Acta Orthopaedica 2019; 90 (1): 15–20

Measurement properties of the OARSI core set of performance-based measures for hip osteoarthritis: a prospective cohort study on reliability, construct validity and responsiveness in 90 hip osteoarthritis patients

Jaap J TOLK [1], Rob P A JANSSEN [1], C (Sanna) A C PRINSEN [2], M (Marieke) C VAN DER STEEN [3], Sita M A BIERMA ZEINSTRA [4,5], and Max REIJMAN [1,5]

[1] Department of Orthopedic Surgery and Trauma, Máxima Medical Center, Eindhoven; [2] Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam Public Health (APH) Research Institute, Amsterdam; [3] Department of Orthopedic Surgery, Catharina Hospital Eindhoven, Eindhoven; [4] Department of General Practice, Erasmus MC, University Medical Center Rotterdam, Rotterdam; [5] Department of Orthopedic Surgery, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands
Correspondence: jaap.tolk@mmc.nl
Submitted 2018-04-12. Accepted 2018-10-11.

Background and purpose — Improvement of physical function is one of the main treatment goals in severe hip osteoarthritis (OA) patients. The Osteoarthritis Research Society International (OARSI) has identified a core set of performance-based tests to assess the construct physical function: 30-s chair stand test (30-s CST), 4x10-meter fast-paced walk test (40 m FPWT), and a stair-climb test. Despite this recommendation, available evidence on the measurement properties is limited. We evaluated the reliability, validity, and responsiveness of these performance-based measures in patients with hip OA scheduled for total hip arthroplasty (THA).

Patients and methods — Baseline and 12-month follow-up measurements were prospectively obtained in 90 end-stage hip OA patients who underwent THA. As there is no gold standard for comparison, the hypothesis testing method was used for construct validity and responsiveness analysis. A test can be assumed valid if ≥ 75% of predefined hypotheses are confirmed. A subgroup (n = 30) underwent test–retest measurements for reliability analysis. The Oxford Hip Score, Hip injury and Osteoarthritis Outcome Score—Physical Function Short Form, pain during activity score, and muscle strength were used as comparator instruments.

Results — Test–retest reliability was appropriate; intraclass correlation coefficient values exceeded 0.70 for all 3 tests. None of the performance-based measures reached 75% hypothesis confirmation for the construct validity or responsiveness analysis.

Interpretation — The performance-based tests have good reliability in the assessment of physical function. Construct validity and responsiveness, using patient-reported measures and muscle strength as comparator instruments, could not be confirmed. Therefore, our findings do not justify their use for clinical practice.

Improvement of physical function is one of the main treatment goals of total hip arthroplasty (THA). Physical function can be assessed using patient-reported and performance-based outcome measurement instruments (Reiman and Manske 2011). Because different domains of the construct physical function are measured, the methods are considered complementary and not competing (Stratford and Kennedy 2006, Reiman and Manske 2011, Dobson et al. 2013).

3 activities have been identified as most relevant for patients with hip OA: sit-to-stand movement, level walking, and stair negotiation (Dobson et al. 2013). Impairment on these domains is classified as "activity limitations" on the World Health Organization International Classification of Functioning, Disability and Health (ICF) (World Health Organization 2001). The Osteoarthritis Research Society International (OARSI) has identified a set of performance-based tests to assess the construct physical function (Dobson et al. 2012, 2013). The core set consists of the 30-s chair stand test (30-s CST) for assessment of sit-to-stand movement, 4x10 meter fast-paced walk test (40 m FPWT) for assessment of level walking, and a stair-climb test to assess stair negotiation (Dobson et al. 2013).

The validity and responsiveness of the OARSI core set have been challenged in knee OA patients (Tolk et al. 2017), but available evidence on the measurement properties in patients with hip OA is insufficient (Dobson et al. 2012, 2013). Measurement properties of a test should be confirmed in the population in which it is to be used, but the recommendation to use the specific tests included in the OARSI core set is based on expert opinion (Dobson et al. 2012, 2013). Therefore, before further implementation of the OARSI core set for hip OA patients can be considered, additional evidence on the measurement properties of these performance measures is essential (Terwee et al. 2006, Dobson et al. 2012). We evaluated the reliability, validity, and responsiveness after THA of the

DOI 10.1080/17453674.2018.1539567

OARSI recommended performance-based measures, for measurement of physical function in patients with severe hip OA.

