

Translation and validation of the Dutch version of the Oxford 12-item knee questionnaire for knee arthroplasty

Daniël Haverkamp, Stefan J M Breugem, Inger N Sierevelt, Leendert Blankevoort and C Nick van Dijk

Academic Medical Center, Department of Orthopedic Surgery, P.O. Box 22660, NL-1100 DD Amsterdam, the Netherlands

Correspondence DH: D.Haverkamp@osteotomie.nl

Submitted 04-07-26. Accepted 04-09-27

Background In 1998, the Oxford 12-item knee questionnaire was developed by Dawson et al. as a self-administered disease- and site-specific questionnaire, specifically developed for knee arthroplasty patients. Since then, it has proven to be an effective outcome questionnaire, and is widely used. Despite the positive psychometric properties for the Total Knee Arthroplasty (TKA) population, the 12-item knee questionnaire has only been translated into a few languages. We thus translated and validated the Oxford 12-item knee questionnaire for the Dutch population.

Methods and results After translation according to a forward/backward protocol, 174 knee arthroplasty patients were asked to answer the questionnaire together with an SF-36, an AKSS and a VAS. The reliability, validity, content validity and the sensitivity to change were all tested. Our Dutch version of the Oxford 12-item knee questionnaire achieved excellent scores in all of these properties.

Interpretation The Dutch Oxford 12-item knee questionnaire proved to be an excellent evaluation instrument for the Dutch orthopedic surgeon and can be used for all total knee arthroplasty patients.

Orthopedic surgeons are becoming increasingly interested in determining the outcome of their surgical interventions, which is reflected in a variety of outcome measures (Dunbar and Gross 1995, Bellamy et al. 1988, Dawson et al. 1998, Dunbar et al. 2000, Liow et al. 2000, Davies 2002).

In 1998, the Oxford 12-item knee questionnaire was developed by Dawson et al. (1998) as a self-administered disease- and site-specific questionnaire, specifically developed for knee arthroplasty patients. Since then, it has proven to be an effective outcome questionnaire and is widely used (Dunbar et al. 2000, Liow et al. 2003, Padua et al. 2003). Despite the positive psychometric properties for the Total Knee Arthroplasty (TKA) population, the 12-item knee questionnaire has only been translated into a few languages and subsequently validated in only two cases: Swedish and Italian (Dunbar et al. 2000, Padua et al. 2003). In both of these languages, the 12-item knee questionnaire has good psychometric properties and has been recommended for the purpose of evaluating TKA results (Dunbar et al. 2000, Padua et al. 2003).

In a large survey of 3 600 knee arthroplasty patients, Dunbar et al. (2000, 2001) compared several site/disease-specific questionnaires and showed that the 12-item knee questionnaire scored the best for burden, feasibility, content validity and reliability. Davies (2002) recently performed a review of all knee-specific questionnaires, concluding that the 12-item knee questionnaire, the 36-item Short-Form health survey (SF-36) and the WOMAC were the most appropriate for the assessment of outcome after total knee arthroplasty. Liow et al. (2003) recently compared studies of different functional rating systems for TKA and the 12-item knee questionnaire emerged as the most reliable, since it eliminates interobserver errors.

The only knee-specific questionnaire that has validated in Dutch is the WOMAC (Roorda et al. 2004). The WOMAC is a commonly used disease- and site-specific questionnaire, but it was not developed specifically to evaluate functional outcome of TKA (Bellamy et al. 1988). The 12-item knee questionnaire has been shown to be reliable and the most appropriate for the assessment of outcome after total knee arthroplasty (Dunbar et al. 2000, 2001, Davies 2002, Liow et al. 2003). A validated Dutch version of the 12-item knee questionnaire would therefore be of value. The purpose of this study was to translate the 12-item knee questionnaire into the Dutch language and to validate it.

