Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 51, No. 1, February 15, 2004, pp 14–19
DOI 10.1002/art.20091
© 2004, American College of Rheumatology

ORIGINAL ARTICLE

Issues in Cross-Cultural Validity: Example From the Adaptation, Reliability, and Validity Testing of a Turkish Version of the Stanford Health Assessment Questionnaire

AYSE A. KÜÇÜKDEVECI,[1] HÜLYA SAHIN,[1] SEBNEM ATAMAN,[1] BRIDGET GRIFFITHS,[2] AND ALAN TENNANT[3]

Objective. **Guidelines have been established for cross-cultural adaptation of outcome measures. However, invariance across cultures must also be demonstrated through analysis of Differential Item Functioning (DIF). This is tested in the context of a Turkish adaptation of the Health Assessment Questionnaire (HAQ).**

Methods. **Internal construct validity of the adapted HAQ is assessed by Rasch analysis; reliability, by internal consistency and the intraclass correlation coefficient; external construct validity, by association with impairments and American College of Rheumatology functional stages. Cross-cultural validity is tested through DIF by comparison with data from the UK version of the HAQ.**

Results. **The adapted version of the HAQ demonstrated good internal construct validity through fit of the data to the Rasch model (mean item fit 0.205; SD 0.998). Reliability was excellent (α = 0.97) and external construct validity was confirmed by expected associations. DIF for culture was found in only 1 item.**

Conclusions. **Cross-cultural validity was found to be sufficient for use in international studies between the UK and Turkey. Future adaptation of instruments should include analysis of DIF at the field testing stage in the adaptation process.**

KEY WORDS. HAQ; Rasch; Cross-cultural; Adaptation; Validity.

INTRODUCTION

During the last decade there has been a rapid development of instruments, often in the form of questionnaires, to measure the health status of patients undergoing treatment. Once developed, it is common for these instruments to be adapted for use in other cultures (1–3). The process of translation and validation for use in different cultures is referred to as cross-cultural validation, and guidelines have been published to facilitate a standard approach (4,5). Generally, the process involves a series of steps in the translation process, field testing, and then research to demonstrate the reliability and validity of the adapted instrument. A study of responsiveness usually follows at a later stage.

It is important to recognize what is being done in this process, and why. Much of the above procedure is concerned with providing a reliable and valid version of the instrument for a new culture. Implicit in the process is an attempt to make the new version a replica of the original, and consequently something that can be used in international clinical trials or other studies. However, modern psychometric approaches suggest that such a process is a necessary, but not sufficient condition for cross-cultural validity when the objective is to compare patients across different countries using adapted versions of the same instrument. In these conditions, a further requirement is that of invariance (6,7). Put simply, invariance means that the probability of a patient in 1 country affirming an item (in the dichotomous case) will be the same as the probability of a patient in another country affirming the item, given that they are both at the same level of the trait or construct being measured. Only under these conditions

Supported by the British Council and the Scientific and Technical Research Council of Turkey.

[1]Ayse A. Küçükdeveci, MD, Hülya Sahin, MD, Sebnem Ataman, MD: Ankara University, Ankara, Turkey; [2]Bridget Griffiths, MD: The Freeman Hospital, Newcastle Upon Tyne, United Kingdom; [3]Alan Tennant, PhD: University of Leeds, Leeds, United Kingdom.

Address correspondence to Dr. Ayse A. Küçükdeveci, Gökkusagi Sitesi, Mor Blok, 12/25, Karakusunlar 06530 Ankara, Turkey. E-mail: ayse@tepa.com.tr.

Submitted for publication January 11, 2002; accepted in revised form November 2, 2002.

can instruments be deemed to be equivalent in a measurement sense, facilitating the pooling of data and so on. Thus, the issue resolves to one of Differential Item Functioning (DIF) (8), which formally tests that such equivalence exists. Consequently, it is perfectly feasible to have a reliable and valid adaptation of an instrument that works well in a given culture, but which, in measurement terms, is not the same instrument when DIF is present for culture. Cross-cultural validation must include an examination of DIF for culture when the objective is to develop an instrument for use in international clinical trials.

The adaptation of the Stanford Health Assessment Questionnaire (HAQ) (9) for use in Turkey offers a chance to examine the implications of invariance for cross-cultural validity, and to consider how the relevant analysis may be subsumed into the adaptation process.