Patients and methods

We performed a prospective cohort study of patients indicated for THA to evaluate the measurement properties of the 30-s CST, 40 m FPWT, and 10-step stair climb test (10-step SCT). The study was conducted following the COSMIN (COnsensus based Standards for the selection of health status Measurement INstruments) checklist (Mokkink et al. 2010b). The COSMIN checklist contains design requirements and preferred statistical methods for studies on measurement properties of health status measurement instruments.

Patient population

Patients were eligible for inclusion if they had unilateral symptomatic hip OA and were scheduled for primary THA. Patients with comorbidity leading to inability to perform the performance-based measures, insufficient knowledge of the Dutch language, and inability to visit follow-up appointments were excluded. All patients in the Máxima Medical Centre meeting these criteria, and willing to participate, signed an informed consent form. The number of patients needed for the analysis was guided by the COSMIN standards (Terwee et al. 2007, Mokkink et al. 2010b). We aimed to include ≥ 50 patients for construct validity and responsiveness analyses, and 30 patients for reliability analyses.

Study procedures

Patient characteristics measured at baseline were: sex, age, and BMI. The assessment of performance-based measures and comparator instruments described below was made at baseline before surgery, and 12 months after THA. The standardized testing procedures were performed by a research nurse strictly according to the manual provided by the OARSI, with a fixed order of tests (Dobson et al. 2013).

Performance-based measures

30-s CST. The 30-s CST aims to quantify a patient's performance on the activity "sit-to-stand movement" (Dobson et al. 2013). From a sitting position, the patient stands up until hips and knees are fully extended, then completely back down. This is repeated for 30 seconds and each full cycle is counted as 1 chair stand (Dobson et al. 2013). A 43-cm high, straight-back chair without armrests was used. For patients with hip OA, good reliability is reported with an intraclass correlation coefficient (ICC) of 0.81 (0.63–0.91) and standard error of measurement (SEM) of 1.27 (Wright et al. 2011). No reports on construct validity are available.

40 m FPWT. The 40 m FPWT is a test for performance on the activity short-distance walking (Dobson et al. 2013). Participants are asked to walk as quickly but as safely as possible, without running, along a 10-meter walkway for a total distance of 40 meters. Walking speed is measured in meters/second (m/s). Use of a walking aid is allowed and recorded. Inter-rater reliability is reported to be good in patients with hip OA, with an ICC of 0.95 (0.90–0.98) and SEM of 1.0 m/s (Wright et al. 2011). There are no reports available on the construct validity.

Stair climb test. The OARSI included a stair-climb test in the core set, but no specific measure is recommended (Dobson et al. 2013). We selected the 10-step stair climb test (10-step SCT), as the stair in the testing area had 10 steps with a step height of 19 cm. Patients were instructed to ascend and descend the flight of stairs as quickly as possible but in a safe manner. The time needed is recorded in seconds (Dobson et al. 2013). To our knowledge, there is no evidence available on measurement properties of the 10-step stair-climb test or comparable stair-climb tests in patients with hip OA.

Comparator instruments

We used a combination of comparator instruments; a specification of these instruments and their measurement properties can be found in a supplementary file. For measurement of physical function 2 joint-specific PROMs were used: the Hip injury and Osteoarthritis Outcome Score—Physical Function Short Form (HOOS—PS) (Davis et al. 2009), and the Oxford Hip Score (OHS) (Dawson et al. 1996). The EuroQol 5D-3L (EQ-5D) was used as a measure of health-related quality of life (Rabin and de Charro 2001). Pain during activity was scored from 0 to 10 using a numerical rating scale (NRS pain) (Ruyssen-Witrand et al. 2011). At 12 months follow-up a 7-point Likert scale anchor question was scored for change in activities of daily living. Preoperatively knee extensor and hip abductor strength of the affected leg was measured using a handheld dynamometer (Holstege et al. 2011, Zeni et al. 2014).

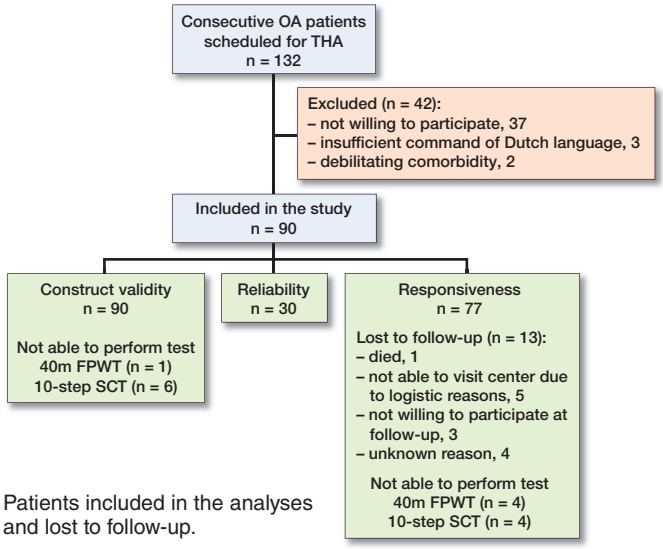
Evaluation of the measurement properties and statistics

