Knee Surg Sports Traumatol Arthrosc (2019) 27:2898–2909
DOI 10.1007/s00167-017-4789-y

KNEE

# The OARSI core set of performance-based measures for knee osteoarthritis is reliable but not valid and responsive

**J. J. Tolk[1] · R. P. A. Janssen[1] · C. A. C. Prinsen[2] · D. A. J. M. Latijnhouwers[1] · M. C. van der Steen[3] · S. M. A. Bierma-Zeinstra[4,5] · M. Reijman[1,5]**

Received: 20 August 2017 / Accepted: 27 October 2017 / Published online: 11 November 2017

**Abstract**

*Purpose* The Osteoarthritis Research Society International has identified a core set of performance-based tests of physical function for use in people with knee osteoarthritis (OA). The core set consists of the 30-second chair stand test (30-s CST), 4 × 10 m fast-paced walk test (40 m FPWT) and a stair climb test. The aim of this study was to evaluate the reliability, validity and responsiveness of these performance-based measures to assess the ability to measure physical function in knee OA patients.

*Methods* A prospective cohort study of 85 knee OA patients indicated for total knee arthroplasty (TKA) was performed. Construct validity and responsiveness were assessed by testing of predefined hypotheses. A subgroup ($n = 30$) underwent test–retest measurements for reliability analysis. The Oxford Knee Score, Knee injury and Osteoarthritis Outcome Score—Physical Function Short Form, pain during activity score and knee extensor strength were used as comparator instruments. Measurements were obtained at baseline and 12 months after TKA.

*Results* Appropriate test–retest reliability was found for all three tests. Intraclass correlation coefficient (ICC) for the 30-s CST was 0.90 (95% CI 0.68; 0.96), 40 m FPWT 0.93 (0.85; 0.96) and for the 10-step stair climb test (10-step SCT) 0.94 (0.89; 0.97). Adequate construct validity could not be confirmed for the three tests. For the 30-s CST, 42% of the predefined hypotheses were confirmed; for the 40 m FPWT, 27% and for the 10-step SCT 36% were confirmed. The 40 m FPWT was found to be responsive with 75% of predefined hypothesis confirmed, whereas the responsiveness for the other tests could not be confirmed. For the

✉ J. J. Tolk
jaap.tolk@mmc.nl

R. P. A. Janssen
R.Janssen@mmc.nl

C. A. C. Prinsen
C.Prinsen@vumc.nl

D. A. J. M. Latijnhouwers
D.Latijnhouwers@student.maastrichtuniversity.nl

M. C. van der Steen
Marieke.vd.Steen@catharinaziekenhuis.nl

S. M. A. Bierma-Zeinstra
S.Bierma-Zeinstra@erasmusmc.nl

M. Reijman
Max.Reijman@mmc.nl

[1] Department of Orthopedic Surgery and Trauma, Máxima Medical Center, Postbus 90052, 5600 PD Eindhoven, The Netherlands

[2] Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam Public Health (APH) Research Institute, De Boelelaan 1089a, 1081 HV Amsterdam, The Netherlands

[3] Department of Orthopedic Surgery, Catharina Hospital Eindhoven, P.O. Box 1350, 5602 ZA Eindhoven, The Netherlands

[4] Department of General Practice, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands

[5] Department of Orthopedic Surgery, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, Netherlands

30 s CST and 10-step SCT, only 50% of hypotheses were confirmed.

*Conclusions* The three performance-based tests had good reliability, but poor construct validity and responsiveness in the assessment of function for the domains sit-to-stand movement, walking short distances and stair negotiation. The findings of the present study do not justify their use for clinical practice.

*Level of evidence* Level 1. Diagnostic study.

**Keywords** Total knee arthroplasty · Osteoarthritis · Functional outcome · Performance-based measures

## Introduction

Knee osteoarthritis (OA) has large societal, psychological and physical burdens for patients affected by the disease [1]. Knee OA patients often experience pain and restrictions in physical functioning [2]. Important goals of knee OA treatment with total knee arthroplasty (TKA) are pain relief and improvement of physical function [3].

The evaluation of treatment outcome after TKA should at least assess the domains pain, function and a global assessment [4]. For the measurement of physical function, self-reported measures of function and testing of the execution of a specific task associated with function (performance-based tests) can be used [5]. Whereas patient reported outcome measures (PROMs) assess what patients perceive they can do, performance-based measures aim to quantify what patients actually can do [6]. When measuring change in physical function after TKA, a discrepancy is observed between the results of these methods [7, 8]. This leads to the idea that these two types of outcome measurement instruments, although being related, measure different aspects of the construct physical functioning [7, 9, 10]. Integration of both types of measurement in an assessment continuum is suggested, and considered complementary in the evaluation of physical function [5, 11].

The functional tasks that are most relevant to measure are pathology and population specific [12]. The three most relevant functional domains for knee OA are level walking, stair negotiation and sit-to-stand movement [13]. Impairment on these domains is classified as ‘activity limitations’ on the World Health Organisation International Classification of Functioning, Disability and Health (ICF) [14].

Based on the currently available evidence and expert consensus, the Osteoarthritis Research Society International (OARSI) identified a set of performance-based tests to assess these functional domains [9, 13]. The aimed construct of measurement is physical function, which is related to the ability to “move around” and “perform daily activities” and can be classified as activities using the ICF model [9, 13, 14]. The core set consist of the 30-second chair stand test (30-s CST), 4 × 10 m fast-paced walk test (40 m FPWT) and a stair climb test [13].

For tests to be usable in both clinical practice and research, measurement properties should be appropriate [15, 16]. Data on the reliability, validity and responsiveness of the OARSI core set of performance-based measures are either unavailable or from low quality studies [9]. Therefore, good quality research investigating measurement properties of these performance measures is necessary [6, 9]. The aim of the current study was to evaluate the reliability, validity and responsiveness of the core set performance-based measures for the measurement of physical function in knee OA patients.