METHODS

The adaptation and validation of an instrument involves several stages. Initially, the translation process provides an initial version of the questionnaire. An examination of reliability usually follows, and finally construct validity. The latter stage will also provide information about the scale performance (for example, item total correlations), which may be used to compare the original and newly adapted instruments. More recently, the notion of internal construct validity has emerged, which is a more detailed examination of the structure of the scale, particularly related to unidimensionality, DIF, and scaling properties (10). Such an evaluation should follow the translation process before reliability is assessed, although the whole process may be subsumed into a single study involving sufficient patients to test both internal and external (construct) validity, as well as a test-retest phase for reliability and internal consistency. This phase may also provide information to explicitly examine cross-cultural validity by comparison of score level attributes between the original and adapted versions (11). The modern psychometric approach to this would be an examination of DIF by culture.

Translation process. An adaptation of the HAQ was made in Ankara University, Turkey to be used in a study investigating the correlation of radiographic joint damage with physical disability in rheumatoid arthritis (RA) (12). For the translation process, using the recent guidelines for cross-cultural adaptation (5), stage I involved 4 bilingual professionals translating the original version. One professional had a clinical background and was thus an "informed" translator. The other 3 translators were 2 English teachers in the university and a bilingual engineer (educated in the US), and were thus "uniformed" translators. Inconsistencies in the translations were resolved (stage II) by discussions between the translators. Back translation (stage III) and further expert review (stage IV) were not undertaken at that time. Following pretesting for face validity (stage V) in a group of 25 patients of variable educational levels with various musculoskeletal disorders, modifications had to be made in 5 items. Item 2 "shampoo your hair" was modified to "wash your hair," because both shampoo and soap are used in Turkey for washing hair. If taken literally, many respondents would have viewed the question as not relevant. Item 7 "open a milk carton" was changed to "open a new milk or a juice carton" because milk cartons were not very commonly used at that time. Item 10 "wash and dry your body" was modified to "wash and dry yourself" to adjust for nuances of the Turkish language. Also, for item 20 "Do chores such as vacuuming or yard work" was translated to "Do the housework such as sweeping the floor or gardening." In the Turkish language, the equivalent of "do chores" does not exist, and all the work associated with the home, internal or external, is considered as "housework." Finally, in item 13 the term "5 pounds" was changed to "2.5 kilos." The reliability and validity of this Turkish adaptation was not reported.

More recently, because of the increasing emphasis on the back translation as an important part of the adaptation process, and an imminent new study, it was thought worthwhile to undertake a back translation (stage III). Two uninformed bilingual translators were involved in this process. The expert review committee, comprising the back translators and 1 of the developers, was convened (stage IV). Slight differences were identified in the structure of the sentences of 2 items. For example, the Turkish version of item 13 ("reach and get down a 5 pound object, such as bag of sugar from just above your head") was back translated to "reach a 2.5 kilogram object (such as a sugar bag) above your head and get it down." However, the expert review committee thought it unnecessary to make further modifications to the existing adapted scale.

Internal construct validity. The principal modern psychometric approach used in health outcome measurement is Rasch analysis (13,14). The Rasch model is a unidimensional model that asserts that the easier the item the more likely it will be passed, and the more able the person, the more likely they will pass an item compared with a less able person. It assumes that the probability that a person will affirm an item or category within an item is a logistic function of the difference between the person's ability (θ) and the difficulty of the item (b), and only a function of that difference.

$$P_i(\theta)=\frac{e^{(\theta-b_i)}}{1+e^{(\theta-b_i)}}$$

where Pi (θ) is the probability that respondents with ability θ will answer item i correctly (or be able to do the task specified by that item), and b is the item difficulty parameter.

From this, the expected pattern of responses to an item set is determined given the estimated θ and b. When the observed response pattern coincides with or does not deviate too much from the expected response pattern, then the items constitute a true Rasch scale (15). Taken with confirmation of local independence of items, that is, no residual associations in the data after the Rasch trait has been removed, this confirms unidimensionality (16,17).