Patients and methods

Translation procedure

A Dutch translation was made using a forward/backward translation protocol according to the guidelines of cross-cultural adaptation from Guillemin et al. (1993). Since no major cultural differences in lifestyle exist between the Dutch and English population, we assumed that cultural adaptation of the questionnaire was not required.

Patients

The study population consisted of 174 patients (mean age 69 (31–92) years), 65% of whom were female. A preoperative 12-item knee questionnaire in Dutch was obtained from 119 patients who were scheduled for a primary TKA for osteoarthritis of the knee joint. From 111 patients, a 12-item knee questionnaire was obtained 1 year after TKA. Within these patient groups, 56 patients took part in both preoperative and postoperative sampling. All patients were examined clinically and radiographs were taken both preoperatively and at follow-up.

Questionnaires

Besides the 12-item knee questionnaire, the outcome measures that we used were the SF36, the American Knee Society (AKS) score, and a 100-mm Visual Analog Scale (VAS) for pain.

The 12-item knee questionnaire is a self-administered disease- and site-specific questionnaire and consists of 12 questions to be answered by the patient concerning the knee, and is therefore

entirely subjective. Each question consists of a 5-point Likert scale, leading to a total score ranging from a best functional score of 12 to the worst functional outcome of 60 (Dawson et al. 1998). By incorporating questions about pain and disability, these items have more weight in the score. For this analysis, we inverted (made reciprocal) the values of the 12-item knee questionnaire in order to simplify the interpretation of the results, since all the other scores have values that are reciprocal to those of the 12-item questionnaire.

The SF-36 is a general health questionnaire and consists of 36 Likert box questions (Brazier et al. 1992, Ware and Sherbourne 1992, Sullivan et al. 1995, Aaronson et al. 1998). The questionnaire contains 8 health concepts: physical functioning (PF), role limitation due to physical problems (RP), bodily pain (BP), perception of general health (GH), energy and vitality (VT), social functioning (SF), role limitation due to emotional problems (RE), and mental health (MH). The results of the SF-36 range from 0 (the worst outcome) to 100 (the best outcome) (Brazier et al. 1992, Sullivan et al. 1995).

The American Knee Society (AKS) score is a clinical rating system and is divided into a knee score and a function score. The knee score is designed to assess pain, stability and range of motion. This score can be influenced by subtracting points for flexion contractures, extension lag and malalignment. The function score is designed to assess walking distance and the ability to climb stairs. Again, points are subtracted for the use of walking aids. Both scores range from 0 to 100, with a high score signifying a good result (Insall et al. 1989).

The VAS is a 10-point visual analog scale (100 mm) and is used to determine the seriousness of the general pain in the affected knee over the preceding month (Price et al. 1983).

Testing

Psychometrics can be defined as “the scientific measurement of mental capacities and processes of personality” (Dunbar 2001). Psychometrics is therefore the “process that allows researchers to apply scientific methodology to the measurement of subjective outcomes” (Dunbar 2001). The psychometric properties of a questionnaire define how well a questionnaire measures what it is supposed

to measure. The aspects that were tested were reliability, validity and responsiveness. The patient burden imposed by administering the 12-item knee questionnaire was considered to be identical to that of the original version of Dawson et al. (1998), and these authors found it to be minimal.

Reliability

Reliability is defined as the ability of a test to “yield the same results on repeated trials under the same conditions” (Liow et al. 2003). To determine the test-retest reliability, a second questionnaire was given to a random sample of 26 patients who were scheduled for total knee arthroplasty. They were asked to complete the questionnaires with 1 week interval and return them immediately after completion. Test-retest reliability was assessed by the intra-class coefficient (ICC) (Bland and Altman 1996), the Pearson correlation coefficient and the coefficient of reliability of Bland and Altman (Bland and Altman 1986). T-tests were performed to determine the systematic difference between the first and second test.