Reliability

Test–retest reliability refers to the extent to which scores for patients who have not changed are the same for repeated measurement over time. For this analysis, test–retest measurements of the 3 performance-based measures were obtained in a subset of the study population. 30 minutes of rest were allowed in between, to allow for full recovery during the resting interval. Performance on the activity under study can assumed to be stable over this testing period. ICC values for absolute agreement with corresponding 95% confidence intervals (CI) were calculated using a 2-way random model with absolute agreement. The threshold for an appropriate ICC is 0.70 (Terwee et al. 2007, Prinsen et al. 2016). SEM and SDC were calculated as described by Atkinson (1998).

Construct validity

Construct validity refers to the degree to which the instruments under study measure the construct they aim to mea-



Patients included in the analyses and lost to follow-up.

sure. This is the recommended method to assess validity when there is no "Gold Standard" available, as is the case for the functional domains level walking, stair negotiation, and sit-to-stand movement in hip OA. Before the start of the study, an expert panel formulated hypotheses on the expected relationships of performance-based measure scores with scores on the comparative instruments (Table 3, see Supplementary data) (Mokkink et al. 2010a, de Vet et al. 2011). Direction and magnitude of the expected results were stated. The expert panel consisted of an orthopedic surgeon (RJ), orthopedic resident and PhD candidate (JT), specialist in measurement property analysis (CP), and a methodologist (MR).

The hypotheses were based on the following predictions—we expected: a moderate correlation of the performance-based measures with PROMs and quadriceps strength; a stronger correlation of PROMs with pain scores than with the performance-based measures; a stronger correlation of the performance-based measures with PROMs measuring functional outcome than with a PROM measuring general health; a stronger correlation of specific questions of the PROMs regarding walking, stair negotiation, and sit-to-stand movement to their respective performance-based measure than to the total PROM score. Correlations on a convergent hypothesis were expected to be at least moderate: ≥ 0.4 or ≤ –0.4. Divergent hypotheses were expected to have a poor correlation (≥ –0.39; ≤ 0.39). Pearson or Spearman correlation coefficients were calculated, depending on normality of data distribution. Construct validity can be assumed adequate if at least 75% of the predefined hypotheses are confirmed (Terwee et al. 2007).

Responsiveness

Responsiveness refers to the ability of the instruments to detect change over time in the construct measured. In the absence of a gold standard, a construct approach is to be used. Hypotheses were formulated a priori by the expert panel, in a similar manner to the construct validity analysis (Table 5) (Terwee et al. 2007, Mokkink et al. 2010a, de Vet et al. 2011).

The hypotheses were formulated according to the following criteria: the anchor question would be moderately correlated to change in the performance-based measures scores (≥ 0.4 or ≤ –0.4) and the change in PROMs would be more correlated to pain than to change in the performance-based measure scores. Pearson or Spearman correlation coefficients were calculated, depending on normality of data distribution. Adequate responsiveness can be assumed if minimally 75% of the predefined hypotheses are confirmed (Terwee et al. 2007).

SPSS statistics version 24.0 was used for the analyses (IBM Corp, Armonk, NY, USA).

Table 1. Patient characteristics. Data are mean (SD) unless otherwise stated

	Total cohort (n = 90)	Reliability analysis cohort (n = 30)
Age, years	69 (9.5)	66 (9.4)
Women, n	61	22
BMI	27 (3.9)	26 (2.7)
Hip abductor strength, N	196 (7.8)	219 (7.9)
Knee extensor strength, N	134 (5.7)	13 (4.3)

Ethics, funding, and potential conflicts of interest

The Máxima Medical Centre Medical Ethics Committee approved the study (registration code 2014-73). No funding was received for the present study. The authors declare that there are no conflicts of interest related to this article.

Results

Patient characteristics

In the period April to October 2015, 90 consecutive patients scheduled for arthroplasty because of hip OA were recruited (Table 1, Figure).

Measurement properties

Reliability analysis

30 randomly selected patients were enrolled in the test–retest study. Test–retest reliability was appropriate; ICC values exceeded 0.70 for all 3 tests (Table 2, see Supplementary data).