## Materials and methods

A prospective cohort study of patients indicated for TKA was performed. Evaluation of measurement properties of the 30-s CST, 40 m FPWT and 10-step chair climb test (10-step SCT) was conducted following the COSMIN methodology (COnsensus based Standards for the selection of health status Measurement INstruments) [15]. The COSMIN checklist is a consensus-based checklist and can be used to evaluate the methodological quality of studies on the measurement properties of health status measurement instruments [15]. The Máxima Medical Center Medical Ethics Committee approved the study (registration code 2014-73).

### Patient population

All symptomatic knee OA patients scheduled for primary TKA in the Máxima Medical Center were eligible for inclusion. Exclusion criteria were comorbidity leading to inability to perform the performance-based tests, insufficient knowledge of the Dutch language leading to inability to fill out the study questionnaires and inability to visit follow-up appointments. If the patient met the criteria and was willing to participate, an informed consent form was signed.

### Study procedures

At baseline the following clinical parameters were recorded; side of operation, gender, age, and body mass index (BMI).

Testing procedures took place at the outpatient clinic of the Máxima Medical Center, in a designated testing area by a research nurse. Measurement of the OARSI core set of performance-based tests was executed strictly according to the manual provided by the OARSI, following a standardized protocol with the following fixed order of tests [13]. Measurements were obtained pre-operatively and 12 months post-operative.

## Performance-based measures

### *30-s CST*

The 30-s CST is a performance-based measure that evaluates the activity 'sit-to-stand movement'[13]. The test is executed by scoring the maximum amount of complete chair stand movements during 30 seconds. A full sit-to-stand and consecutive stand-to-sit cycle is counted as one chair stand. A 43 cm high, straight back chair without arm rests was used. To date no previous reliability reports specifically for knee OA patients are available. In a combined group of hip and knee OA patients, an excellent reliability is reported, with an intraclass correlation coefficients (ICC) of 0.95 (SD 0.93–0.97), and a standard error of measurement (SEM) of 0.7 repetitions [17]. Construct validity and responsiveness have not been reported previously in knee OA patients.

### *40 m FPWT*

The 40 m FPWT assesses the activity 'walking short distances' [13]. It scores the maximal walking speed on a marked walkway of four times 10 m, excluding turns. The result is expressed as speed in m/s. There are no previous reports on the reliability of this version of the 40 m FPWT [18]. Kennedy et al. report on a similar walk test, scoring walking speed using a walkway of two times 20-m. Their results show good reliability with an ICC of 0.91 (SD 0.81, 0.97) and SEM of 1.73 m/s (SD 1.39–2.29) [18]. No previous reports on construct validity of the 40 m FPWT are available in the literature [9].

### *10-step SCT*

For assessment of the activity 'stair negotiation', no specific stair climb test is advised by the OARSI [13]. In the present study, the 10-step stair climb test (10-step SCT) was selected, as the stair in the testing area had ten steps. The step height was 18.8 cm and depth 22.4 cm. The time needed to ascend and descent these steps is recorded in seconds.

No previous reports on reliability of the 10-step SCT are available. Almeida et al. reported excellent reliability with an ICC of 0.94 (SD 0.55–0.98) and a SEM of 2.35 s for the 11-step stair test in knee OA patients [19]. The 11-step SCT is essentially the same test as the 10-step version, with the only difference that the stairway used has one step more.

## Comparator instruments

### *KOOS-PS*

The Knee injury and Osteoarthritis Outcome Score—Physical Function Short Form (KOOS-PS) Dutch version is a 7-item questionnaire that assesses the construct physical function. From a 5-point Likert scale question, a normalized score is calculated (0 indicating no symptoms and 100 indicating extreme symptoms) [20]. KOOS-PS has good reliability, face and content validity and ability to detect change over time in knee OA patients [20–23].

### *OKS*

The Dutch version of the Oxford Knee Score (OKS) is a 12-item PROM designed to measure function and pain after TKA. 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 [24]. It is short, reproducible, valid and sensitive to clinically important changes [24]. The OKS has adequate internal consistency and test retest reliability, good face, content and construct validity and good sensitivity and responsiveness in knee OA patients [21].

### *EQ-5D*

The Dutch version of the EuroQol 5D-3L (EQ-5D) is a 5-item PROM, measuring generic health status [25]. Scoring the lowest score on the EQ-5D index indicates the worst health state possible and a score of 1 represents the best possible health state [25]. The EQ-5D has good reliability and validity in knee OA patients [26].

### *NRS pain*

Numerical Rating Scale (NRS) for pain during activity (NRS pain) was used to measure level of pain during activity. The scale consists of 11 points in which the patient can score the pain during activities in general from 0 to 10. A score of 0 represented 'no pain' and a score of 10 represented 'worst imaginable pain'. The NRS has good reliability and responsiveness [23].

### *Anchor question*

At 12-month postoperative follow-up, a 7-point Likert scale anchor question was scored for change in activities of daily living. Response options ranged from 1 (a lot worse) to 7 (very much improved).

*ROM*

Range of motion (ROM) of the affected knee was measured in supine position using a goniometer, considering the bony landmarks of the greater trochanter, lateral femoral condyle, and lateral malleolus. Maximal flexion was scored as positive value and an extension deficit was scored as negative value. In knee OA, ROM measurement has adequate reliability with a reliability coefficient of 0.81 for extension and 0.96 for flexion [27].

*Quadriceps strength*

To determine quadriceps strength of the affected leg, maximal isometric knee extensor strength was measured using a handheld dynamometer (HHD). Testing took place in an upright position. The HHD was positioned perpendicular to the anterior aspect of the tibia, 5 cm proximal of the medial malleolus. A protective shin guard was used for patient comfort as well as standardisation of HHD placement. Three consecutive measurements were obtained, of which the highest value was used for analysis. An HHD is a widely used, reliable, and valid instrument to measure knee extensor strength, with good reliability in OA patients (ICC 0.94) [28].

