EXTENDED REPORT

Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty

L D Roorda, C A Jones, M Waltz, G J Lankhorst, L M Bouter, J W van der Eijken, W J Willems, I C Heyligers, D C Voaklander, K D Kelly, M E Suarez-Almazor

Ann Rheum Dis 2004;**63**:36–42. doi: 10.1136/ard.2002.001784

See end of article for authors' affiliations

Correspondence to: Dr L D Roorda, VU University Medical Centre, Department of Rehabilitation Medicine, PO Box 7057, 1007 MB Amsterdam, The Netherlands; ld.roorda@vumc.nl

Accepted 12 May 2003

Background: Cross cultural validity is of vital importance for international comparisons.
Objective: To investigate the validity of international Dutch-English comparisons when using the Dutch translation of the Western Ontario and McMaster Universities osteoarthritis index (WOMAC).
Patients and Methods: The dimensionality, reliability, construct validity, and cross cultural equivalence of the Dutch WOMAC in Dutch and Canadian patients waiting for primary total hip arthroplasty was investigated. Unidimensionality and cross cultural equivalence was quantified by principal component and Rasch analysis. Intratest reliability was quantified with Cronbach's α, and test-retest reliability with the intraclass correlation coefficient. Construct validity was quantified by correlating sum scores of the Dutch WOMAC, Arthritis Impact Measurement Scales (Dutch AIMS2), Health Assessment Questionnaire (Dutch HAQ), and Harris Hip Score (Dutch HHS).
Results: The WOMAC was completed by 180 Dutch and 244 English speaking Canadian patients. Unidimensionality of the Dutch WOMAC was confirmed by principal component and Rasch analysis (good fit for 20/22 items). The intratest reliability of the Dutch WOMAC for pain and physical functioning was 0.88 and 0.96, whereas the test-retest reliability was 0.77 and 0.92, respectively. Dutch WOMAC pain sum score correlated 0.69 with Dutch HAQ pain, and 0.39 with Dutch HHS pain. Dutch WOMAC physical functioning sum score correlated 0.46 with Dutch AIMS2 mobility, 0.62 with Dutch AIMS2 walking and bending, 0.67 with Dutch HAQ disability, and 0.49 with Dutch HHS function. Differential item functioning (DIF) was shown for 6/22 Dutch items.
Conclusions: The Dutch WOMAC permits valid international Dutch-English comparisons after correction for DIF.

The past 25 years have seen the development of a large number of health related quality of life (HR QOL) measurement instruments, and these instruments are being used increasingly in research, with a growing emphasis on multinational applications.[1] A goal of HR QOL instruments is the use of these instruments cross nationally which stems, in part, from the need to pool data from multinational studies.[2] In addition, there is increasing international collaboration in HR QOL research.[3]

The cross national use of HR QOL instruments has created a need for cross culturally valid instruments for outcome assessment.[3] Many studies that investigate cross cultural validity of an HR QOL instrument report on the careful translation process, and investigate the clinimetric properties (dimensionality, reliability, and construct validity) of the translated instrument. This is an approach based on classical test theory.[4] It deals with the issue of whether the items (questions) from the translated measurement instrument measure only one (dimensionality), and the same (reliability) construct that is related to other measures as hypothesised (construct validity). This approach, however, does not answer the key question of whether scores on the translated items can be compared with scores on the original items. For instance, Dutch patients with the same HR QOL as Canadian patients, and responding to the Dutch version of an HR QOL questionnaire should attain the same score as Canadian patients responding to the English version of the HR QOL questionnaire. Evaluating different cultural (including language) groups (for example, Dutch and English speaking patients) without significant bias requires a study of cross cultural equivalence.[5] The investigation of cross cultural equivalence is based on item response theory.[4] ''Equivalence'' in this context refers to the absence of item bias[5] or differential item functioning (DIF).

The Western Ontario and McMaster Universities osteoarthritis index (WOMAC) is increasingly used for international research in patients with hip osteoarthritis (OA).[6] Pain and physical functioning are among the major determinants of HR QOL in patients with OA.[7 8] The WOMAC measures these determinants by assessing five pain related activities and 17 functional activities.[9] The clinimetric properties of the original English version of the WOMAC are well known.[6 9] An increasing number of translations are available.[10–14] Recently, we have made a Dutch translation of the WOMAC.

The purpose of this study was to investigate the dimensionality, reliability, construct validity, and the cross cultural equivalence of the Dutch translation of the WOMAC in patients with OA waiting for primary total hip arthroplasty (THA).

Abbreviations: AIMS, Arthritis Impact Measurement Scale; CI, confidence interval; DIF, differential item functioning; HAQ, Health Assessment Questionnaire; HHS, Harris Hip Score; HR QOL, health related quality of life; ICC, intraclass correlation coefficient; OA, osteoarthritis; SD, standard deviation; SE, standard error; THA, total hip arthroplasty; WOMAC, Western Ontario and McMaster Universities osteoarthritis index

METHODS

Patients

An inception cohort of Dutch patients indicated for primary THA in two general hospitals (Onze Lieve Vrouwe Gasthuis and Medisch Centrum Alkmaar) and one university hospital (VU University Medical Centre) was assembled. Selection criteria were related to the time of placement on the waiting list rather than the time of surgery. We selected consenting patients who (*a*) were scheduled for elective primary THA; (*b*) were placed on the hospital waiting list between February and May 1996; (*c*) were aged 90 years or younger; (*d*) had a diagnosis of primary or secondary OA; (*e*) were able to complete a questionnaire; and (*f*) were living independently. Finally, the surgeon had to agree that the patient was invited to participate in the study.