The formula can be expressed as a logit model:

$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right)=\theta_n-b_i$$

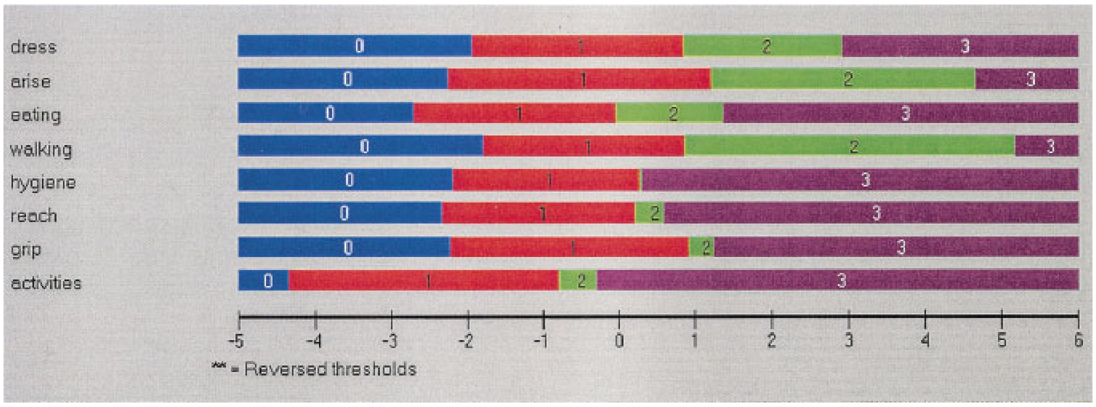

Figure 1. Category responses of each item (subscale) of the Health Assessment Questionnaire (0–3) across the underlying metric trait.

In other words, the log of the odds of a yes response, compared with a no response is, as stated previously, the difference between person ability and item difficulty. Thus, the Rasch analysis "currency" is the logit (log odds unit). A logit is the distance along the line of the variable that increases the odds of observing the event by a factor of 2.718, and it is this logit scale that provides interval level measurement for data that fit the model.

The model can be extended to cope with items with more than 2 categories (7), such as the HAQ, and this involves an explicit "threshold" parameter (τ), where the threshold represents the 0.5 probability point between any 2 adjacent categories within an item.

$$\ln\left(\frac{P_{nik}}{1-P_{nik-1}}\right)=\theta_n-b_i-\tau_k$$

It is easy to see how the approach was readily adopted in rehabilitation in the late 1980s (18–21). Patients undergoing rehabilitation have a given ability level, and they are presented with a range of tasks with differing degrees of disability. The language of ability and difficulty easily transferred from education to rehabilitation. From this early work, the approach quickly moved into the mainstream of health status measurement (10,22).

The early published work on Rasch analysis in rehabilitation explored issues of unidimensionality (18) and this has remained a central theme (23–25). However, Rasch analysis allows for much more than an empirical test for unidimensionality. Following Lord and Novick's work (26), and their explication of Item Response Theory, examination of DIF became routine. The basis of the DIF approach lies in the item response function, the S-shaped trace of the proportion of individuals at the same ability level who answer a given item correctly. Under the assumptions that the ability under consideration is unidimensional and that the item measures the same ability, then, except for random variations, the same curve is found irrespective of the nature of the group for whom a function is plotted (8). Thus, DIF is a formal test of invariance of the scale (across cultures) and this would be reflected in both similarity in slope of the response function of the item, as well as location (difficulty level of item). DIF can be considered to be uniform (where the same difference is observed across the trait), or nonuniform, where the difference in probability between groups differs across the trait.

Thus, this analysis is central to issues of cross-cultural validity, and using this approach, it is possible to make a formal test of whether or not a scale works in the same way across cultures. Consequently, in the present study, internal construct validity was tested by fit of the data to the Rasch model, and by testing for DIF for age, sex, disease duration, and culture. Due to the number of repeated tests, the significance level of 0.5 was adjusted by Bonferroni correction to 0.006.

Reliability. Where a scale is found to have internal construct validity, an examination of reliability can be made. For questionnaires of this type, it is usual to examine internal consistency through coefficient alpha, test-retest reliability, and intraclass correlation coefficient (ICC) (27). In the present study coefficient alpha and ICC were tested.

External construct validity. External construct validity is determined by testing for expected associations between the adapted instrument and other valid measures through the process of convergent construct validity (28). In this study, the following associations were considered: C-reactive protein (CRP), pain intensity by visual analog scale (VAS), duration of morning stiffness, and American College of Rheumatology functional stages (29).