### Evaluation of the measurement properties

*Reliability*

Reliability is defined as the extent to which scores for patients, who have not changed, are the same for repeated measurement under similar conditions [15]. To evaluate the reliability of the three performance-based tests, test–retest measurements were obtained in a random subset of patients. After initial measurement (T0) patients rested for 30 min, after which a second round of testing was performed (T0_1). This test–retest design was considered appropriate as the resting period allows full recovery from the performed tests, and the tested function can be assumed to remain stable over the testing period. Circumstances, setting, order of the three tests and instructions in the retest setting were identical to the first round of testing.

Reliability analysis consisted of determining ICC for absolute agreement with the corresponding 95% confidence intervals (CI), SEM, and smallest detectable change (SDC). An ICC value $>0.70$ is considered appropriate [29, 30].

*Construct validity*

There is no ‘Gold Standard’ available for the assessment of the functional domains level walking, stair negotiation and sit-to-stand movement in knee OA. Therefore, determining construct validity is the designated method to analyse the degree to which the studied measurement instruments are measuring the constructs that they aim to measure [15, 16, 31]. This method is internationally accepted and recommended by the COSMIN for these circumstances [15, 16, 31]. Predefined hypotheses were formulated on the relationships of performance-based tests scores with scores on other instruments measuring similar or dissimilar construct [16, 31]. A panel comprising of four experts in the field of outcome measurement in knee OA (orthopedic surgeon, orthopedic resident and Ph.D. candidate, specialist in measurement property analysis and methodologist), formulated 11 to 15 hypotheses for each measurement instrument under study. An overview of the hypotheses can be found in Table 3.

The predefined hypotheses consisted of both convergent and discriminant validity hypotheses, and comparative hypothesis on a closer relationship with similar compared to dissimilar constructs. The hypothesis included direction and magnitude of the expected results. In general, we hypothesised the following. The performance-based measures would be moderately correlated to PROMs and quadriceps strength. PROMs have a stronger correlation with pain scores than with the performance-based measures. Performance-based measures were expected to have a stronger correlation with PROMs measuring functional outcome than with a PROM measuring general health. Specific questions of the PROMs regarding walking, stair negotiation and sit-to-stand movement were expected to correlate stronger to their respective performance-based measure than to the total score of the PROM.

Correlations of measurements with similar constructs were expected to be at least moderate $\geq 0.4$ or $\leq -0.4$. Measurements that were unrelated or had different constructs were expected to have a poor correlation $[-\geq 0.39; \leq 0.39]$. The performance-based tests are assumed valid if at least 75% of the predefined hypotheses are confirmed [29, 30].

*Responsiveness*

Responsiveness is defined as the ability of the instruments to detect change over time in the construct measured [15, 16, 31]. In the absence of a gold standard, the assessment of responsiveness relies on hypotheses testing (i.e. a construct approach) [15, 16, 31]. These hypotheses concern the expected relationships between changes on the studied instruments and changes on other instruments that measure similar or different constructs with adequate responsiveness [16, 30, 31]. These hypotheses, with expected direction and magnitude of the correlations, were formulated a priori.

The performance-based tests are assumed to be adequately responsive if minimally 75% of the predefined hypotheses are confirmed [29, 30]. The responsiveness hypothesis can be found in Table 5. In summary, it was hypothesised that the anchor question was moderately correlated to change in the performance-based measures scores. Only a moderate correlation was expected, because experienced change in functional ability is not exactly the same construct as actual change in execution of the task. Furthermore, we hypothesised that the change in PROMs is more correlated to pain, than to change in the performance-based test scores.

### Statistical analysis

Statistical analysis was performed with SPSS statistics version 24.0 (IBM corporation). The reliability analysis was performed using a two-way random model with absolute agreement. SEM was calculated using the formula: *Standard Deviation* (*SD*) *difference/*√*n*. Where *n* represents the number of measurement repetitions; $n=2$ for the present study. The SDC was calculated as $1.96 \times \sqrt{2} \times \mathrm{SEM}$ [32]. For the construct validity and responsiveness analysis Pearson or Spearman correlation coefficients were calculated, depending on normality of data distribution. Comparison of performance-based measures and PROM scores before and after TKA was conducted using a paired samples *t* test or Wilcoxon signed-rank test, depending on normality of data distribution. The sample size was based on the COSMIN criteria, aiming for a good score for the construct validity and responsiveness analysis (≥ 50 patients) and fair for reliability assessment (≥ 30 patients) [15, 30].

## Results

### Patient characteristics

Between April and October 2015, 85 consecutive patients with knee OA were included. The baseline characteristics are described in Table 1. Number of patients included in the reliability, construct validity and responsiveness analysis and reasons for loss to follow-up are summarised in Fig. 1.

### Measurement properties

*Reliability analysis*

Test–retest measurements were performed in a random subgroup consisting of the first 30 patients that were included in the study. Mean test scores and reliability parameters are presented in Table 2.

**Table 1** Baseline characteristics

| | Total cohort (*n* = 85) | Reliability analysis cohort (*n* = 30) |
|---|---|---|
| Age (years) | 69.3 (± 8.2) | 67.8 (± 7.7) |
| Gender, female [*n* (%)] | 46 (57) | 13 (43) |
| Side affected, right [*n* (%)] | 41 (48) | 17 (57) |
| BMI (kg/m$^2$) | 29.6 (± 5.0) | 29.9 (± 5.6) |
| Maximal flexion (°) | 110 (± 17.0) | 106 (± 18.9) |
| Extension deficit (°) | 4 (± 7.0) | 4 (± 6.5) |

Data are presented as mean and standard deviation between parentheses, or reported otherwise as mentioned

*Construct validity (hypothesis testing)*

Spearman's correlation coefficients for the construct validity analysis are presented in Table 3. Confirmation of 75% or more of the predefined hypotheses was achieved by none of the three performance-based measures. 5/12 (42%) were confirmed for the 30-s CST, 4/15 (27%) for the 40 m FPWT and 4/11 (36%) for the 30-s SCT.