Among the 245 Dutch patients on the waiting list, eight patients did not receive surgery (either for medical or personal reasons). In addition, 31 patients were excluded. Reasons for exclusion were: 2 were older than 90 years; 1 was not diagnosed with OA; 18 could not complete the questionnaire (language (5), emotional (2), cognitive (7), logistic (3, not living in the Netherlands) or other (1) reasons); 3 were not living in the community; 2 had surgeons who refused participation; and 5 for other reasons. Of the 206 patients who were eligible, 22 refused or could not be contacted to request participation, 1 was operated on in another hospital, and 3 were lost to follow up. The final Dutch study group comprised 180 patients, corresponding to a participation rate of 87% of the eligible patients.

For the purpose of studying cross cultural equivalence, data were used from the study of Jones and colleagues[15–18] because their study design and patient selection criteria were similar to those of the Dutch study. The Canadian study was a prospective longitudinal study that followed up a consecutive community based cohort of patients who were scheduled to receive a primary THA. Like the Dutch cohort, patients were enrolled when they were recommended for surgery. Selection criteria were related to time of placement on the health regional waiting list. The selection criteria included patients who (*a*) were scheduled for elective primary THA; (*b*) were placed on the health regional waiting list between December 1995 and January 1997; (*c*) were placed on the waiting list at least seven days before their surgery; (*d*) resided within the health region; and (*e*) were English speaking. Patients who resided in long term care institutions were excluded.

Among the 319 Canadian patients on the waiting list, 29 had their surgery cancelled either for medical reasons or by personal choice. Of the 290 eligible patients, 32 patients refused, could not be contacted, or had already had their surgery, and 14 patients were lost to follow up. The final Canadian study cohort comprised 244 patients, corresponding to a participation rate of 84% of the eligible patients. Although no upper age limit was defined, no patients were 90 years or older at the time of surgery. The vast majority of Canadian patients were diagnosed with primary or secondary OA.

The study protocol was reviewed and approved by the VU University Medical Centre Institutional Review Board, Ethics Committee, and the University of Alberta Health Research Ethics Board.

Measurements

Sociodemographic characteristics and diagnosis were extracted from the medical records. Furthermore, all Dutch patients completed the Dutch WOMAC, the Dutch Arthritis Impact Measurement Scales (AIMS2), and the Dutch Health Assessment Questionnaire (HAQ) on the same day, and within one week before surgery. The questionnaires were self administered for the Dutch patients. In addition, the Harris Hip Score (HHS) was administered by interview for the Dutch patients one or two days before surgery. The Canadian patients completed the English WOMAC within one month before surgery. The questionnaire was self administered with the interviewer assisting when needed for the Canadian patients.

The WOMAC is well tested, and its reliability, validity, and responsiveness are satisfactory.[6 9] The Dutch translation was made by two bilingual and bicultural translators, using a double (back-) translation procedure.[5 19] The WOMAC consists of three dimensions: pain (5 items), stiffness (2 items), and physical functioning (17 items). Because pain and physical functioning are the major determinants of HR QOL in patients with OA of the hip, only the pain and physical functioning dimensions of the WOMAC were studied. The five point Likert version[9] of the WOMAC was used. Item responses range from ''none'' to ''extreme''.

The AIMS2 is a revision of the AIMS,[20] a widely used, well tested, and highly recommended outcome measure for arthritis research.[21] It is expected that this revision of the AIMS will be the preferred form of the instrument.[21] A validated Dutch version is available.[22] The questionnaire assesses multiple dimensions. We used the arthritis pain scale (five items) in this study. In addition, we used the scales dealing with mobility (five items) and walking and bending (five items), because these scales specifically measure physical functioning related to the legs. Responses are based on a Likert scale with five response options ranging from ''all days'' to ''no days''.

The HAQ[23] is also a widely used, well tested, and highly recommended outcome measure for arthritis research.[21] Most published data of the HAQ concern the HAQ Disability Index. The clinimetric properties of the HAQ discomfort (pain) dimension are not very clear.[21] Pain is measured on a single 15 cm horizontal visual scale with terminal markers anchored to ''no pain'' and ''very severe pain''. The validated Dutch version of the HAQ[24] Disability Index assesses patients' functional ability, and covers eight fields of activity: dressing, arising, eating, walking, hygiene, reach, grip, and outside activities. The Dutch HAQ has 20 items with four response options ranging from ''independent without difficulty'' to ''completely dependent''.

The HHS[25] is a widely used, yet poorly tested, outcome measure for hip arthroplasty research. To our knowledge there is no official validated Dutch translation of the HHS. Pain is measured with one item and has six response options. Function (physical functioning) is measured with seven items, and has a varying number of response options (two to six).