Cross-cultural validity. Cross-cultural validity is examined by looking at the property of invariance through DIF analysis for culture. For purposes of this analysis, secondary analysis of a data set from the UK was used involving patients recruited to examine the relationship between genetic markers and disease severity in rheumatoid arthritis (30). The version of the HAQ used was adapted and validated by Kirwan and Reeback for use in Great Britain (31).

RESULTS

Characteristics of patients. Seventy-five outpatients, all meeting American College of Rheumatology (ACR; formerly American Rheumatism Association) 1987 revised criteria for RA (32), were recruited from the outpatient RA

Table 1. Differential Item Functioning by age, sex, duration, and country*

Item (subscale)	Age		Sex		Duration		Country	
	U	NU	U	NU	U	NU	U	NU
Dress	0.517	0.019	0.297	0.064	0.155	0.546	0.154	0.052
Arise	0.566	0.658	0.035	0.501	0.730	0.105	0.645	0.832
Eating	0.928	0.082	0.546	0.960	0.124	0.600	0.083	0.514
Walking	0.886	0.554	0.010	0.596	0.976	0.545	0.279	0.751
Hygiene	0.656	1.000	0.105	1.000	0.578	1.000	0.223	0.511
Reach	0.226	1.000	0.672	0.139	0.609	0.894	0.250	0.701
Grip	0.141	1.000	0.005	0.033	0.890	0.240	0.584	0.009
Activities	0.200	1.000	0.175	0.580	0.148	0.147	0.000	0.196

* Bonferroni adjusted level of < 0.006, expressed as significance level for each Uniform (U) and Nonuniform (NU) Differential Item Functioning.

clinic of a university hospital in Turkey. Their mean age was 49 years (SD 13.4 years), and 85% were female. The mean disease duration was 10.3 years (SD 9.3 years), the mean CRP was 48.4 mg/liter (SD 147.4), the mean pain intensity by VAS was 53.2 (SD 28.0), and duration of morning stiffness was 82.4 minutes (SD 98.4 minutes). The mean HAQ was 1.48 (SD 0.90).

In the original UK study, 174 patients were recruited; all met ACR criteria for RA, had a mean age of 51 years (SD 10.0 years), and 82% were female. Mean HAQ was 1.72 (SD 0.94).

Internal construct validity. The internal construct validity of the adapted Turkish version of the scale is confirmed by excellent fit to the Rasch model. Mean item fit was 0.205 (SD 0.998) and Person fit was 0.125 (SD 0.779), where fit statistics are standardized to a mean of 0 and standard deviation of 1. Consequently, observed data closely follow the model expectation, and the scale constitutes a true Rasch scale. Item trait interaction (χ^2 = 4.117, 8 degrees of freedom [df], $P = 0.846$) shows invariance of the scale for patients at different levels of disability. Person separation is high at 0.945, showing that the scale is able to discriminate across several groups of patients. The category structure of the scale is also working properly, with increases in item score between, for example, 2 and 3, representing an increase in disability on the underlying trait (Figure 1).

The scale is largely free of DIF for age, sex, and disease duration (Table 1). Only 1 item, grip, shows any significant difference for uniform DIF (Bonferroni corrected at 0.006) for sex. Otherwise, the scale items are invariant across groups and consequently the item response function is identical for the different groups (e.g., see Figure 2, the eating item).

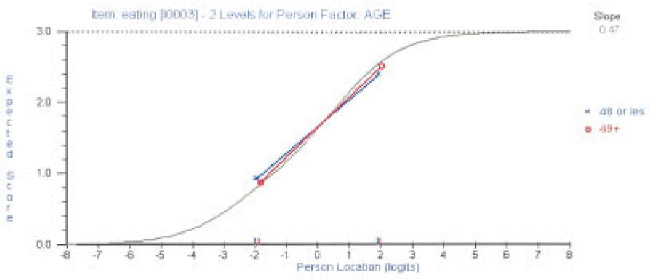

Figure 2. Differential item functioning for the "eating" item (subscale) by age. Color figure can be viewed in the online issue, which is available at http://www.arthritisrheum.org.

The scale has all the hallmarks of the classic ordinal scale. The thresholds are distributed unevenly across the construct with gaps between and clusters of thresholds (Figure 3). Patients will thus lose points (i.e., improve) in a haphazard manner and, depending on where they start on the scale (a high score is particularly vulnerable), may show either no ordinal-based improvement for some time, despite improvement of the metric scale (because their baseline position was just at the start of a long gap in the thresholds), or rapid ordinal improvement for little underlying metric improvement (because the baseline position was just above a cluster of thresholds).