### Responsiveness

The scores of the performance-based measures at baseline and after TKA at 12-month follow-up are presented in Table 4. All performance-based measures, PROMs and the NRS pain score showed significant improvement at 12-month follow-up. Only the use of a handrail during the 10-step SCT did not show significant change. On the anchor question for change in activities of daily living the mean score at 12 month follow-up was 6.2 (95% CI 5.9–6.5), this represents 'much improved'. Spearman' s correlation coefficients for responsiveness analysis are presented in Table 5. For the 30-s CST, 4/8 (50%) of the hypothesis were confirmed, for the 40 m FPWT, 6/8 (75%) and for the 10-step SCT, 4/8 (50%).

## Discussion

The present study showed good reliability of the OARSI recommended core set of performance-based measures. However, based on a low percentage of confirmation of our predefined hypotheses, construct validity and responsiveness of the tests were poor.

Test–retest reliability of the three performance-based measures is adequate, as the presented ICC values are well above 0.70, which is considered acceptable [33]. This is

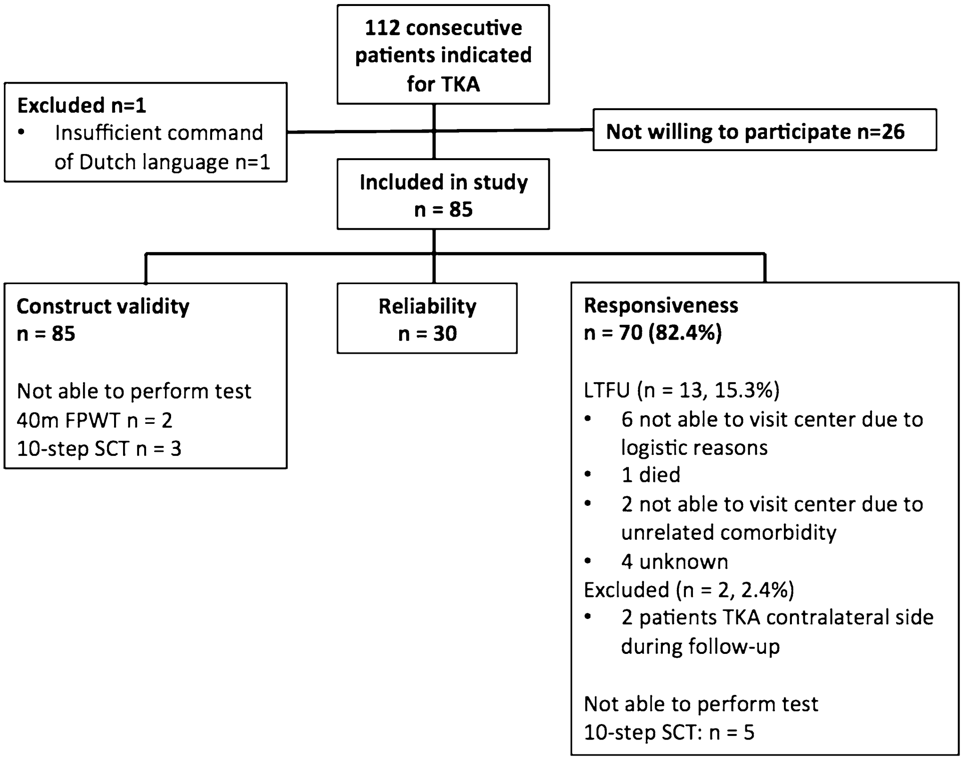
112 consecutive patients indicated for TKA
Excluded n=1
• Insufficient command of Dutch language n=1
Not willing to participate n=26
Included in study n = 85
Construct validity n = 85
Not able to perform test
40m FPWT n = 2
10-step SCT n = 3
Reliability n = 30
Responsiveness n = 70 (82.4%)
LTFU (n = 13, 15.3%)
• 6 not able to visit center due to logistic reasons
• 1 died
• 2 not able to visit center due to unrelated comorbidity
• 4 unknown
Excluded (n = 2, 2.4%)
• 2 patients TKA contralateral side during follow-up
Not able to perform test
10-step SCT: n = 5

**Fig. 1** Number of patients included in analysis and reasons for loss to follow-up. *LTFU* lost to follow-up

**Table 2** Reliability analysis ($n=30$)

| | Mean score baseline | Mean retest score | Mean of difference (baseline–retest score) | ICC | SEM | SDC |
|---|---|---|---|---|---|---|
| 30-s CST (stands) | 9.0 (7.9–10.1) | 9.8 (8.6–11.0) | 0.8 (0.4 to 1.3) | 0.90 (0.68–0.96) | 0.85 | 2.4 |
| 40 m FPWT (m/s) | 1.30 (1.16–1.44) | 1.32 (1.20–1.45) | − 0.02 (− 0.08 to 0.03) | 0.93 (0.85–0.96) | 0.10 | 0.27 |
| 10-step SCT (s) | 16.7 (13.5–19.9) | 16.6 (13.5–19.7) | 0.1 (− 1.0 to 1.1) | 0.94 (0.89–0.97) | 1.98 | 5.5 |

Data are presented as mean and 95% confidence interval between parentheses, or reported otherwise as mentioned

*ICC* intraclass correlation coefficient, *SEM* standard error of measurement, *SDC* smallest detectable change

in line with previous reports on test–retest reliability for these tests [18, 19]. The SDC values reported in the present study for the 30-s CST and 10-step SCT are similar to those reported in the literature [17, 19]. There is no consensus on what SDC value is acceptable [32]. From a clinical point of view, the SDC's of 2.5 stands for the CST and 0.27 m/s for the 40 m FPWT reported in the present study seem reasonable. This is different however for the 10-step SCT. With an SDC of 5.5 s, an individual patient has to improve or deteriorate almost 1/3 of the mean time taken for the initial test, to be certain a change has occurred. From a clinical perspective, this seems quite a large difference, resulting in a low sensitivity to change on the tested functional domain.

