Reproducibility and validity of the DynaPort KneeTest

Mokkink, L.B.; Terwee, C.B.; van der Slikke, R.M.; van Lummel, R.C.; Benink, R.J.; Bouter, L.M.; de Vet, H.C.W.

published in
Arthritis & Rheumatism
2005

DOI (link to publisher)
10.1002/art.21167

document version
Publisher's PDF, also known as Version of record

Arthritis & Rheumatism (Arthritis Care & Research)
Vol. 53, No. 3, June 15, 2005, pp 357–363
DOI 10.1002/art.21167

ORIGINAL ARTICLE

Reproducibility and Validity of the DynaPort KneeTest

LIDWINE B. MOKKINK,[1] CAROLINE B. TERWEE,[1] RIENK M. A. VAN DER SLIKKE,[2] ROB C. VAN LUMMEL,[2] ROB J. BENINK,[3] LEX M. BOUTER,[1] AND HENRICA C. W. DE VET[1]

Objective. **To determine the reproducibility and validity of the DynaPort KneeTest, a performance-based test that measures quality of movement of patients undergoing total knee replacement (TKR).**

Methods. **A total of 92 patients with osteoarthritis (OA) of the knee performed the KneeTest twice on the same day; 94 healthy controls performed the KneeTest once. During the test, 29 activities were performed with accelerometers attached to the body. Relevant functional parameters were extracted from the accelerometers. A selection of parameters was used to calculate activity scores, based on the ability of parameters to discriminate between patients and controls (regression analyses). Based on internal consistency analyses (Cronbach's α), redundant activities were removed. Four cluster scores and a total KneeScore were calculated from the remaining activity scores. Reproducibility and validity of the cluster scores and the total KneeScore 2 were assessed.**

Results. **Based on internal consistency analyses, the test was reduced to 23 activities. Inter- and intraobserver reliability using intraclass correlation coefficients were 0.90 (0.83–0.94) and 0.95 (0.83–0.98), respectively. Limits of inter- and intraobserver agreement were –8.3 to 11.3 and –4.2 to 9.0. Construct validity was confirmed by expected correlations with the Western Ontario and McMaster University Osteoarthritis Index physical functioning (0.55), Medical Outcomes Study Short Form-36 Health Survey physical functioning (0.62), and Knee Society Score function (0.64).**

Conclusion. **The KneeTest is a useful performance-based measure for research in patients with knee OA undergoing TKR, with good reliability and validity. Further research is required to improve its usefulness for clinical practice.**

KEY WORDS. Total knee replacement; DynaPort KneeTest; Osteoarthritis.

INTRODUCTION

Functioning is one of the key outcome measures for patients with osteoarthritis (OA) of the knee undergoing total knee replacement (TKR) (1). Functioning can be assessed either by performance-based measures (supervised by a health care professional), or by self-report questionnaires (completed by the patient). Both methods provide information about different aspects of functioning. Performance-based measures intend to measure what a patient can do, while self-report questionnaires measure what a person thinks he or she can do.

A number of arguments have been proposed in favor of both methods. For example, authors have argued that self-report questionnaires (2–5) are easier to use, less time consuming, less of a burden to patients, and cannot be influenced by observer bias (1,6). Performance-based measures are reported to be more objective and less influenced by patient expectations, culture, education, or cognitive impairments (2–5), but they measure functioning in an artificial situation (7,8). We recently showed that self-reported functioning, especially the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) (9), is highly correlated with pain, whereas performance-based functioning is not (Terwee CB: unpublished observation). Because of these differences, it has been suggested that both performance-based measures and self-report questionnaires are required to comprehensively evaluate functioning in TKR (10,11).

Self-report questionnaires such as the WOMAC or the Medical Outcomes Study Short Form-36 Health Survey

Supported by McRoberts, The Hague, The Netherlands, and the Royal Dutch Association for Physical Therapy.

