point damage scores with observed progression of radiographic damage during the first 6 years of rheumatoid arthritis. J Rheumatol 2003;30:705–13.
23. Paulus HE, Wiesner J, Bulpitt KJ, Patnaik M, Law J, Park GS, et al. Autoantibodies in early seropositive rheumatoid arthritis, before and during disease modifying antirheumatic drug treatment. J Rheumatol 2002;29:2513–20.
24. 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.
25. Van der Heijde DM, van 't Hof MA, van Riel PL, Theunisse LA, Lubberts EW, van Leeuwen MA, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 1990;49:916–20.

26. Blalock SJ, DeVellis RF, Brown GK, Wallston KA. Validity of the Center for Epidemiological Studies Depression scale in arthritis populations. Arthritis Rheum 1989;32:991–7.
27. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: dimensions and practical applications. Health Qual Life Outcomes 2003;1:1–6.
28. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77.
29. Schafer JL. NORM. Version 2.03. url: http://www.stat.psu.edu/~jls/misoftwa.html.
30. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Stat Med 1991;10:585–98.
31. Cole JC, Motivala SJ, Dang J, Lucko A, Lang N, Levin MJ, et al. Structural validation of the Hamilton Depression Rating Scale. J Psychopathol Behav Assess 2004;26:241–54.
32. Bollen K, Stine RA. Bootstrapping goodness-of-fit measures in structural equation models. Sociol Methods Res 1992;21:205–29.
33. Nevitt J, Hancock GR. Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. J Exp Educ 2000;68:251–68.
34. Comrey AL, Lee HB. A first course in factor analysis. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum; 1992.
35. Arbuckle JL. Amos. Version 4.02. Chicago: Small Waters; 2003.
36. Bentler PM, Bonett DG. Significance tests and goodness-of-fit in the analysis of covariance structures. Psychol Bull 1980; 88:588–606.
37. Hu LT, Bentler PM. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol Methods 1998;3:424–53.
38. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equation Model 1999;6:1–55.
39. Arbuckle JL, Wothke W. Amos 4.0 user's guide. 4.01 ed. Chicago: Small Waters; 1999.
40. Marcoulides GA. Introduction to structural equation modeling. In: Annual meeting of the American Educational Research Association, 1998; San Diego (CA): American Educational Research Association; 1998.
41. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale (NJ): Lawrence Erlbaum; 1988.
42. Palmer RF, Graham JW, Taylor B, Tatterson J. Construct validity in health behavior research: interpreting latent variable models involving self-report and objective measures. J Behav Med 2002;25:525–50.
43. Anastasi A, Urbina S. Psychological testing. 7th ed. Upper Saddle River (NJ): Prentice Hall; 1998.
44. Wothke W. Models for multitrait-multimethod matrix analysis. In: Marcoulides GA, Schumacker RE, editors. Advanced structural equation modeling: issues and techniques. Mahwah (NJ): Lawrence Erlbaum; 1996. p. 7–56.
45. Haynes SN, Richard DC, Kubany ES. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol Assess 1995;7:238–47.
46. Byrne BM. Structural equation modeling with AMOS: basic concepts, applications, and programming. Mahwah (NJ): Lawrence Erlbaum; 2001.
47. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of items response theory. Newbury Park (CA): Sage; 1991.
48. Wright BD. A history of social science measurement. Educ Meas Issues Pract 1997;16:33–45.
49. Linacre JM. Structure in Rasch residuals: why principal components analysis? Rasch Meas Trans 1998;12:636.
50. Van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer; 1997.