In the construct validity assessment, the necessary 75% hypotheses confirmation was achieved by none of the performance-based tests. The main reason for this was the rejection of all convergent hypotheses for correlations between the performance-based measures and the patient reported measures of function. As PROMs are, by definition, subjective measures of function, only a moderate correlation with the more objective measurements of the performance-based measures was expected. For example, PROMs are known to be more related to pain than to actual execution of the task at hand [7, 8, 34] as was also found in the present study. Self-reported and performance-based assessment of activities are inherently linked, considering that both methods aim to measure the same 'activities' defined in the ICF theoretical framework [14]. In our view for performance-based measures to be clinically relevant, some relation between experienced performance and the result of the performance-based measure of this activity should be present.

**Table 3** Construct validity

| Predefined hypotheses | 30-s CST | | 40 m FPWT | | 10-step SCT (CC in opposite direction[a]) | |
|---|---|---|---|---|---|---|
| | Correlation coefficient | Hypothesis confirmed | Correlation coefficient | Hypothesis confirmed | Correlation coefficient[a] | Hypothesis confirmed |
| Moderate correlation with KOOS-PS (≤ −0.4)[a] | −0.33 | No | −0.25 | No | 0.26 | No |
| Moderate correlation with OKS (≥ 0.4)[a] | 0.35 | No | 0.32 | No | −0.33 | No |
| Moderate correlation with quadriceps strength (≥ 0.4)[a] | 0.60 | Yes | 0.64 | Yes | −0.74 | Yes |
| Unrelated with EQ-5D [−0.35; 0.35] | 0.25 | Yes | 0.18 | Yes | −0.18 | Yes |
| Correlation with KOOS-PS is minimal 0.1 stronger than with EQ-5D | −0.33/0.25 | No | −0.25/0.18 | No | 0.26/−0.18 | No |
| Correlation with OKS is minimal 0.1 stronger than with EQ-5D | 0.35/0.25 | Yes | 0.32/0.18 | Yes | −0.33/−0.18 | Yes |
| 'Absolute' correlation between NRS pain and KOOS-PS is minimal 0.1 higher than between performance-based measure and NRS pain | 0.37/−0.10 | Yes | 0.37/−0.07 | Yes | 0.37/0.01 | Yes |
| 'Absolute' correlation between NRS pain and OKS is minimal 0.1 higher than performance-based measure and NRS pain | −0.45/−0.10 | Yes | NA | | NA | |
| 'Absolute' correlation 30-s CST with KOOS-PS Question 3 is minimal 0.1 higher than with KOOS-PS (total score) | −0.21/−0.33 | No | NA | | NA | |
| 'Absolute' correlation 30-s CST with KOOS-PS Question 3 is minimal 0.1 higher than with OKS | −0.21/0.35 | No | NA | | NA | |
| 'Absolute' correlation 30-s CST with KOOS-PS Question 3 is minimal 0.1 higher than with EQ-5D Score | −0.21/0.25 | No | NA | | NA | |
| Moderate correlation 30-s CST with KOOS-PS Question 3 (≤ −0.4) | −0.21 | No | NA | | NA | |
| 'Absolute' correlation 40 m FPWT with EQ-5D Question 1 is minimal 0.1 stronger than with KOOS-PS | NA | | −0.09/0.26 | No | NA | |
| 'Absolute' 40 m FPWT with EQ-5D Question 1 is minimal 0.1 stronger than with Oxford Knee Score | NA | | −0.09/0.32 | No | NA | |
| 'Absolute' correlation 40 m FPWT with EQ-5D Question 1 is minimal 0.1 higher than with EQ-5D Score | NA | | −0.09/0.18 | No | NA | |

**Table 3** (continued)

| Predefined hypotheses | 30-s CST | | 40 m FPWT | | 10-step SCT (CC in opposite direction[a]) | |
|---|---|---|---|---|---|---|
| | Correlation coefficient | Hypothesis confirmed | Correlation coefficient | Hypothesis confirmed | Correlation coefficient[a] | Hypothesis confirmed |
| 'Absolute' correlation 40 m FPWT with OKS Question 6 is minimal 0.1 stronger than with KOOS-PS | NA | | − 0.03/− 0.25 | No | NA | |
| 'Absolute' correlation 40 m FPWT with OKS Question 6 is minimal 0.1 stronger than with OKS | NA | | − 0.03/0.32 | No | NA | |
| 'Absolute' correlation 40 m FPWT with OKS Question 6 is minimal 0.1 stronger than with EQ-5D Score | NA | | − 0.03/0.18 | No | NA | |
| Moderate correlation 40 m FPWT with EQ-5D Question 1 ($\leq -0.4$) | NA | | − 0.09 | No | NA | |
| Moderate correlation 40 m FPWT with OKS Question 6 ($\leq -0.4$) | NA | | − 0.03 | No | NA | |
| 'Absolute' correlation 10-step SCT with OKS Question 12 is minimal 0.1 stronger than with KOOS-PS | NA | | NA | | 0.22/0.26 | No |
| 'Absolute' correlation 10-step SCT with OKS Question 12 is minimal 0.1 stronger than with OKS | NA | | NA | | 0.22/− 0.33 | No |
| 'Absolute' correlation 10-step SCT with OKS Question 12 is minimal 0.1 stronger than with EQ-5D | NA | | NA | | 0.22/− 0.18 | No |
| Moderate correlation 10-step SCT with OKS Question 12 ($\leq -0.4$) | NA | | NA | | 0.22 | No |
| Hypothesis confirmed | 5/12 | 42% | 4/15 | 27% | 4/11 | 36% |

*KOOS-PS* Knee injury and Osteoarthritis Outcome Score—Physical Function Short Form, *OKS* Oxford Knee Score, *NA* not applicable, *NRS pain* Numerical Rating Scale for pain during activity, *30-s CST* 30-second Chair Stand Test, *40 m FPWT* 40-m Fast-Paced Walk Test, *10-step SCT* 10-step Stair Climb Test

[a] The 10-step SCT is scored in the opposite direction of the 30-s CST and 40 m FPWT (better performance is a lower score), therefore the hypothesised correlations are in the opposite directions

However, even the moderate correlations we predicted were not met, resulting in poor construct validity.