[1]Lidwine B. Mokkink, MSc, Caroline B. Terwee, PhD, Lex M. Bouter, PhD, Henrica C. W. de Vet, PhD: Institute for Research in Extramural Medicine, VU University Medical Center, Amsterdam, The Netherlands; [2]Rienk M. A. van der Slikke, MSc, Rob C. van Lummel, MSc: McRoberts, The Hague, The Netherlands; [3]Rob J. Benink, MD: Gemini Hospital, Den Helder, The Netherlands.

Address correspondence to Caroline B. Terwee, PhD, Institute for Research in Extramural Medicine (EMGO Institute), VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands. E-mail: cb.terwee@vumc.nl.

Submitted for publication August 11, 2004; accepted in revised form January 3, 2005.

(SF-36) (12) are used extensively; however, performance-based measures are used less often. This might be due to a lack of valid, easy-to-use, performance-based measures. In a recent systematic review aimed at finding all performance-based tests that have been evaluated for their measurement properties in patients with hip or knee OA or patients undergoing hip or knee replacement, we found 24 different performance-based tests (Terwee CB: unpublished observation). None of these measures had been tested for reproducibility or validity and responsiveness. Also, the measurement properties were often assessed inadequately, using too small a sample size or with an inadequate analysis method.

The DynaPort KneeTest (McRoberts, The Hague, The Netherlands) is a promising performance-based measure because of its apparent content validity for patients with knee OA undergoing TKR (13). Patients perform a standardized set of 29 activities in 15–30 minutes, while accelerometers are used to measure functional parameters. The KneeTest is more economical and easier to use than gait analysis. In contrast to gait analysis or more simple performance-based tests (e.g., 6-minute walk test, stair climb tests) (14–16), the KneeTest contains multiple activities (e.g., walking, stair climbing, sitting and rising, lifting and carrying objects, and picking up objects from the floor) that were seleced to represent activities of daily living that are considered difficult for patients with knee problems and that patients in focus-group discussions have identified as being important (2). Unlike other performance-based tests (except gait analysis), the KneeTest also measures different aspects of functioning than just the time it takes to perform the test, e.g., movement intensity, accelerations, and joint angles during movement. Time alone has been suggested to be inadequate to represent the concept of functional status (17). Finally, the scoring of the KneeTest is based on identification of the most significant test parameters to separate patients and controls, an approach that has been shown to be useful for the evaluation of functioning in patients undergoing TKR (18). Therefore, the KneeTest appears to have good content validity. The aim of this study was to determine the reproducibility and construct validity of the DynaPort KneeTest.

MATERIALS AND METHODS

Design. A cross-sectional study was performed in 2 hospitals, Gemini Hospital, Den Helder and Atrium Medical Centre, Brunssum, the Netherlands. Four test days in each hospital were organized. Preoperative and postoperative patients were asked to perform the DynaPort KneeTest (McRoberts) twice, and healthy controls were asked to perform the KneeTest once. Additional information was collected through a questionnaire and physical examination. The Medical Ethics Committee of the VU University Medical Center, Amsterdam approved the study protocol.

Research population. Inclusion criteria for patients were as follows: diagnosis of knee OA from an orthopedic surgeon; registered on the waiting list for a primary total or unicondylar TKR, or had received a TKR 3 months to 5 years ago; able to speak and read Dutch; and have given written informed consent. Inclusion criteria for controls included no indication for TKR (i.e., no OA, rheumatoid arthritis, or severe trauma of the knee), no contraindication for (knee) surgery, between the ages of 45 and 90 years (the common age range for TKR), able to speak and read Dutch, and had given written informed consent.

A total of 96 patients (24 preoperative and 24 postoperative patients in each hospital) were recruited. In addition, 96 age- and sex-matched controls were recruited via patients, employees of the hospital, and acquaintances of the authors in the region of the 2 hospitals.

Procedures. Both patients and controls received standardized KneeTest instructions on video. Patients were randomized to either perform the KneeTest twice under supervision of the same physical therapists (intraobserver reliability) or under supervision of 2 different physical therapists (interobserver reliability). In each hospital, 3 different physical therapists supervised the KneeTest.