Internal consistency was assessed to determine whether all 12 items cover the same construct. It was calculated from the data from the preoperative ($n = 119$) and postoperative ($n = 111$) populations separately, and data from the total population ($n = 174$). Internal consistency was assessed by Cronbach’s (1955) alpha. Cronbach’s alpha addresses the homogeneity of the questions included in a questionnaire and is complimentary to the ICC as a measure of reliability. An alpha of 0.7 is considered to represent a fair degree of internal consistency, 0.8 is considered to represent good internal consistency, and 0.9 would represent excellent internal consistency (Dunbar et al. 2000).

Validity

Validity relates to the ability of a questionnaire to measure the outcome parameter of interest. Criterion validity refers to comparison of the new test with the “gold standard”. Unfortunately, there is no “gold” standard for knee arthroplasty. Consequently, questionnaires for knee arthroplasty are usually validated against a postulated effect that should result from the intervention. Such a postulation is referred to as a construct (Dunbar et al. 2001). Construct validity was tested by comparing

the 12-item knee questionnaire to a 100-mm VAS scale for pain, the AKS knee and function score and the relevant domains of the SF-36 scores. The scores of the 8 domains of the SF-36 were used to assess convergent and divergent validity of the 12-item knee questionnaire. Construct validity was evaluated using Pearson correlation coefficients between the 12-item knee questionnaire and the 100-mm VAS, the AKS knee and function score and the SF-36.

We evaluated convergent and divergent validity by hypothesizing that correlation coefficients between the study questionnaire and SF-36 domains bodily pain (BP), role of physical limitations (RP) and physical functioning (PF), were higher than correlations with the other domains.

Content validity addresses whether a questionnaire has enough items and adequately covers the domain of interest. Content validity was evaluated by assessing the distribution and floor and ceiling effects of the 12-item knee questionnaire. The floor effect occurs when the patient scores the lowest possible score (12), and therefore the patient appears to be very satisfied. The ceiling effect is the highest possible score, and is thus the opposite of the floor effect (Dunbar et al. 2000). Construct validity and content validity were tested on the scores of the preoperative and postoperative populations separately.

Responsiveness

Responsiveness is a measure of the ability of a questionnaire to determine sensitivity to change, before and after an intervention. A sample of 56 patients with preoperative and postoperative scores was used to determine sensitivity to change. Effect sizes of all the questionnaires were calculated and compared. Effect size is defined as the ratio of the mean change in pre- and postoperative scores; this value is then divided by the standard deviation of the preoperative score. It represents a standardized measure of change in a group, and can be used to compare different clinical measures. Effect sizes of 0.2, 0.5 and 0.8 are considered to be small, moderate and large, respectively (Dawson et al. 1998, Dunbar et al. 2000).

Statistics

Statistical analysis was done using SPSS (version

Table 1. Pearson correlation coefficients of the Oxford 12-item knee score, and VAS, AKS and SF36 scores

	Preoperatively	1-year follow-up
Pain VAS	0.62	0.81
AKSS knee score	0.33	0.62
AKSS function	0.31	0.48
SF-36		
Physical functioning (PF)	0.64	0.64
Role limitation due to physical problems (RP)	0.40	0.63
Role limitation due to emotional problems (RE)	0.26	0.50
Social functioning (SF)	0.46	0.57
Mental health (MH)	0.27	0.40
Energy and vitality (VT)	0.14	0.48
Bodily pain (BP)	0.68	0.67
General health perception (HP)	0.31	0.39
All differences between preoperative and follow-up values are significant with $p < 0.01$.		

12.01; SPSS Inc., IL), and a p -value of less than 0.05 was considered to be statistically significant.

Results

Reliability

The ICC of the questionnaire was high at 0.97 ($p < 0.01$) (95% CI: 0.94–0.99), and we found a mean difference of 2.6 between the 2 assessments ($p < 0.01$). The reliability coefficient of Bland and Altman was 4.6 (95% CI: –7.2 to 2.0). Cronbach's alpha showed that the questionnaire had strong internal consistency, with values of 0.87 and 0.90 preoperatively and at 1-year follow-up, respectively. The value of Cronbach alpha for the total population was 0.94.