Sum scores for each scale were calculated after imputing the corrected item mean in cases of item non-response.[26] Furthermore, scale sum scores were standardised (0–100), with high values indicating less pain or better physical functioning.

Statistics

Patients

Differences in sociodemographic characteristics between the Dutch and the Canadian patients were investigated with an independent samples *t* test (age) and a χ^2 test (sex).

Unidimensionality

Unidimensionality indicates that items assess a single underlying construct.[4] In this study unidimensionality was investigated by principal component analysis.[4] In a first analysis the dimensionality of the pain items was investigated. In a second analysis the dimensionality of the physical functioning items was investigated.

Unidimensionality was also investigated by Rasch rating scale analysis,[27–29] which relates to the item response theory.

The Rasch rating scale analysis provides estimates with standard errors (SEs) of person ability (measures) and item difficulty (calibrations) along a common measurement continuum. For instance, the analyses may demonstrate that a person with unilateral involvement of the hip has a higher level of ability than another person with bilateral involvement of the hips, and that item A (about walking stairs) is more difficult than item B (about rising from a chair). The person measures and item calibrations can be estimated independently of one another by means of conditional maximum likelihood. Person measures and item calibrations are expressed in log-odd units (logits). The logit is a unit of interval measurement which is defined within the context of a set of items.[30] Unidimensionality of an item set, and also compliance with the other assumptions of the Rasch model, is determined by the pattern of item goodness of fit statistics. The goodness of fit statistics are indices of how well the item calibration, as estimated for the entire sample, fits the data with respect to all of the individual subjects in the sample. In this analysis we report both the infit and the outfit statistic. The infit statistic focuses on the central performance of an item. The outfit statistic is an outlier sensitive fit statistic. Low fit statistics indicate that the item measures redundant or overlapping content areas. High fit statistics (residual between observed *v* predicted score), generally speaking fit statistics >1.30,[31] may indicate that the item is not as closely related to the overall construct as predicted. Rasch analyses were performed using BIGSTEPS, version 2.82.[30]

Reliability

Reliability concerns the degree to which the results of measurement are consistent across repeated measurements.[32] The intratest reliability or internal consistency of the Dutch WOMAC was quantified by Cronbach's α. Test-retest reliability was determined in a subgroup of patients who had been on the waiting list for longer than one month. This subgroup consisted of 28 patients who completed the questionnaire preoperatively twice within 22 days (range 8–28 days). To estimate the test-retest reliability of the Dutch WOMAC sum scores, intraclass correlation coefficients (ICCs) with 95% confidence interval (95% CI) were calculated, using a two way mixed model. Patients were considered to be random effects, while the measure effect was a fixed effect. The ICC is generally considered to be excellent at 0.75 and above.[33]

Construct validity

Construct validity is concerned with the extent to which a particular measure relates to other measures consistent with theoretically derived hypotheses for the constructs that are being measured.[32] In this study it was suggested that the standardised sum scores of Dutch WOMAC pain, Dutch AIMS2 pain, Dutch HAQ pain, and Dutch HHS pain would be moderately to strongly, $r>0.35$,[34] and positively correlated. Secondly, it was suggested that the standardised sum scores of Dutch WOMAC physical functioning, Dutch AIMS2 mobility, Dutch AIMS2 walking and bending, Dutch HAQ disability, and Dutch HHS function would be moderately to strongly, and positively correlated. To evaluate the construct validity of the Dutch WOMAC, Pearson's correlation coefficients with 95% CI[35] were calculated.

Cross cultural equivalence

Cross cultural equivalence refers to the equivalence of measurement across different cultural groups of people.[5] Equivalence of measurement necessitates a careful translation process.[5] In addition, one can examine the equivalence of measurement across culture and language by Rasch analyses[5] if the data fit the Rasch model. Having a means to determine item difficulty independently of person abilities enables one to detect and diagnose DIF.[5]

According to the recommendations of Cella and colleagues,[5] the Dutch WOMAC was administered to Dutch patients and the English WOMAC to Canadian patients. Owing to the careful translation process none of the Dutch items were expected to demonstrate DIF caused by the translation process. Calibration of the items was performed for the pain scale and the physical functioning scale, and separately for the Dutch and the English speaking patients. The calibrated item difficulties resulting from the separate analysis of each sample for each scale were centred and plotted against each other, with the Dutch items on the y axis and the English items on the x axis. An identity line was drawn through the origin of each plot, with a slope of 1. Statistical control lines with 95% CI are drawn around the identity lines to guide interpretation.[27] Finally, the plots are examined to determine whether any items fall outside the control lines, suggesting DIF.

Corrected sum scores

To make valid comparisons between Dutch and Canadian patients, sum scores corrected for DIF should be used. The influence of DIF on a patient's sum scores can be corrected when items fit the Rasch model. An important property of the Rasch model is that, generally speaking, the measurement of patients is ''test-free''.[27] This means that using different subsets of the items will result in approximately the same measure for a patient. Suppose that only item 3 of the five WOMAC pain items shows DIF. In such a case, the translated and the original item 3 are handled as two different items:[5] item 3D (the translated Dutch item about pain at night), and item 3E (the original English item about pain at night). Consequently, item 3 is entered into the dataset as two items (3D and 3E) with missing data coded on English speaking patients for the Dutch version (item 3D) and on Dutch speaking patients for the English versions (item 3E). Subsequently, the calibrated item difficulties from the combined analysis of both samples are calculated. The measurement of the Dutch patients is based on the following subset of items: items 1, 2, 3D, 4, and 5. The measurement of the Canadian patients is based on a different subset of items: items 1, 2, 3E, 4, and 5. Summarising, quantitatively sound Dutch-English comparisons can be made by creating two items (3D and 3E) from one item showing DIF (item 3), anchoring these two items by the calibrations of the other items not showing DIF (items 1, 2, 4, and 5), and using different subsets of items for the measurement of the Dutch and the Canadian patients.