After each test the physical therapist was asked to give an overall rating of the patients' quality of movement on a 5-point scale, based on their own interpretation of the quality of the patient performance. Between the 2 tests, patients completed a questionnaire, which consisted of the SF-36 (12), the WOMAC (9), and an overall question about quality of movement on a 5-point scale. A fourth physical therapist per hospital administered the Knee Society Score (KSS) (19,20). Controls performed the KneeTest once, and their age and sex were recorded.

KneeTest. Patients performed the 29 activities of the KneeTest (see Appendix A available at the *Arthritis Care & Research* Web site http://www.interscience.wiley.com/jpages/0004-3591:1/suppmat/index.html) while wearing 6 acceleration sensors strapped around the trunk and the legs. The activities were categorized into 4 clusters: locomotion (walking), rising and descending (stair climbing or stepping up blocks), lifting and moving objects (carrying a bag or walking with a shopping trolley), and transfers (sitting down or picking up a weight). Standardized equipment was supplied such as blocks, stairs, and a slope.

Thirty functional parameters per activity (i.e., accelerations, angles, durations, step number, step frequencies, relative speed, and asymmetry) were extracted from the signals of the acceleration sensors (see Appendix B available at the *Arthritis Care & Research* Web site http://www.interscience.wiley.com/jpages/0004-3591:1/suppmat/index.html). The values of these parameters were transformed into a score for each activity and a total functional score for the entire test (KneeScore), which is considered to be a measure of quality of movement. The extraction of the parameters and calculation of the KneeScore was performed at McRoberts Company using an automated procedure. The KneeTest has been previously described (13).

Scoring system. The scoring algorithm was modified from the method described by van den Dikkenberg et al (13). Scoring of the functional parameters was based on discrimination between patients and controls. We used

multiple logistic regression analyses, adjusted for age and sex, to select those functional parameters per activity that could significantly separate patients and controls.

In advance, a preselection of parameters was made, based on 3 considerations: 1) gait parameters were excluded from the activities "stepping up blocks" and "sit down and stand up" (activities 9–14 and 17,18); 2) if 2 parameters had a high correlation ($r > 0.90$), the least conceptually-relevant parameter was excluded; and 3) only parameters with $P < 0.25$ in univariate linear regression analyses were included in the multivariate model (see Appendix C available at the *Arthritis Care & Research* Web site http://www.interscience.wiley.com/jpages/0004-3591:1/suppmat/index.html for a list of functional paramaters).

For the selected parameters, we used norm-based scoring to calculate parameter scores, by subtracting mean parameter values for the control group from each patient's value and dividing the difference by the SD of the control group. The parameter scores were averaged into a total activity score for each activity. If a parameter score was missing (e.g., due to a detached sensor), the activity score was calculated as the average of the available parameter scores. The activity scores were multiplied by 20 and added up by 50 to achieve positive scores. An activity score of 50 is therefore equal to the mean score of the controls, and an activity score of 30 implies that a person scores one SD below (i.e., worse than) the mean of the controls. If a person did not perform an activity because of physical inability, an activity score of 3 SD below the mean of the patients was assigned, indicating that the patient had a bad score for that activity in comparison with the group of patients.

The activity scores were averaged into 4 cluster scores and a total KneeScore was calculated as the average of the 4 cluster scores. In the original KneeScore method described by van den Dikkenberg et al (13), the assignment of activities to clusters was based on conceptual considerations only. In our modified KneeScore (KneeScore 2), the activities were assigned slightly differently, based on internal consistency analyses (see below).

Assessment of measurement properties. *Internal consistency.* The Cronbach's alpha and item-total correlations were calculated to determine the internal consistency of the activities within the clusters. In case of a very high Cronbach's alpha (>0.95), indicating that some activities were redundant (21), the Spearman-Brown formula (22) was used to assess the amount of activities that would be sufficient for adequate internal consistency. In case of a low Cronbach's alpha (<0.70) or low item-total correlations (<0.40), the Spearman-Brown formula was used to remove activities with a low item-total correlation from the cluster. Cronbach's alpha between 0.70–0.95 was considered to be adequate.