Reliability. Internal consistency of the scale was assessed with coefficient alpha with a value of 0.97, which demonstrates adequate homogeneity of items in the scale. The ICC (one way effect random model) (33) was 0.95.

External construct validity. Correlation (Spearman's rho) between the Turkish version of the HAQ and CRP was 0.44; correlations of 0.33 for pain intensity (VAS) and 0.68 for duration of morning stiffness were found. The strengths of these correlations are as expected for the association between impairments and disability. A strong association was found between ACR functional stages and the HAQ (Kruskal-Wallis $\chi^2 = 55.8$, $P < 0.01$).

Cross-cultural validity. The cross-cultural validity of the scale is formally tested by checking the invariance of the scale across different language versions. Data from the UK data set were first fitted to the Rasch model to ensure internal construct validity. The results were very similar to the Turkish version with excellent fit to the model. The mean item fit was 0.198 (SD 0.957) and Person fit was 0.218 (SD 1.005). Item trait interaction chi-square was 18.62 (df 24, $P = 0.772$), showing invariance across groups of patients. Person separation was excellent at 0.942. Similar results have been found previously (10).

Invariance across countries was supported by the absence of DIF across all items except "activities" (Table 1). For this item, patients at the same level of disability in

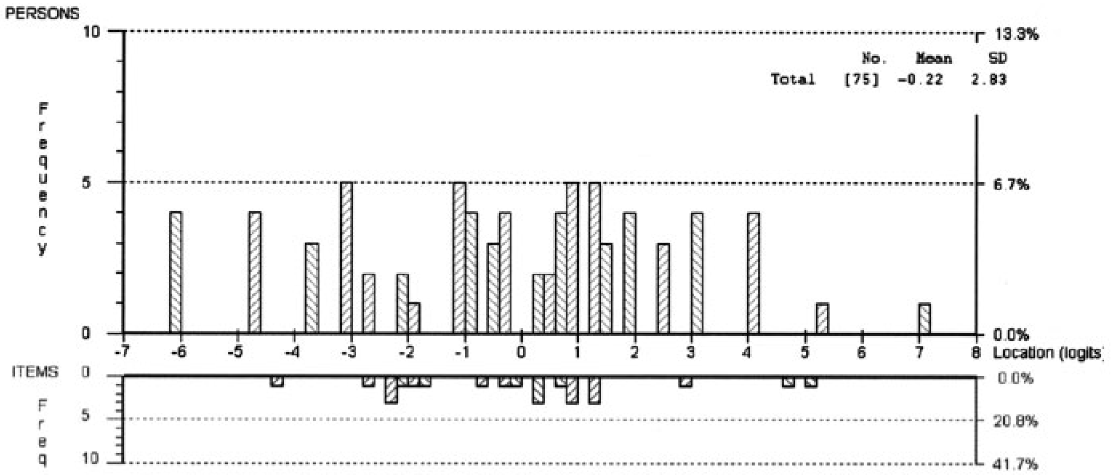

Figure 3. Distribution of persons and item (subscale) thresholds across the physical disability metric trait. (Grouping set to interval length of 0.2 making 75 groups).

Turkey will give a slightly higher response to this item (a consistent difference indicating uniform DIF; Figure 4).

Given the unique way in which the HAQ is scored, that is 20 underlying tasks contributing to the 8 items (subscales), it may be instructive to look explicitly at the tasks that contribute to the activities item. The pattern of difference between cultures found overall for the activities item is apparent for each of the 3 tasks that make up the item: “run errands,” “get in and out of car,” and “do chores.” No individual task is mostly responsible for this slight variation in response function between cultures.

DISCUSSION

Although guidelines have been produced for cross-cultural adaptation of instruments (5), there is not yet clear consensus on the most appropriate approach. For example, Herdman and colleagues have recently proposed a universalist approach to equivalence in the cultural adaptation of health-related quality of life instruments (34). This was in response to what they describe as an absolutist approach, which, they argue, makes unsupportable assumptions about the equivalence of concepts across cultures, instead concentrating on relatively technical issues of linguistic equivalence. This Turkish adaptation falls somewhere between the 2 approaches, incorporating conceptual equivalence and adjustment of language for variable educational levels into the technical process.