Differences between the Dutch and the Canadian patient scores for WOMAC pain and physical functioning were investigated with an independent samples t test, while using the standardised (0–100) Rasch scores, corrected for DIF, with high values indicating less pain or better physical functioning.

RESULTS

Patients

The mean (SD) age of the Dutch patients was 67.9 (10.7) years. Sixty three (35%) of the Dutch patients were male. The mean (SD) age of the Canadian patients was 67.4 (11.9) years and 100 (41%) were male. Differences in age and sex between the Dutch and the Canadian patients were not significant (two tailed $p = 0.70$ and $p = 0.21$, respectively).

Unidimensionality

In the first principal component analysis the pain items of the Dutch WOMAC loaded on one component. In the second analysis the physical functioning items also loaded on one

Table 1 Item calibrations with standard error (SE), infit statistics, and outfit statistics for the items of the Dutch WOMAC and the English WOMAC pain and physical functioning dimensions, according to the item order of the WOMAC

	Dutch				English			
WOMAC	**Calibration**	**SE**	**Infit**	**Outfit**	**Calibration**	**SE**	**Infit**	**Outfit**
Pain								
1. Walking on a flat surface	−0.47	0.13	1.10	1.11	−0.11	0.10	0.78	0.80
2. Going up or down stairs	−1.14	0.14	0.90	0.89	−1.19	0.10	0.92	0.91
3. At night while in bed	0.63	0.13	*1.31*	1.29	0.28	0.10	1.23	1.24
4. Sitting or lying	0.85	0.13	0.73	0.73	1.03	0.10	0.83	0.82
5. Standing upright	0.13	0.13	0.85	0.83	−0.01	0.10	1.19	1.18
Physical functioning								
8. Descending stairs	0.27	0.12	1.04	1.03	0.19	0.10	0.98	0.97
9. Ascending stairs	−0.65	0.12	0.94	0.91	−0.58	0.10	0.83	0.84
10. Rising from sitting	0.00	0.12	0.58	0.57	−0.01	0.10	0.90	0.92
11. Standing	0.32	0.12	1.16	1.16	0.39	0.10	0.97	0.97
12. Bending to floor	−1.09	0.12	1.24	1.21	−0.77	0.10	1.02	0.97
13. Walking on flat	0.41	0.12	1.13	1.14	0.33	0.10	0.82	0.81
14. Getting in/out of car	−0.91	0.12	0.78	0.77	−0.69	0.10	0.67	0.68
15. Going shopping	−0.87	0.12	0.91	0.88	−0.90	0.10	0.80	0.77
16. Putting on socks/stockings	−0.82	0.12	*1.56*	*1.51*	−0.94	0.10	1.29	1.27
17. Rising from bed	0.56	0.12	0.71	0.70	0.59	0.10	0.76	0.76
18. Taking off socks/stockings	−0.14	0.12	1.20	1.18	−0.17	0.10	1.17	1.18
19. Lying in bed	1.29	0.12	1.23	1.21	0.99	0.09	*1.35*	*1.34*
20. Getting in/out of bath	0.28	0.14	1.10	1.11	−0.20	0.10	1.16	1.24
21. Sitting	1.54	0.12	1.00	0.96	1.56	0.09	0.96	0.98
22. Getting on/off toilet	0.51	0.12	0.68	0.67	1.03	0.09	1.11	1.09
23. Heavy domestic duties	−1.56	0.13	0.85	0.87	−1.72	0.10	1.11	1.06
24. Light domestic duties	0.86	0.12	0.73	0.73	0.90	0.09	1.04	1.02

Calibrations are expressed in logits. Negative calibrations indicate easier items; positive calibrations indicate more difficult items.
Infit and outfit statistics >1.30 indicate items that do not contribute to the underlying construct. Statistics >1.30 are shown in italics. Infit and outfit statistics <0.70 indicate items that are muted.

component. In Rasch analysis of the Dutch WOMAC (table 1) the infit statistic for one out of five pain items (pain at night while in bed) was above 1.30 (1.31), indicating that this item was not closely related to the underlying construct.[31] The SEs of the pain item calibrations ranged from 0.13 to 0.14 logits, indicating that the item calibrations were equally well assessed for the entire range of items. Sixteen of the 17 physical function items fitted the Rasch model (table 1). Both the infit (1.56) and the outfit (1.51) statistic were greater than 1.30 for one item (difficulty putting on socks/stockings) for the Dutch WOMAC. The SE of the item calibration of the physical functioning items ranged from 0.12 to 0.14 logits.