An explanation for the poor construct validity might be that timed measures of performance did not fully capture impairment on the activities at hand. The time taken to execute a task is not the only factor in the performance of this task in daily living. A patient might execute the activity swiftly, but if the quality of performance is affected by, for example, limping or instability, it can still be considerably impaired [11, 35, 36]. Such an impairment cannot be captured by merely timing the activity [35, 36]. Another explanation for discordance between self-reported and performance-based measurement of function can be underrepresentation [37]. Whereas the OKS and KOOS-ps measure the general construct physical function, the performance-based tests under study aim to quantify performance on specific functional tasks. The narrower construct of the performance test might not be represented by these two PROMs used as comparative instruments [37]. If underrepresentation were the case, the specific questions addressing the functional tasks measured by the respective tests would be likely to correlate stronger to these tests. To

**Table 4** Performance-based measures and PROM scores before and after TKA

| | Baseline | 12-month follow-up after TKA | *p* value |
|---|---|---|---|
| 30-s CST (stands) | 9.2 (8.4–10.0) | 11.3 (10.3–12.4) | < 0.001 |
| 40 m FPWT (m/s) | 1.25 (1.16–1.34) | 1.38 (1.25–1.50) | 0.001 |
| Use of assistive device during 40 m FPWT (patients, *n*) | 2 | 0 | NA |
| 10-step SCT (s) | 21.8 (18.4–25.1) | 15.5 (13.9–17.1) | 0.007 |
| Use of handrail 10-step SCT (patients, *n*) | 39 | 24 | 0.40 (n.s.) |
| KOOS-PS score | 54.2 (50.8–57.5) | 28.9 (24.6–33.1) | < 0.001 |
| OKS | 21.7 (20.2–23.2) | 40.1 (38.1–42.1) | < 0.001 |
| EQ5D | 0.48 (0.42–0.55) | 0.84 (0.79–0.89) | < 0.001 |
| NRS pain | 7.6 (7.2–7.9) | 2.1 (1.6–2.7) | < 0.001 |

Data are presented as mean and 95% confidence interval between parentheses, or reported otherwise as mentioned

*n.s.* non-significant

**Table 5** Responsiveness

| | 30-s CST (change score) | | 40 m FPWT (change score) | | 10-step SCT (change score) | |
|---|---|---|---|---|---|---|
| Predefined hypotheses | Correlation coefficient | Hypothesis confirmed | Correlation coefficient | Hypothesis confirmed | Correlation coefficient | Hypothesis confirmed |
| Moderate correlation with anchor question (≥ 0.4) | 0.22 | No | 0.40 | Yes | − 0.25 | No |
| Moderate correlation with change score NRS pain during activity (≤ − 0.4) | − 0.20 | No | − 0.36 | No | 0.08 | No |
| Moderate correlation with change score KOOS-PS (≤ − 0.4) | − 0.26 | No | − 0.28 | No | 0.27 | No |
| Moderate correlation with change OKS (≥ 0.4) | 0.22 | No | 0.43 | Yes | − 0.36 | No |
| Correlation between change scores NRS pain and KOOS-PS is minimal 0.1 stronger than between NRS pain and performance-based test | 0.56/− 0.20 | Yes | 0.56/− 0.36 | Yes | − 0.56/0.08 | Yes |
| Correlation between change scores NRS pain and KOOS-PS is minimal 0.1 stronger than between KOOS-PS and performance-based test | − 0.56/− 0.26 | Yes | 0.56/− 0.28 | Yes | − 0.56/0.27 | Yes |
| Correlation between changes scores NRS pain and OKS minimal 0.1 stronger than between NRS pain and performance-based test | − 0.70/− 0.20 | Yes | − 0.70/− 0.36 | Yes | − 0.70/− 0.08 | Yes |
| Correlation between change scores NRS pain and OKS is minimal 0.1 stronger than between OKS and performance-based test | − 0.70/− 0.22 | Yes | − 0.70/0.40 | Yes | − 0.70/− 0.36 | Yes |
| Hypothesis confirmed | 4/8 | 50% | 6/8 | 75% | 4/8 | 50% |

*KOOS-PS* Knee injury and Osteoarthritis Outcome Score—Physical Function Short Form, *OKS* Oxford Knee Score, *NA* not applicable, *NRS pain* Numerical Rating Scale for pain during activity, *30-s CST* 30-s Chair Stand Test, *40 m FPWT* 40-m Fast-Paced Walk Test, *10-step SCT* 10-step Stair Climb Test account for this, we made hypothesis on correlations with these specific questions. The correlations found on these hypotheses were even lower, making underrepresentation as an explanation unlikely.

The strong relationship between pain and self-reported function found in the construct validity analysis was even more obvious in the responsiveness analysis. The change in NRS pain score was strongly related to the change in subjective scores, but unrelated to the performance-based measures. This supports claims that performance-based measures are less pain driven than PROMs, and provide a more objective view on the task performed [7, 8]. On the other hand, it is our opinion that for a test to be clinically relevant some relationship between actual change and experienced change in performance on the functional task at hand should exist. Therefore, we hypothesised that the overall change in PROM scores 1 year after TKA would correlate moderately to the change in performance-based measures. Only for the 40 m FPWT, most hypotheses in this regard were confirmed. For the other two tests, no such relationship was found. As mentioned earlier for the construct validity, underrepresentation and the inability of timed measures to fully capture impairment on the tested domains might explain the lack of responsiveness of the 30-s CST and 10-step SCT.