Validity

The 12-item knee questionnaire correlated very well with the VAS before the TKA operation and at 1-year follow-up. However, the correlation between the 12-item knee questionnaire and the 2 AKS scores was low before the operation and moderate 1 year after TKA. The highest correlation coefficients with the domains of the SF-36 were seen for the PF, RP and BP. However, the correlation coefficient of the RP before the operation was moderate (Table 1).

Content validity

The scores of the questionnaire before operation and at 1-year follow up were normally distributed. Ceiling and floor effects were not observed preoperatively. At 1-year follow-up, a slight floor effect was observed: 10 patients (9%) scored the lowest possible score. A ceiling effect was not seen postoperatively.

Sensitivity to change

A statistically significant improvement of –16 (95% CI –18; –13) points for the 12-item knee questionnaire was observed 1 year after total knee arthroplasty ($p < 0.01$). The scores of the other outcome measures improved significantly too ($p < 0.01$), except Mental Health in the SF-36 ($p = 0.5$). The effect size of the 12-item knee questionnaire 1 year postoperatively was –2.03, which was larger than the effect sizes of the other outcome measures (Table 2).

Discussion

The aim of this study was to translate the original disease- and site-specific 12-item knee questionnaire on the subjective input of patients with total knee arthroplasty, into the Dutch language and to validate it. A mere translation of the questionnaire into another language—without proper validation of the translated version—would not suffice (Guillemin et al. 1993). The translation procedure did not create any problems, since the items in this questionnaire are universal for all knee arthroplasty patients and there is little or no cultural difference between Dutch and British patients.

According to the ICC and correlation coefficient, the test-retest reliability was good. However, a statistically significant difference of 2.85 was observed between the two assessments. Although this difference is small and not relevant clinically, it can be explained by the fact that it was not specifically asked whether the complaints of the patients had changed at the second assessment. After checking the patient files, it appeared that 29 patients had been treated with physiotherapy or analgesics. Taking this into account, the results seem to indicate good test-retest reliability.

Table 2. Effect size of the Oxford 12-item knee score, and VAS, AKS and SF36 scores

Outcome measure	Preoperative mean (SD)	1 year follow-up mean (SD)	Difference	95% CI	Effect size
Oxford 12-item (n = 56)	34 (7.5)	49 (9.0)	–15	–18; –13	–2.0
VAS	47 (27)	84 (25)	–37	–45; –29	–1.4
AKSS knee score	59 (17)	87 (15)	–28	–36; –20	–1.6
AKSS function score	54 (20)	76 (22)	–22	–30; –15	–1.1
SF-36					
Physical functioning (PF)	33 (19)	58 (26)	–25	–32; –18	–1.3
Role limitation due to physical problems (RP)	17 (28)	57 (42)	–41	–53; –28	–1.5
Role limitation due to emotional problems (RE)	57 (43)	74 (42)	–17	–28; –5.4	–0.38
Social functioning (SF)	59 (22)	73 (21)	–13	–20; –6.1	–0.6
Mental health (MH)	78 (19)	80 (20)	–1.9	–6.8; 3.0	–0.1
Energy and vitality (VT)	61 (22)	70 (23)	–9.2	–14; –4.5	–0.4
Bodily pain (BP)	39 (20)	71 (29)	–32	–40; –24	–1.6
General health (GH)	72 (19)	65 (23)	7.5	3.1; 12	0.39

The study questionnaire has a very high internal consistency for the total population (Cronbach's alpha 0.94). This could indicate redundancy, but since the Cronbach's alpha did not exceed 0.9 when assessed for the preoperative and postoperative populations separately, no questions needed to be excluded from the list. Dunbar et al. (2000) and Dawson et al. (1998) obtained comparable results.