All pain items of the English WOMAC fitted the model (table 1); and 16 of the 17 items pertaining to physical function fitted the Rasch model (table 1). Both the infit (1.35) and the outfit (1.34) statistic were greater than 1.30 for the item concerning difficulty lying in bed. The SE of the item calibrations of the pain and physical functioning items ranged from 0.09 to 0.10 logits.

Reliability

Cronbach's α for Dutch WOMAC pain was 0.88, and for Dutch WOMAC physical functioning it was 0.96, indicating good intratest reliability.[4] The ICC (95% CI) for Dutch WOMAC pain and Dutch WOMAC physical functioning were 0.77 (0.56 to 0.89) and 0.92 (0.83 to 0.96), respectively.

Construct validity

The Pearson correlation coefficient (95% CI) of the Dutch WOMAC pain (mean (SD) 42.4 (20.5)) standardised sum score and the Dutch HAQ pain (mean (SD) 35.8 (22.3)) was 0.69 (0.60 to 0.76). A correlation value of 0.39 (0.26 to 0.51) was seen with the Dutch WOMAC pain and the Dutch HHS pain (mean (SD) 45.1 (17.3)) standardised sum scores. Dutch AIMS2 pain was dropped from the analysis owing to a high mean percentage of item non-response (48%) compared with the Dutch WOMAC pain (2%) and Dutch HAQ pain (6%). The correlation between the Dutch WOMAC physical functioning (mean (SD) 39.6 (18.9)) standardised sum score and the Dutch AIMS2 mobility standardised score (mean (SD) 69.2 (22.9)) was 0.46 (0.33 to 0.56), while the correlation with the standardised sum scores of the Dutch AIMS2 walking and bending (mean (SD) 25.5 (19.7)) was 0.62 (0.52 to 0.70), with the Dutch HAQ disability (mean (SD) 71.0 (15.5)) it was 0.67 (0.58 to 0.74), and with the Dutch HHS function (mean (SD) 56.8 (17.4)) it was 0.49 (0.35 to 0.60). All sum scores were moderately to strongly, and positively correlated, indicating a satisfactory construct validity.

Cross cultural equivalence

There was DIF for two out of five pain items of the Dutch WOMAC: item 1 (pain walking on a flat surface) was easier, and item 3 (pain at night while in bed) was more difficult than the original English item (fig 1). In addition, there was DIF for four of the 17 Dutch physical functioning items: item 12 (difficulty bending to floor) and item 22 (difficulty getting on/off toilet) were easier, whereas item 19 (difficulty lying in bed) and item 20 (difficulty getting in/out of bath) were more difficult than the original English items (fig 2).

Corrected sum scores

Table 2 summarises the sum scores for the Dutch and Canadian patients. The (standardised) corrected Rasch scores should be used to make valid Dutch-English comparisons. The mean (SD) standardised corrected Rasch pain score was 45.0 (18.3) for the Dutch patients and 45.6 (14.1) for the Canadian patients. The mean (SD) standardised corrected Rasch physical functioning score was 54.0 (16.6) for the Dutch patients and 53.3 (12.6) for the Canadian patients. Differences in pain and physical functioning scores between the Dutch and the Canadian patients were not significant (two tailed $p = 0.70$ and $p = 0.67$, respectively).

DISCUSSION

In this study we demonstrated unidimensionality and, with a few exceptions, good fit with the Rasch model for the pain

Table 2 Mean with SD WOMAC sum scores for the Dutch and Canadian patients

	Dutch		Canadian	
WOMAC	**Mean**	**SD**	**Mean**	**SD**
Pain				
Raw sum score	16.5	4.1	16.4	3.2
Standardised raw sum score	42.4	20.5	43.0	15.9
Corrected Rasch score	0.7	2.3	0.7	1.7
Standardised and corrected Rasch score	45.0	18.3	45.6	14.1
Physical functioning				
Raw sum score	58.1	12.9	58.9	10.1
Standardised raw sum score	39.6	18.9	38.4	14.9
Corrected Rasch score	0.9	1.8	1.0	1.4
Standardised and corrected Rasch score	54.0	16.6	53.3	12.6

Raw sum scores, with lower scores indicating less pain (range 5–25) or better physical functioning (range 17–85).
Standardised raw sum scores (range 0–100), with higher scores indicating less pain or better physical functioning.
Corrected Rasch scores, with lower scores indicating less pain (range −6.04–6.26) or better physical functioning (range −4.18–6.82).
Standardised and corrected Rasch scores (range 0–100), with higher scores indicating less pain or better physical functioning.

and physical functioning items of the Likert version of the Dutch WOMAC, and the Likert version of the English WOMAC. Good fit with the Rasch model was demonstrated in patients with hip OA scheduled for elective primary THA, who are patients with severe OA. An important property of the Rasch model is that the item calibrations are ''sample-free''.[27] This implies that, generally speaking, the use of patient groups with less pain or limitations in physical functioning, such as patients with mild OA, will result in similar item calibrations. In addition, the precision of the assessment of the item calibrations is calculated, and expressed as an SE. In this study, the SE of the item calibrations within one dimension, were similar, which indicates that all item calibrations were equally well assessed. So the inclusion of only patients with severe OA does not seem to have a major impact on the precision of the assessment of the item calibrations, or consequently on the presence of DIF.