Reproducibility. Both reliability and agreement were assessed (23) for the activity scores, the cluster scores, and KneeScore 2. Intraclass correlation coefficients (ICCs) were calculated to assess interobserver reliability and intraobserver reliability (21,24). An ICC >0.70 was considered to be satisfactory (25).

To assess intraobserver and interobserver agreement, mean differences between the 2 measurements were calculated with limits of agreement (defined as mean difference $\pm$ 1.96 $\times$ SD of the difference) according to the method described by Bland and Altman (26). The limits of agreement define the range within which 95% of the differences between the 2 measurements lie. In the internal consistency and reproducibility analyses, no values were imputed for activities that were not performed by patients, because imputation would artificially increase the internal consistency and reproducibility.

Construct validity. To assess construct validity, KneeScore 2 of the patients was compared with several other measures. Spearman correlations were calculated between KneeScore 2 and the KSS (scores for functioning and pain), the WOMAC (subscores for physical functioning and pain), the SF-36 (subscores for physical functioning and pain), the physical therapists' opinion of the patients' quality of movement and the patient's own opinion of his or her quality of movement, the mean duration of the activities of the test, and the active range of motion of the (to be) operated leg.

Specific hypotheses about the expected relationships between KneeScore 2 and the other measures were specified in advance. In addition, it was expected that postoperative patients would have higher cluster scores and a higher KneeScore 2 than preoperative patients. A moderate to large (>0.50) effect size for TKR was expected (defined as the difference in mean scores between preoperative and postoperative patients, divided by the SD of the mean score of the preoperative patients). Construct validity was considered satisfactory if at least 75% of the hypotheses were confirmed (27).

RESULTS

Participants. During 8 test days, 92 patients (41 preoperative patients and 51 postoperative patients) performed the KneeTest twice, and 101 controls performed the test once. The test of one patient, the retest of another patient, and the test of 4 healthy controls failed because of technical problems. Three controls turned out to be too young and were excluded from all analyses, leaving a total of 94 controls in the analyses. All patients completed the questionnaire. The mean $\pm$ SD age of the patients was 67 $\pm$ 9.9 years and 66 $\pm$ 7.4 years for the controls. The number of men versus women was 25/67 in the patient group and 38/56 in the control group, respectively.

Selection of functional parameters. The number of functional parameters that were considered in the multivariate analyses varied from 9 to 26 per activity. The definite selection of parameters varied from 3 to 10 per activity (see Appendix C).

Internal consistency. The original 8 activities of the locomotion cluster had a high Cronbach's alpha (0.98) with high item-total correlations (>0.85). It was estimated (based on the Spearman-Brown formula) (22) that 4 activ-

Table 1. Reliability of cluster scores and KneeScore 2 in 90 patients*

	Interobserver reliability		Intraobserver reliability	
	ICC (95% CI)	n	ICC (95% CI)	n
Locomotion	0.88 (0.80–0.93)	57	0.94 (0.86–0.97)	33
Rise and descend	0.85 (0.75–0.91)	56	0.87 (0.32–0.96)	32
Transfers	0.66 (0.48–0.79)	56	0.89 (0.78–0.94)	32
Lift and move objects	0.84 (0.75–0.90)	56	0.93 (0.85–0.96)	33
KneeScore 2	0.90 (0.83–0.94)	57	0.95 (0.83–0.98)	33

* ICC = intraclass correlation coefficient; 95% CI = 95% confidence interval.

ities would be sufficient to get a scale with a good internal consistency. We chose to retain the activities "walk 9 meters" (twice), "walk 9 meters and back," and "walk a longer distance" on the basis of conceptual and practical considerations. The new locomotion cluster with these 4 activities had a Cronbach's alpha of 0.95.

The original 10 activities of the cluster "rise and descend" had a Cronbach's alpha of 0.91, with item-total correlations >0.50, and were therefore retained.