A remark has to be made on the comparative instruments used for the construct validity and responsiveness analysis. These consisted of a combination of objective and subjective measurements of function and general health with good reliability, construct validity and responsiveness in a knee OA population [38–41]. Other options for comparison could have been objective measures such as optoelectric- or inertia-based motion analysis systems. These measures are suitable for a strictly kinematic analysis, but their clinical relevance has not been clarified [35, 42]. Therefore, we believe that they are not suitable of the construct validity analysis in this regard. In our view, the comparative measurement instruments in the present study were the most appropriate instruments available.

To our knowledge, this is the first study assessing the most important measurement properties of the OARSI recommended core set of performance-based measures. A strength of the present study is the strictly followed, state-of-the-art methodology [15]. We report on an unselected, consecutive group of knee OA patients awaiting total knee arthroplasty in a general hospital. Previous reports on measurement properties often included a combined population of knee and hip OA patients, resulting in a more heterogeneous population [17, 18]. Combining these distinctly different groups reduces the accuracy of the previously reported data. Our findings can be considered representative for knee OA patients.

The sample size can be considered good for the construct validity and responsiveness analysis and fair for reliability assessment [15, 30]. A limitation of this research was the incomplete 12-month follow-up, the 82.4% follow-up achieved is however acceptable. For the subset of patients with incomplete data, no difference in preoperative demographics or baseline measurement was observed. Therefore, a systematic bias because of the loss to follow-up seems unlikely. The results for the reliability assessment should be interpreted with some caution as a subgroup of only 30 patients was used. There is concurrent evidence on test–retest measurements from others studies, with similar results [18, 19]. When combining these data, stronger evidence for an adequate reliability can be obtained. As mentioned earlier, the SDCs in the present study are relatively large, especially for the 10-step SCT. Test–retest measurements in a larger population would have resulted in a more precise determination of the SDC; it might be smaller than reported in the present study.

## Conclusion

The OARSI core set of performance-based measures was advised to obtain a more complete view of the functional performance of knee OA patients [13]. The 30-s CST, 40 m FPWT and 10-step SCT have good reliability, but poor construct validity and responsiveness in the assessment of function and change in function for the domains sit-to-stand movement, walking short distances and stair negotiation respectively. The findings of the present study do not justify their use for clinical practice.

**Acknowledgements** We would like to sincerely thank Christa van Doesburg, Hilke Cox and Mathias Mariam for their work in administrative and testing procedures.

**Author contribution** JT and MR have contributed to the conception and design of the study and drafting of the article. CP provided methodological support. DL has contributed to collection and assembly of data. All authors have contributed to interpretation of the data, critically revising the article for important intellectual content and have approved the final version of the article.

**Compliance with ethical standards**

**Conflict of interest** All authors have no conflicts of interest to report.

**Ethical review committee statement** The study has been performed in accordance with the ethical standards in the 1964 Declaration of Helsinki.