Good fit of the WOMAC items with the Rasch model was also demonstrated in other studies. Wolfe and Kong demonstrated good fit with the Rasch model for the visual analogue scale version (converted to an 11 point Likert scale) of the English WOMAC in patients with knee OA (n = 655), including patients with mild OA.[8] They found misfit for two physical functioning items (item 8: difficulty descending stairs, and item 20: difficulty getting in/out of bath). Ryser and colleagues demonstrated good fit with the Rasch model for an 11 point Likert version of the German WOMAC in patients with hip and knee OA (n = 158).[36] A misfit for one item (item 3: pain at night while in bed) was reported, which was also the only misfitting pain item in the Dutch WOMAC. An explanation for the misfit of item 3 may be that the pain is nocturnal rather than daytime pain. Rojkovich and Gibson demonstrated in patients with rheumatoid arthritis that nocturnal pain scores correlated best with measures of joint inflammation, whereas daytime pain scores, both at rest and on movement, seemed to be influenced by the degree of permanent joint damage.[37] In summary, a good fit with the Rasch model was demonstrated for different translations and different versions of the WOMAC, in different patients

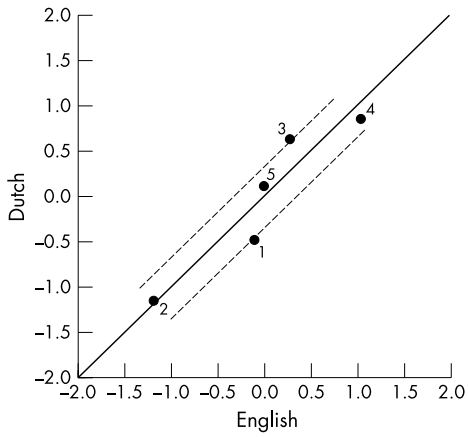

Figure 1 Pain item calibrations for the Dutch WOMAC and the English WOMAC. English calibrations on the x axis, and Dutch on the y axis. Calibrations are expressed in logits. Negative calibrations indicate easier items. Positive calibrations indicate more difficult items. An identity line is drawn through the origin, with a slope of 1. Statistical control lines (95% CI) are drawn around the identity line (dotted lines). The area between the dotted lines depicts acceptable item deviation. Items outside the control lines demonstrate DIF. The numbers near the data points refer to the WOMAC item numbers. For a full explanation of the items see table 1.

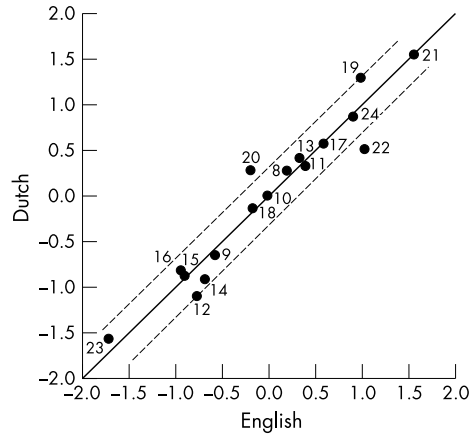

Figure 2 Physical functioning item calibrations for the Dutch WOMAC and the English WOMAC. English calibrations on the x axis, and Dutch on the y axis. Calibrations are expressed in logits. Negative calibrations indicate easier items. Positive calibrations indicate more difficult items. An identity line is drawn through the origin, with a slope of 1. Statistical control lines (95% CI) are drawn around the identity line (dotted lines). The area between the dotted lines depicts acceptable item deviation. Items outside the control lines demonstrate DIF. The numbers near the data points refer to the WOMAC item numbers. For a full explanation of the items see table 1.

groups, and with different severities of OA. Good fit with the Rasch model offers excellent opportunities to study the cross cultural equivalence of translations of the WOMAC by assessing DIF.

The testing of reliability and construct validity showed satisfactory results. The high percentage of item non-response for the AIMS pain scale was unexpected. We believe that this was because the instructions for patients stated that the questions were about ''rheumatoid pain''. Most patients waiting for THA in the Netherlands are well aware of the fact that they have OA, and not rheumatoid arthritis. Responsiveness or sensitivity to change[38] of the Dutch WOMAC was not examined in this paper. This clinimetric aspect of the Dutch WOMAC should be addressed in further research. The responsiveness of a translated measurement instrument, just like dimensionality, reliability, and construct validity, does not prove that valid cross cultural comparisons can be made. Moreover, comparable responsiveness of the original and the translated instrument does not address the validity of cross cultural comparisons. For instance, if the sum scores both preoperatively and postoperatively are higher on the original than on the translated instrument owing to DIF, then the change in score on the original and translated instruments, and the responsiveness of both instruments may be the same, despite a lack of comparability of the scores on the original and translated instruments.