The original 6 activities of the transfer cluster had a Cronbach's alpha of 0.51. We removed the activities "sit down and stand up from 20 cm" and "lie down and stand up," because of low item-total correlations and the high number of patients and controls that could not perform these activities. This increased the Cronbach's alpha to 0.73. The activity "sit down and stand up from 30 cm" has an item-total correlation of 0.31, but was retained on a conceptual basis.

Originally, 2 activities ("picking up a weight with non-affected and affected leg") were included in the transfer cluster as well as in the lift and move objects cluster. We removed these activities from the lift and move objects cluster. The new lift and move objects cluster (now containing 5 activities) had a Cronbach's alpha of 0.93. In total, after removing 6 redundant activities, 23 activities were included in KneeScore 2.

Reliability. The ICCs for the cluster scores and KneeScore 2 were >0.80, except for the interobserver reliability of the transfer cluster (0.66, Table 1). Most ICCs (32 of 46, 69%) for the inter- and intraobserver reliability of the individual activity scores were >0.70 (data not shown). The interobserver reliability of 9 activities was <0.70 (range 0.32–0.66), and the intraobserver reliability of 5 activities (4 activities of the rise and descend cluster and one activity of the transfer cluster) was <0.70 (range 0.58–0.69).

Agreement. Mean ± SD interobserver difference in KneeScore 2 was 1.5 ± 5.0 points (n = 57). Limits of interobserver agreement ranged from −8.3 to 11.3. Mean ± SD intraobserver difference in KneeScore 2 was 2.4 ± 3.4 points (n = 33). Limits of intraobserver agreement ranged from −4.2 to 9.0 (Table 2).

Construct validity. Correlations between KneeScore 2 and other measures were generally as anticipated (Table 3). We found high correlations with other measures of functioning and low correlations with measures of pain, although the correlation with the Knee Society Pain Score and with the mean duration of the activities was higher than expected. The correlation with the patient's rating of his or her quality of movement was lower than expected. As expected, postoperative patients had higher cluster scores and KneeScores than preoperative patients, and effect sizes were >0.50 (Table 4).

In an attempt to explain the high correlation between

Table 2. Limits of interobserver (n = 57) and intraobserver (n = 33) agreement for the KneeScore 2 in 90 patients*

	Measurement 1	Measurement 2	Difference	Limits of agreement, range
Interobserver agreement				
Locomotion	34.7 ± 14.3	35.5 ± 15.1	0.8 ± 7.3	−13.5, 15.1
Rise and descend	32.2 ± 11.0	34.3 ± 11.7	1.7 ± 6.0	−10.1, 13.4
Transfers	35.7 ± 10.4	38.6 ± 11.1	2.5 ± 8.6	−14.3, 19.4
Lift and move objects	34.7 ± 13.0	36.2 ± 13.6	1.1 ± 7.5	−13.5, 15.8
KneeScore 2	43.3 ± 11.1	35.9 ± 11.8	1.5 ± 5.0	−8.3, 11.3
Intraobserver agreement				
Locomotion	32.0 ± 16.2	34.9 ± 17.1	2.4 ± 5.6	−8.6, 13.3
Rise and descend	30.3 ± 12.2	35.1 ± 10.4	4.4 ± 4.2	−3.8, 12.5
Transfers	34.4 ± 11.5	35.5 ± 10.6	1.0 ± 5.3	−9.3, 11.3
Lift and move objects	33.2 ± 13.5	35.2 ± 12.3	1.6 ± 4.7	−7.7, 10.8
KneeScore 2	31.8 ± 13.6	34.6 ± 12.6	2.4 ± 3.4	−4.2, 9.0

* Unless otherwise indicated, values are the mean ± SD.