**Informed consent** Informed consent was obtained from all individual participants included in the study.

## References

1. Litwic A, Edwards MH, Dennison EM, Cooper C (2013) Epidemiology and burden of osteoarthritis. Br Med Bull 105:185–199
2. Hoy D, Cross M, Smith E, Nolte S, Ackerman I, Fransen M, Bridgett L, Williams S, Guillemin F, Hill CL, Laslett LL, Jones G, Cicuttini F, Osborne R, Vos T, Buchbinder R, Woolf A, March L (2014) The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis 73:1323–1330
3. Zhang W, Moskowitz RW, Nuki G, Abramson S, Altman RD, Arden N, Bierma-Zeinstra S, Brandt KD, Croft P, Doherty M, Dougados M, Hochberg M, Hunter DJ, Kwoh K, Lohmander LS, Tugwell P (2008) OARSI recommendations for the management of hip and knee osteoarthritis, Part II: OARSI evidence-based, expert consensus guidelines. Osteoarthr Cartil 16:137–162
4. Pham T, van der Heijde D, Altman RD, Anderson JJ, Bellamy N, Hochberg M, Simon L, Strand V, Woodworth T, Dougados M (2004) OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. Osteoarthr Cartil 12:389–399
5. Reiman MP, Manske RC (2011) The assessment of function: how is it measured? A clinical perspective. J Man Manip Ther 19:91–99
6. Terwee CB, Mokkink LB, Steultjens MPM, Dekker J (2006) Performance-based methods for measuring the physical function of patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. Rheumatology (Oxford) 45:890–902
7. Mizner RL, Petterson SC, Clements KE, Zeni J a, Irrgang JJ, Snyder-Mackler L (2011) Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments: a longitudinal analysis of outcomes. J Arthroplast 26:728–737
8. Stevens-Lapsley JE, Schenkman ML, Dayton MR (2011) Comparison of self-reported knee injury and osteoarthritis outcome score to performance measures in patients after total knee arthroplasty. PM R 3:541–549
9. Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL (2012) Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthr Cartil 20:1548–1562
10. Stratford PW, Kennedy DM, Woodhouse LJ (2006) Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 86:1489–1496
11. Stratford PW, Kennedy DM (2006) Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 59:160–167
12. Hegedus EJ, Vidt ME, Tarara DT (2014) The best combination of physical performance and self-report measures to capture function in three patient groups. Phys Ther Rev 19:196–203
13. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford P, Davis a M, Buchbinder R, Snyder-Mackler L, Henrotin Y, Thumboo J, Hansen P, Bennell KL (2013) OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthr Cartil 21:1042–1052
14. World Health Organization (2001) International Classification of Functioning, Disability and Health. World Health Organization, Geneva, Switzerland
15. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, De Vet HCW (2010) The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 19:539–549
16. de Vet HCW, Terwee CB, Mokkink LB, Knol DL (2011) Measurement in medicine: a practical guide. Cambridge University Press, Cambridge
17. Gill S, McBurney H (2008) Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. Physiother Res Int 13:141–152
18. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D (2005) Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord 6:3
19. Gustavo J, Almeida PT M, Carolyn A. Schroeder B, Alexandra B, Gil PT M, Kelley G, Fitzgerald PT P, Piva SR (2010) Interrater reliability and validity of the stair ascend/descend test in individuals with total knee arthroplasty. Arch Phys Med Rehabil 91:932–938
20. Perruccio AV, Lohmander LS, Canizares M, Tennant A, Hawker GA, Conaghan PG, Roos EM, Jordan JM, Maillefert J, Dougados M, Davis AM (2008) The development of a short measure of physical function for knee OA KOOS-Physical Function Shortform (KOOS-PS)—an OARSI/OMERACT initiative. Osteoarthr Cartil 16:542–550
21. Collins NJ, Misra D, Felson DT, Crossley KM, Roos EM (2011) Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS). Arthritis Care Res (Hoboken) 63(Suppl 1):S208-228
22. de Groot IB, Favejee MM, Reijman M, Verhaar JA, Terwee CB (2008) The Dutch version of the Knee Injury and Osteoarthritis Outcome Score: a validation study. Health Qual Life Outcomes 6:16
23. Ruyssen-Witrand A, Fernandez-Lopez CJ, Gossec L, Anract P, Courpied JP, Dougados M (2011) Psychometric properties of the OARSI/OMERACT osteoarthritis pain and functional impairment scales: ICOAP, KOOS-PS and HOOS-PS. Clin Exp Rheumatol 29:231–237
24. Haverkamp D, Breugem SJM, Sierevelt IN, Blankevoort L, van Dijk CN (2005) Translation and validation of the Dutch version of the Oxford 12-item knee questionnaire for knee arthroplasty. Acta Orthop 76:347–352
25. Rabin R, de Charro F (2001) EQ-5D: a measure of health status from the EuroQol Group. Ann Med 33:337–343
26. Conner-Spady BL, Marshall DA, Bohm E, Dunbar MJ, Loucks L, Khudairy A, Al Noseworthy TW (2015) Reliability and validity of the EQ-5D-5L compared to the EQ-5D-3L in patients with osteoarthritis referred for hip and knee replacement. Qual Life Res 24:1775–1784
27. Cibere J, Thorne A, Bellamy N, Greidanus N, Chalmers A, Mahomed N, Shojania K, Kopec J, Esdaile JM (2004) Reliability of the knee examination in osteoarthritis. Arthritis Rheum 50:458–468
28. Holstege MS, Lindeboom R, Lucas C (2011) Preoperative quadriceps strength as a predictor for short-term functional outcome after total hip replacement. Arch Phys Med Rehabil 92:236–241
29. Prinsen CAC, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, Williamson PR, Terwee CB, Chalmers I, Glasziou P, Williamson P, Altman D, Blazeby J, Clarke M, Devane D, Gargon E, Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, Clarke M, Schmitt J, Apfelbacher C, Spuls P, Thomas K, Simpson E, Furue M, Prinsen C, Vohra S, Rose M, King-Jones S, Ishaque S, Bhaloo Z, Boers M, Kirwan J, Wells G, Beaton D, Gossec L, D'Agostino

M, Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, Murphy M, Black N, Lamping D, McKee C, Sanderson C, Askham J, Chiarotto A, Deyo R, Terwee C, Boers M, Buchbinder R, Corbin T, Verhagen A, Vet H, Bie R, Kessels A, Boers M, Bouter L, Jones J, Hunter D, Terwee C, Bot S, Boer M, Windt D, Knol D, Dekker J, Gargon E, Gurung B, Medley N, Altman D, Blazeby J, Clarke M (2016) How to select outcome measurement instruments for outcomes included in a 'Core Outcome Set'—a practical guideline. Trials 17:449

30. Terwee CB, Bot SDM, de Boer MR, van der Windt D a, Knol WM, Dekker DL, Bouter J, de Vet LM HCW (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60:34–42
31. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, Bouter LM, de Vet HC (2010) The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 10:22
32. Atkinson G NA (1998) Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sport Med 26:217–238
33. Nunally JC, Bernstein IH (1994) Psychometric theory, 3rd edn. McGraw-Hill, New York
34. Terwee CB, van der Slikke RMA, van Lummel RC, Benink RJ, Meijers WGH, de Vet HCW (2006) Self-reported physical functioning was more influenced by pain than performance-based physical functioning in knee-osteoarthritis patients. J Clin Epidemiol 59:724–731
35. Bolink SAAN, Grimm B, Heyligers IC (2015) Patient-reported outcome measures versus inertial performance-based outcome measures: a prospective study in patients undergoing primary total knee arthroplasty. Knee 22:618–623
36. Steultjens MPM, Dekker J, van Baar ME, Oostendorp R, a. B, Bijlsma JWJ (1999) Internal consistency and validity of an observational method for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 12:19–25
37. Piva SR, Fitzgerald GK, Irrgang JJ, Bouzubar F, Starz TW (2004) Get up and go test in patients with knee osteoarthritis. Arch Phys Med Rehabil 85:284–289
38. Conaghan PG, Emerton M, Tennant A (2007) Internal construct validity of the Oxford knee scale: evidence from Rasch measurement. Arthritis Rheum 57:1363–1367
39. Davis AM, Perruccio AV, Canizares M, Hawker GA, Roos EM, Maillefert J-F, Lohmander LS (2009) Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthr Cartil 17:843–847
40. Ramkumar PN, Harris JD, Noble PC (2015) Patient-reported outcome measures after total knee arthroplasty: a systematic review. Bone Jt Res 4:120–127
41. Roos EM, Toksvig-Larsen S (2003) Knee injury and Osteoarthritis Outcome Score (KOOS)—validation and comparison to the WOMAC in total knee replacement. Health Qual Life Outcomes 1:17
42. Mills K, Hunt M a, Ferber R (2013) Biomechanical deviations during level walking associated with knee osteoarthritis: a systematic review and meta-analysis. Arthritis Care Res (Hoboken) 65:1643–1665