Irrespective of the approach to adaptation, from a measurement perspective, where the purpose is to adapt instruments for use in international studies, only invariance is a sufficient condition for cross-cultural validity. Only under these circumstances will equivalent scores represent equivalent levels of the construct across countries. Fitting data to the Rasch model allows a formal test of this property of fundamental measurement (35) but has the added advantage that when data do fit the model, a linear transformation of the construct (in this case physical disability) is achieved.

It is possible, using this approach, to have versions of the instrument that are valid in each country, but that work differently, thus negating cross-cultural use. Alternatively, it is possible to have scales that are comprised of quite different worded items (and technically it could be completely different items through item banking [36]) but that share the same item response function, so facilitating cross-cultural use. It is also possible that extensive adaptations of an instrument, such as the HAQ, may result in several groups of countries, where some meet cross-cultural requirements within groups but not across groups. This is a matter for empirical investigation. What is crucial is that some agency or person (which may be the original developer) should take responsibility for collating the information about adaptations and the level of cross-cultural utility (stage VI of the adaptation process).

For future adaptation of instruments, the question arises as to if, how, and when tests of invariance by culture should be incorporated into the adaptation process. The standard approach as defined by Beaton and colleagues (5) is a necessary, but not sufficient, condition for cross-cultural validity. Undertaking DIF analysis after the closure of the translation procedure (stage VI) would simply confirm the failure of the process, and it would seem that the challenge is to incorporate the analysis into the existing stages. An iterative loop may be necessary within stage V to identify problems and subsequently return to the field

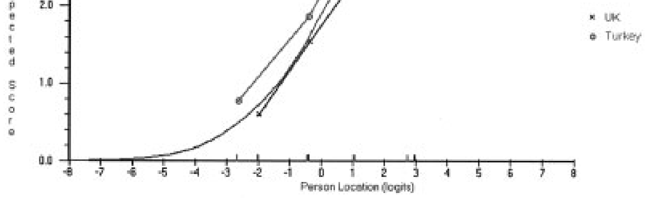

Figure 4. Differential item functioning of “activity” item (subscale) by country.

testing stage to collect further data (and items may be further modified at this point), and repeat the process until such a time that DIF is absent. Alternatively, a more sophisticated item banking approach allowing for variations of item difficulties across cultures may be utilized. Whichever approach is adopted, it is necessary to have data from the original language version available for the DIF analysis. Given this, we would propose that DIF by culture is incorporated into stage V as a formal test to demonstrate that the adaptation process has been successful.

In summary, the Turkish adaptation of the HAQ shows good internal consistency and construct validity. The adaptation process, although somewhat disjointed over a period of time, eventually conformed to recent guidelines for scale adaptation. Some conceptual modifications were required, and language had to be adjusted to accommodate the wide educational levels found in the Turkish population. Modern psychometric methods confirmed a level of crosscultural validity sufficient for use in international studies between the UK and Turkey. Future adaptation of instruments should include this analysis at the field testing stage, and if necessary, enter an iterative loop to ensure the absence of cross-cultural DIF. A Web site collating all this relevant information would help the rheumatology community in its quest for truly cross-cultural instruments for use in international clinical trials.