We found DIF for six of the 22 items, which does not automatically indicate inaccurately translated items. There are many other reasons for DIF, apart from a poor translation. Firstly, DIF may be due to chance, and this may be demonstrated when the study is replicated. When a study has been replicated several times, and DIF for a certain item was found in only one of these studies, chance seems to be a reasonable explanation for the DIF. In this study one might expect that DIF of one or two items is due to chance. It is, however, very unlikely that the DIF of six out of 22 items is due to chance. Secondly, large sample sizes will introduce DIF, because small irrelevant differences in item calibrations will become statistically significant differences. In our opinion, about 400 patients, evenly divided between two groups, is a reasonable sample size. Thirdly, DIF can reflect real differences between different cultural and language settings. For instance, the standard seat height for chairs in the Netherlands is higher than the standard height in Canada. Therefore, rising from a chair should be easier in the Netherlands. Fourthly, one may argue that the DIF can also be caused by differences in age, sex, or disease characteristics between the Dutch and Canadian groups. For example, men have more difficulty getting on/off the toilet, due to their longer legs, which makes the seat ''lower'' for them. Therefore, a difference in sex ratio between the two comparison groups will induce DIF. In this study, the differences in age and sex ratios were very small, and not statistically significant, and differences in disease characteristics were minimised by including only patients with OA waiting for primary THA. Lastly, DIF can be caused by different administration modes. In this study, different administration modes were used for the Dutch (self administered) and the Canadian patients (self administered with interviewer assistance). Cella *et al* argue that in studies that compare self administered data with interview administered data little difference is found in the responses given.[5] This still remains a point of concern in our study.

DIF can indicate an incorrect translation of an item, including an incorrect translation of the answer categories. An incorrect translation (for example, the activity in the Dutch translation has become a more comprehensive, thus a more difficult activity, than in the original English translation) will result in DIF. We resubmitted the six malfunctioning items to three native English speakers who had been living in the Netherlands for at least 10 years. They suggested that the translation of items 1, 12, and 19 could be improved. These three items will be rephrased, and will therefore have to be tested in further research.

For an item showing DIF after translation, the translation should be improved, or corrections should be made for the differences in item difficulty. In this study, DIF was corrected in order to make quantitatively sound Dutch-English comparisons. The (standardised) corrected Rasch pain and physical functioning scores of the Dutch and Canadian patients were remarkably comparable. This might indicate that the method of selecting patients for THA in the Netherlands and Canada is quite similar.

Despite satisfactory dimensionality, reliability, and construct validity, our study showed that DIF was found for some items of the Dutch WOMAC. In addition, our study demonstrated that the Dutch WOMAC permits valid Dutch-English comparisons after correction for DIF.

ACKNOWLEDGEMENTS

This study was supported by a grant from the SGO Health Research Promotion Programme.

Authors' affiliations

L D Roorda, G J Lankhorst, Department of Rehabilitation Medicine, VU University Medical Centre, Amsterdam, The Netherlands
L D Roorda, G J Lankhorst, L M Bouter, Institute for Research in Extramural Medicine, VU University Medical Centre, Amsterdam, The Netherlands
C A Jones, Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB, Canada
M Waltz, Rheumatology Department, St Willibrord Hospital, Emmerich, Germany
J W van der Eijken, Department of Orthopaedic Surgery, Onze Lieve Vrouwe Gasthuis, Amsterdam, The Netherlands
W J Willems, Department of Orthopaedic Surgery, Medisch Centrum Alkmaar, Alkmaar, The Netherlands
I C Heyligers, Department of Orthopaedic Surgery, VU University Medical Centre, Amsterdam, The Netherlands
D C Voaklander, K D Kelly, British Columbia Rural and Remote Health Research Institute, College of Arts, Social and Health Sciences, University of Northern British Columbia, Prince George, BC, Canada
M E Suarez-Almazor, Section of Health Services Research, Department of Medicine, Baylor College of Medicine, Houston, TX, USA

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the authors or upon any organisation with which the authors are associated.

REFERENCES

1 **Acquadro C**, Jambon B, Ellis D, Marquis P. Language and translation issues. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*, 2nd ed. Philadelphia: Lippincott-Raven, 1996:575–85.
2 **Anderson RT**, Aaronson NK, Leplege AP, Wilkin D. International use and application of generic health-related quality of life instruments. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*, 2nd ed. Philadelphia: Lippincott-Raven, 1996:613–32.
3 **Bullinger M**, Power MJ, Aaronson NK, Cella DF, Anderson RT. Creating and evaluating cross cultural instruments. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*, 2nd ed. Philadelphia: Lippincott-Raven, 1996:659–68.
4 **Nunnally JC**, Bernstein IH. *Psychometric theory*, 3rd ed. New York: McGraw-Hill, 1994.
5 **Cella DF**, Lloyd SR, Wright BD. Cross cultural instrument equating: current research and future directions. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*, 2nd ed. Philadelphia: Lippincott-Raven, 1996:707–15.
6 **McConnell S**, Kolopack P, Davis AM. The Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC): a review of its utility and measurement properties. *Arthritis Rheum* 2001;**45**:453–61.
7 **Bellamy N**, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, *et al*. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. *J Rheumatol* 1997;**24**:799–802.