Table 3. Construct validity of KneeScore 2 in 92 patients, based on prior hypotheses about the expected correlations*

	Prior hypotheses		Results	Hypothesis confirmed
	r_s direction	r_s magnitude†	r_s	
Range of motion	Positive	Moderate	0.06	Yes
KSS functioning	Positive	High	0.64	Yes
WOMAC physical functioning	Positive	High	0.55	Yes
SF-36 physcial functioning	Positive	High	0.62	Yes
QOM physical therapists	Positive	High	0.68	Yes
QOM patient	Positive	Moderate	0.32	No
Mean duration of activities	Positive	High	0.93	Yes
KSS pain	Negative	Low	−0.47	No
SF-36 pain	Negative	Low	−0.32	Yes
WOMAC pain	Negative	Low	−0.35	Yes

* KSS = Knee Society Score (ref. 19, 20); WOMAC = Western Ontario and McMaster University Osteoarthritis Index; SF-36 = Medical Outcomes Study Short Form-36; QOM = quality of movement (scored on a 5-point scale by the physical therapist after the first test and by the patient in the questionnaire); range of motion = active range of the (to be) operated leg.
† Spearman correlation magnitude high = $r > 0.50$, moderate = r 0.35–0.50, low = $r < 0.35$.

KneeScore 2 and the mean duration of the activities, we subsequently calculated the correlations between the cluster scores and the mean duration of the activities in the cluster, the correlations between the activity scores and the duration of each individual activity, and the correlations between the functional parameter scores of each activity and the duration of the activity. Correlations between the cluster scores and the mean duration of the activities in the cluster ranged from 0.84 to 0.91. Correlations between the activity scores and the duration of the activities ranged from 0.39 to 0.91 (mean 0.77). Six percent of the functional parameter scores correlated >0.90 with the duration of the activity in which the parameter was assessed.

DISCUSSION

To our knowledge, the DynaPort KneeTest is one of the best validated, performance-based measures for the assessment of functioning of patients with knee OA, undergoing TKR. We shortened the KneeTest from 29 to 23 activities and the internal consistency of 3 out of 4 clusters of activities was considered to be good. KneeScore 2 was found to be highly reliable, but the limits of agreement were rather wide. Construct validity of the KneeTest was supported by expected correlations with other measures of functioning and pain. Correlations of 0.55–0.64 with self reports show that, although the concepts are related, the KneeTest measures a different aspect of functioning.

Two previous studies provide some evidence on the reproducibility and construct validity of the KneeTest. Van den Dikkenberg et al reported an ICC of 0.81 for interobserver reliability of the orignial KneeScore in a group of 37 healthy controls (13). Witvrouw et al provided some evidence for construct validity, by comparing the original KneeScore with the WOMAC subscores for physical functioning and pain (11).

No previous studies reported the absolute measurement error (e.g., limits of agreement) of the KneeTest. We found rather wide limits of agreement, indicating that the absolute measurement error of one individual measurement is quite large. The absolute measurement error might be decreased by repeating activities and improving standardization of test performance. Although the high reliability indicates that the test is useful for research purposes, the wide limits of agreement limit the use of the test for individual patient monitoring.

In fact, many performance-based tests as well as self-report questionnaires have large absolute measurement error. However, this is often not reported. In our systematic review (Terwee CB: unpublished observation), absolute measurement error was reported for only 10 of 24 (42%) performance-based tests, and only 1 test (the Iowa Level of Assistance Scale) (3) received a satisfactory rating. Limits of agreement of the WOMAC ranged from mean difference $\pm$ 17–29 points (scale 0–100) in previous studies (28,29). In the only Dutch validation study of the WOMAC (30), the limits of agreement ranged from –10.7 to

Table 4. Effect sizes of KneeScore 2 in 92 patients*

	Preoperative	Postoperative	Effect size
Locomotion	25.2 ± 14.5	39.2 ± 14.0	0.97
Rise and descend	20.4 ± 13.6	31.1 ± 15.0	0.79
Transfers	28.0 ± 12.7	35.1 ± 14.0	0.56
Lift and move objects	27.6 ± 13.3	38.4 ± 12.9	0.81
KneeScore 2	25.3 ± 11.9	35.9 ± 12.8	0.89

* Values are mean ± SD. Effect size is mean difference between preoperative and postoperative patients divided by the SD of preoperative patients.