Pearson correlation coefficients between the study questionnaire and VAS were good, low preoperatively and moderate postoperatively with AKSS. These moderate preoperative correlations are comparable to the results of Dawson et al. (1998), who did not mention the postoperative values. However, convergent and divergent validity is confirmed, since the study questionnaire correlated better with the domains BP, PF and RP than with the other domains of the SF-36. The low correlation coefficient with the PF domain preoperatively can be explained by a considerable floor effect of the PF; 73 patients (64%) had the minimum score. Considering these results, we conclude that the questionnaire has good construct validity. The 12-item knee questionnaire is highly sensitive, since the effect size of the study questionnaire 1 year postoperatively was –2.0, which was higher than for any of the individual parts of the SF-36 (Table 2).

Aaronson N K, Muller M, Cohen P D, Essink-Bot M L, Fekkes M, Sanderman R, Sprangers M A, te V A, Verrips E. Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *J Clin Epidemiol* 1998; 51: 1055–68.

Bellamy N, Buchanan W W, Goldsmith C H, Campbell J, Stitt L W. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol* 1988; 15: 1833–40.

Bland J M, Altman D G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307–10.

Bland J M, Altman D G. Measurement error and correlation coefficients. *BMJ* 1996; 313: 41–2.

Brazier J E, Harper R, Jones N M, O'Cathain A, Thomas K J, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *BMJ* 1992; 305: 160–4.

Davies A P. Rating systems for total knee replacement. *Knee* 2002; 9: 261–6.

Dawson J, Fitzpatrick R, Murray D, Carr A. Questionnaire on the perceptions of patients about total knee replacement. *J Bone Joint Surg (Br)* 1998; 80: 63–9.

Dunbar M J. Subjective outcomes after knee arthroplasty. *Acta Orthop Scand (Suppl 301)* 2001; 72: 1–63.

Dunbar M J, Gross M. Critical steps in total knee arthroplasty. A method of analyzing operative procedures. *Int Orthop* 1995; 19: 265–8.

Dunbar M J, Robertsson O, Ryd L, Lidgren L. Translation and validation of the Oxford-12 item knee score for use in Sweden. *Acta Orthop Scand* 2000; 71: 268–74.

- Dunbar M J, Robertsson O, Ryd L, Lidgren L. Appropriate questionnaires for knee arthroplasty. Results of a survey of 3600 patients from The Swedish Knee Arthroplasty Registry. *J Bone Joint Surg (Br)* 2001; 83: 339-44.
- Guillemin F, 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.
- Insall J N, Dorr L D, Scott R D, Scott W N. Rationale of the Knee Society clinical rating system. *Clin Orthop* 1989; (248): 13-4.
- Liow R Y, Walker K, Wajid M A, Bedi G, Lennox C M. The reliability of the American Knee Society Score. *Acta Orthop Scand* 2000; 71: 603-8.
- Liow R Y, Walker K, Wajid M A, Bedi G, Lennox C M. Functional rating for knee arthroplasty: comparison of three scoring systems. *Orthopedics* 2003; 26: 143-9.
- Padua R, Zanoli G, Ceccarelli E, Romanini E, Bondi R, Campi A. The Italian version of the Oxford 12-item Knee Questionnaire-cross-cultural adaptation and validation. *Int Orthop* 2003; 27: 214-6.
- Price D D, McGrath P A, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain* 1983; 17: 45-56.
- Roorda L D, Jones C A, Waltz M, Lankhorst G J, Bouter L M, van der Eijken J W, Willems W J, Heyligers I C, Voaklander D C, Kelly K D, Suarez-Almazor M E. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. *Ann Rheum Dis* 2004; 63: 36-42.
- Sullivan M, Karlsson J, Ware J E, Jr. The Swedish SF-36 Health Survey. I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. *Soc Sci Med* 1995; 41: 1349-58.
- Ware J E, Jr., Sherbourne C D. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; 30: 473-83.