REFERENCES

1. Badia X, Alonso J. Rescaling the Spanish version of the Sickness Impact Profile: an opportunity for the assessment of cross-cultural equivalence. J Clin Epidemiol 1995;48:949–57.
2. Wiesinger GF, Nuhr M, Quittan M, Ebenbichler G, Wöhlfl G, Fialka-Moser V. Cross-cultural adaptation of the Roland-Morris Questionnaire for German-speaking patients with low back pain. Spine 1999;24:1099–103.
3. Salaffi F, Piva S, Barreca C, Cacace E, Ciancio G, Leardini G, et al. Validation of the Italian version of the Arthritis Impact Measurement Scales 2 (ITALIAN-AIMS2) for patients with osteoarthritis of the knee. Rheumatology 2000;39:720–7.
4. Guillemin EG, Bombardier C, Beaton D. Cross cultural adaptation of health related quality of life measures: literature review and proposed guidelines. J Clin Epidemiol 1993;46: 1417–32.
5. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;25:3186–91.
6. Thurstone LL. A method of scaling psychological and educational tests. J Educ Psychol 1925;16:433–51.
7. Andrich D. Relationships between the Thurstone and Rasch approaches to item scaling. Appl Psychologic Meas 1978;2: 449–60.
8. Angoff WH. Perspectives on differential item functioning methodology. In: Holland PW, Wainer H, editors. Differential item functioning. Hillsdale (NJ): Lawrence Erlbaum; 1993. p. 3–23.
9. Fries JF, Spitz P, Kraines RK, Holman H. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23:137–45.
10. Tennant A, Hillman M, Fear J, Pickering A, Chamberlain MA. Are we making the most of the Stanford Health Assessment Questionnaire? Br J Rheumatol 1996;35:574–8.
11. Ware JE Jr, Gandek B. Methods for testing data quality, scaling comparisons and reliability: the IQOLA Project approach: International Quality of Life Assessment. J Clin Epidemiol 1998;51:945–52.
12. Dalyan M, Küçükdeveci A, Evcik E, Ergin S. Radiological joint damage in rheumatoid arthritis: relationship with certain variables. J Ankara Med Sch 1996;18:33–8.
13. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960 (reprinted 1980).
14. Andrich D. Rasch models for measurement. Thousand Oaks (CA): Sage Publications; 1988.
15. van Alphen A, Halfens R, Hasman A, Imbos T. Likert or Rasch? Nothing is more applicable than a good theory. J Adv Nurs 1994;20:196–201.
16. Banerji M, Smith RM, Dedrick RF. Dimensionality of an early childhood scale using Rasch analysis and confirmatory factor analysis. J Outcome Meas 1997;1:56–85.
17. Smith RM. Fit analysis in latent trait measurement models. J Appl Meas 2000;2:199–218.
18. Silverstein B, Kilore KM, Fisher WP, Harley JP, Harvey RF. Applying psychometric criteria to functional assessment in medical rehabilitation: I. Exploring unidimensionality. Arch Phys Med Rehabil 1991;72:631–7.
19. Silverstein B, Fisher WP, Kilgore KM, Harley P, Harvey RF. Applying psychometric criteria to functional assessment in medical rehabilitation: II. Defining interval measures. Arch Phys Med Rehabil 1992;73:507–18.
20. Fisher AG. The assessment of IADL motor skills: an application of many-faceted Rasch analysis. Am J Occup Ther 1993; 47:319–29.
21. Fisher WP, Fisher AG. Application of Rasch analysis to studies in occupational therapy. Phys Med Rehabil Clin North Am 1993;4:551–69.
22. Haley SM, McHorney CA, Ware JA. Evaluation of the SF-36 physical functioning scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. J Clin Epidemiol 1994;47:671–84.
23. Tennant A, Geddes JLM, Chamberlain MA. The Barthel Index: an ordinal score or interval level measure? Clin Rehabil 1996; 10:301–8.
24. Prieto L, Alonso J, Lamarca R, Wright BD. Rasch measurement for reducing the items of the Nottingham Health Profile. J Outcome Meas 1998;2:285–301.
25. Shulman JA, Wolfe EW. Development of a nutrition self-efficacy scale for prospective physicians. J Appl Measure 2000;1:107–30.
26. Lord FM, Novick MR. Statistical theories of mental test scores. Reading (MA): Addison-Wesley; 1968.
27. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 2nd Edition. Oxford: Oxford University Press; 1995.
28. Nunnally JC. Psychometric theory. New York: McGraw-Hill; 1978.
29. Hochberg MC, Chang RW, Dwosh I, Lindsey S, Pincus T, Wolfe F. The American College of Rheumatology 1991 revised criteria for the classification of global functional status in rheumatoid arthritis. Arthritis Rheum 1992;35:498–502.
30. Griffiths B, Situmayake RD, Clarke B, Salmon M, Emery P. Racial origin and its effect on disease expression and immunogenetics in patients with RA: a matched cross-sectional study. Rheumatology 2000;39:857–64.
31. Kirwan JR, Reeback JS. Stanford Health Assessment Questionnaire modified to assess disability in British patients with rheumatoid arthritis. Br J Rheumatol 1986;25:206–9.
32. 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.
33. Shrout PE, Fleiss JL. Intraclass correlation coefficient: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
34. Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. Quality Life Res 1998;7:323–35.
35. Ellis B. Basic concepts in measurement. Cambridge: Cambridge University Press; 1966.
36. Dobby J, Duckworth D. Objective assessment by means of item banking. In: Schools council examinations bulletin 40. London: Methuen Educational; 1979.