18.7 (data obtained from LD Roorda). This measurement error is much larger than the minimal important difference, defined as 6–7 points for the WOMAC (31), which limits the use of the WOMAC for individual patient monitoring. We did not find any data on the reproducibility of the SF-36 in patients with knee OA.

We agree with other authors (10,11) that performance-based measures and self-report questionnaires are complementary and both useful to comprehensively evaluate functioning in TKR. Although the KneeTest requires special equipment and takes 15–30 minutes to perform, the KneeTest provides a more comprehensive and valid assessment of functioning as compared with simple tests due to the multiple activities that are assessed. In this study, these aspects were all summarized into KneeScore 2, but individual item scores and specific parameter scores for each activity are available and can be used to assess specific aspects of functioning. The value of the KneeTest could be further improved by developing clinically relevant subscores that can reveal which aspects of quality of movement are affected (e.g., endurance, coordination, or range of motion).

A limitation of our study should be acknowledged. To assign activities to the clusters, ideally factor analyses should have been performed, but this was not possible because of a lack of statistical power. Other divisions of activities into clusters may be possible, but a different cluster division will have no implications for the total KneeScore 2.

KneeScore 2 was found to have a high correlation with the mean duration of the 23 activities, which therefore seems to be a good predictor of the KneeScore. Because time alone has been reported to be an inadequate representation of concept of functional status (17), a lower correlation is desirable. The high correlation might be due to the scorings algorithm. The correlations between the selected functional parameters of each activity and the duration of that activity were moderate (mean 0.49; range 0.27–0.69), and only 6% of these correlations were >0.90. After averaging the selected parameter scores into activity scores, the correlation between those activity scores and the duration of the activities increased (mean 0.77; range 0.39–0.91). Further study of the scoring algorithm is recommended. Furthermore, it might be interesting to extract other parameters from the signals of the accelerometers (i.e., coordination or energy expenditure) that are not related to duration.

The shortened version KneeTest is a useful performance-based measure for research in patients with knee OA undergoing TKR, with good reliability and validity. Further research is required to improve its usefulness in clinical practice.

ACKNOWLEDGMENTS

We would like to thank the physiotherapists who participated in this study: Simon de Witte, Renier Bremer, Linus Jan Luikens (Gemini Hospital Den Helder), Ton van Os (Erasmus University Rotterdam), Leo Wetzels, Lisette Verkijk, Henk Creusen, and Bert Eyck (Atrium Medical Center Brunssum). We would also like to thank Sander Heikens and Bas Pijls (McRoberts, The Hague) for their technical support and supervision during and after the test days; orthopedic surgeon Wil Meijers for his help with including patients for the study, Sita Bierma (Erasmus University Rotterdam); Robert Pütz, Markus Wallach, and Christian Zimpel (Hogeschool Zuyd, Heerlen) for their help on the test days; and Nicolette van den Dikkenberg and Onno Meijer (Faculty of Movement Sciences VU University) for their help with the design of this study.