8 **Wolfe F**, Kong SX. Rasch analysis of the Western Ontario MacMaster questionnaire (WOMAC) in 2205 patients with osteoarthritis, rheumatoid arthritis, and fibromyalgia. *Ann Rheum Dis* 1999;**58**:563–8.
9 **Bellamy N**. *WOMAC osteoarthritis index: a user's guide*. London (Canada): University of Western Ontario, 1995.
10 **Stucki G**, Meier D, Stucki S, Michel BA, Tyndall AG, Dick W, *et al*. Evaluation einer deutschen version des WOMAC (Western Ontario und McMaster Universities) arthroseindex [German]. *Z Rheumatol* 1996;**55**:40–9.
11 **Wigler I**, Neumann L, Yaron M. Validation study of a Hebrew version of WOMAC in patients with osteoarthritis of the knee. *Clin Rheumatol* 1999;**18**:402–5.
12 **Soderman P**, Malchau H. Validity and reliability of Swedish WOMAC osteoarthritis index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). *Acta Orthop Scand* 2000;**71**:39–46.
13 **Thumboo J**, Chew LH, Soh CH. Validation of the Western Ontario and McMaster University osteoarthritis index in Asians with osteoarthritis in Singapore. *Osteoarthritis Cartilage* 2001;**9**:440–6.
14 **Bae SC**, Lee HS, Yun HR, Kim TH, Yoo DH, Kim SY. Cross cultural adaptation and validation of Korean Western Ontario and McMaster Universities (WOMAC) and Lequesne Osteoarthritis Indices for Clinical Research. *Osteoarthritis Cartilage* 2001;**9**:746–50.
15 **Jones CA**, Voaklander DC, Johnston DW, Suarez-Almazor ME. Health related quality of life outcomes after total hip and knee arthroplasties in a community based population. *J Rheumatol* 2000;**27**:1745–52.
16 **Kelly KD**, Voaklander D, Kramer G, Johnston DW, Redfern L, Suarez-Almazor ME. The impact of health status on waiting time for major joint arthroplasty. *J Arthroplasty* 2000;**15**:877–83.
17 **Kelly KD**, Voaklander DC, Johnston DW, Newman SC, Suarez-Almazor ME. Change in pain and function while waiting for major joint arthroplasty. *J Arthroplasty* 2001;**16**:351–9.
18 **Kelly KD**, Voaklander DC, Johnston WC, Suarez-Almazor ME. Equity in waiting times for major joint arthroplasty. *Can J Surg* 2002;**45**:269–76.
19 **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.
20 **Meenan RF**, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. *Arthritis Rheum* 1992;**35**:1–10.
21 **Bellamy N**. *Musculoskeletal clinical metrology*. Dordrecht: Kluwer Academic, 1993.
22 **Riemsma RP**, Taal E, Rasker JJ, Houtman PM, Van Paassen HC, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. *Br J Rheumatol* 1996;**35**:755–60.
23 **Fries JF**, Spitz PW, Young DY. The dimensions of health outcomes: the health assessment questionnaire, disability and pain scales. *J Rheumatol* 1982;**9**:789–93.
24 **Bijlsma JW**, Oude Heuvel CB, Zaalberg A. Development and validation of the Dutch questionnaire capacities of daily life (VDF) for patients with rheumatoid arthritis. *J Rehabil Sci* 1990;**3**:71–4.
25 **Harris WH**. Traumatic arthritis of the hip after dislocation and acetabular fractures: treatment by mold arthroplasty. An end-result study using a new method of result evaluation. *J Bone Joint Surg Am* 1969;**51**:737–55.
26 **Huisman M**. Imputation of missing item responses: some simple techniques. *Quality and Quantity* 2000;**34**:331–51.
27 **Wright BD**, Stone MH. *Best test design*. Chicago: MESA Press, 1979.
28 **Wright BD**, Masters GN. *Rating scale analysis*. Chicago: MESA Press, 1982.
29 **Andrich D**. *Rasch models for measurement*. Newbury Parks: Sage Publications, 1988.
30 **Linacre JM**, Wright BD. *A user's guide to BIGSTEPS: Rasch-model computer program*. Chicago: MESA Press, 1998.
31 **Wright BD**, Linacre JM. Reasonable mean-square fit values. In: Linacre JM, ed. *Rasch measurement transactions*, Part 2. Chicago: MESA, 1996:370.
32 **Carmines EG**, Zeller RA. *Reliability and validity assessment*. Beverly Hills: Sage Publications, 1979.
33 **Rosner B**. *Fundamentals of biostatistics*, 4th ed. Belmont: Duxbury Press, 1995.
34 **Juniper EF**, Guyatt GH, Jeaschke R. How to develop and validate a new health-related quality of life instrument. In: Spilker B, ed. *Quality of life and pharmacoeconomics in clinical trials*, 2nd ed. Philadelphia: Lippincott-Raven, 1996:49–56.
35 **Ruben H**. Some new results on the distribution of the sample correlation coefficient. *Journal of the Royal Statistical Society, Series B: Methodological* 1966;**28**:513–25.
36 **Ryser L**, Wright BD, Aeschlimann A, Mariacher-Gehler S, Stucki G. A new look at the Western Ontario and McMaster Universities Osteoarthritis Index using Rasch analysis. *Arthritis Care Res* 1999;**12**:331–5.
37 **Rojkovich B**, Gibson T. Day and night pain measurement in rheumatoid arthritis. *Ann Rheum Dis* 1998;**57**:434–6.
38 **Liang MH**. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care* 2000;**38**(suppl 9):II84–90.