REFERENCES

1. Lingard E, Hashimoto H, Sledge C. Development of outcome research for total joint arthroplasty. J Orthop Sci 2000;5:175–7.
2. Rejeski WJ, Ettinger WH Jr, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67.
3. Shields RK, Enloe LJ, Evans RE, Smith KB, Steckel SD. Reliability, validity, and responsiveness of functional tests in patients with total joint replacement. Phys Ther 1995;75:169–79.
4. Hoeymans N, Wouters ER, Feskens EJ, van den Bos GA, Kromhout D. Reproducibility of performance-based and self-reported measures of functional status. J Gerontol A Biol Sci Med Sci 1997;52:M363–8.
5. Elam JT, Graney MJ, Beaver T, el Derwi D, Applegate WB, Miller ST. Comparison of subjective ratings of function with observed functional ability of frail older persons. Am J Public Health 1991;81:1127–30.
6. Steultjens MP, Roorda LD, Dekker J, Bijlsma JW. Responsiveness of observational and self-report methods for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 2001;45:56–61.
7. Sager MA, Dunham NC, Schwantes A, Mecum L, Halverson K, Harlowe D. Measurement of activities of daily living in hospitalized elderly: a comparison of self-report and performance-based methods. J Am Geriatr Soc 1992;40:457–62.
8. Myers AM, Holliday PJ, Harvey KA, Hutchinson KS. Functional performance measures: are they superior to self-assessments? J Gerontol 1993;48:M196–206.
9. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. 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.
10. Kennedy D, Stratford PW, Pagura SM, Walsh M, Woodhouse LJ. Comparison of gender and group differences in self-report and physical performance measures in total hip and knee arthroplasty candidates. J Arthroplasty 2002;17:70–7.
11. Witvrouw E, Victor J, Bellemans J, Rock B, van Lummel R, van der Slikke R, et al. A correlation study of objective functionality and WOMAC in total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc 2002;10:347–51.
12. Ware JE Jr, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.
13. Van den Dikkenberg N, Meijer OG, van der Slikke RM, van Lummel RC, van Dieen JH, Pijls B, et al. Measuring functional abilities of patients with knee problems: rationale and construction of the DynaPort KneeTest. Knee Surg Sports Traumatol Arthrosc 2002;10:204–12.
14. Kreibich DN, Vaz M, Bourne RB, Rorabeck CH, Kim P, Hardie R, et al. What is the best way of assessing outcome after total knee replacement? Clin Orthop Relat Res 1996;331:221–5.
15. Parent E, Moffet H. Comparative responsiveness of locomotor tests and questionnaires used to follow early recovery after total knee arthroplasty. Arch Phys Med Rehabil 2002;83:70–80.
16. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW. Get

up and go test in patients with knee osteoarthritis. Arch Phys Med Rehabil 2004;85:284–9.
17. Stratford PW, Kennedy D, Pagura SM, Gollish JD. The relationship between self-report and performance-related measures: questioning the content validity of timed tests. Arthritis Rheum 2003;49:535–40.
18. Chao EY, Laughman RK, Stauffer RN. Biomechanical gait evaluation of pre and postoperative total knee replacement patients. Arch Orthop Trauma Surg 1980;97:309–17.
19. Insall JN, Dorr LD, Scott RD, Scott WN. Rationale of the Knee Society clinical rating system. Clin Orthop Relat Res 1989; 248:13–4.
20. Lingard EA, Katz JN, Wright RJ, Wright EA, Sledge CB, and the Kinemax Outcomes Group. Validity and responsiveness of the Knee Society Clinical Rating System in comparison with the SF-36 and WOMAC. J Bone Joint Surg Am 2001;83-A:1856–64.
21. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. New York: Oxford University Press; 1989.
22. Lord FM, Novick MR. Statistical theories of mental test scores. Reading (MA): Addison-Wesley; 1968.
23. De Vet HC. Observer reliability and agreement. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. Boston University: John Wiley & Sons Ltd; 1998. p. 3123–8.
24. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
25. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996;18:979–92.
26. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.
27. Bot SD, Terwee CB, van der Windt DA, Bouter LM, Dekker J, de Vet HC. Psychometric evaluation of self-report questionnaires: the development of a checklist. In: Ader HJ, Mellenbergh GJ, editors. Second workshop on research methodology; 25–27 June, 2003. Amsterdam: VU University Amsterdam; 2003. p. 161–8.
28. Faucher M, Poiraudeau S, Lefevre-Colau MM, Rannou F, Fermanian J, Revel M. Algo-functional assessment of knee osteoarthritis: comparison of the test-retest reliability and construct validity of the WOMAC and Lequesne indexes. Osteoarthritis Cartilage 2002;10:602–10.
29. 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.
30. Roorda LD, Jones CA, Waltz M, Lankhorst GJ, Bouter LM, van der Eijken JW, et al. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann Rheum Dis 2004;63:36–42.
31. Angst F, Aeschlimann A, Stucki G. Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum 2001;45:384–91.