Construct validity (hypothesis testing)

None of the 3 performance-based measures reached confirmation of 75% or more of the predefined hypotheses. 4/9 were confirmed for the 30-s CST, 6/17 for the 40m FPWT, and 6/17 for the 10-step SCT (Table 3, see Supplementary data).

Responsiveness

The mean score on the anchor question for change in activities of daily living (7-point Likert scale) at 12-month follow-

Table 5. Responsiveness

Predefined hypotheses	30-s chair stand test (change score) Spearman correlation coefficient	Hypothesis confirmed	40 m fast-paced walk test (change score) Spearman correlation coefficient	Hypothesis confirmed	10-step stair climb test (change score) Spearman correlation coefficient [a]	Hypothesis confirmed
1. Moderate correlation with anchor question (≥ 0.4)	0.37	No	0.28	No	–0.18	No
2. Moderate correlation with change score NRS pain during activity (≤ –0.4)	–0.04	No	–0.13	No	0.14	No
3. Moderate correlation with change score HOOS-PS (≤ –0.4)	0.30	No	0.21	No	–0.35	No
4. Moderate correlation with change OHS (≥ 0.4)	0.23	No	0.27	No	–0.26	No
5. Correlation between change scores NRS pain and HOOS-PS is minimal 0.1 stronger than between NRS pain and performance-based test	–0.45/–0.04	Yes	–0.45/–0.13	Yes	–0.45/–0.18	Yes
6. Correlation between change scores NRS pain and HOOS-PS is minimal 0.1 stronger than between HOOS-PS and performance-based test	–0.45/0.30	Yes	–0.45/0.21	Yes	–0.45/–0.35	Yes
7. Correlation between changes scores NRS pain and OHS minimal 0.1 stronger than between NRS pain and performance-based test	–0.66/–0.04	Yes	–0.66/–0.13	Yes	–0.66/–0.18	Yes
8. Correlation between change scores NRS pain and OHS is minimal 0.1 stronger than between OHS and performance-based test	–0.66/0.23	Yes	–0.66/0.27	Yes	–0.66/–0.26	Yes
Hypothesis confirmed	4/8		4/8		4/8	

up was 6.2 (5.9–6.4), which represents "much improvement." Results of the responsiveness analysis are presented in Table 5. For the 30-s CST, 4/8 of the hypothesis were confirmed, for the 40m FPWT 4/8, and for the 10-step SCT 4/8 (Table 4, see Supplementary data).

Discussion

To our knowledge, this is the first thorough assessment of the measurement properties of the OARSI-recommended core set of performance-based measures in patients with severe hip OA. The reliability analysis showed excellent test–retest reliability, which is in line with previous reports (Wright et al. 2011, Dobson et al. 2017). Construct validity and responsiveness could not be confirmed. These findings are in accordance with recently published work on the OARSI core set of performance-based measures in knee OA patients (Tolk et al. 2017).

All 3 performance-based measures scored poorly on the construct validity and responsiveness analysis. One of the reasons is that almost all convergent hypotheses with PROMs measuring physical function were rejected. Although both methods aim to quantify related constructs, previous research has shown that PROMs assessing physical function do not measure the exact same domain as performance-based measures (Stratford and Kennedy 2006, Reiman and Manske 2011, Dobson et al. 2013). This potentially limits the strength of the conclusions that can be drawn from the present study. For example, PROMs are known to have a higher dependency on pain scores than performance based-measures (Stratford and Kennedy 2006). When—in the absence of a gold standard—the construct approach is to be used, it is inherently so that there is a discrepancy between the test under study and the comparator instruments (de Vet et al. 2011). Furthermore, PROMs were not the only comparative instruments used, and hypotheses predicting a higher correlation of the performance-based measure scores with related construct compared with less related constructs were largely rejected as well. Therefore, in our opinion, the conclusion on the construct validity and responsiveness should be interpreted more broadly than only showing the known discrepancy between PROMs and these measures.

As an alternative to the comparator instruments used for construct validity and responsiveness in the present study, 3-D motion analysis or inertia-based motion analysis could be used. These methods allow for a kinematic analysis in patients with hip OA, but their clinical relevance has not been defined (Kolk et al. 2014, Bolink et al. 2016). Therefore, we believe these alternative methods are not suitable for comparison purposes in a clinical perspective. The comparative instruments used in the present study were considered the most suitable instruments available.

The findings on construct validity of the performance-based measures might be affected because impairment on the tested activities in daily living is not fully appreciated by merely timing the performance (Steultjens et al. 1999, Stratford and Kennedy 2006). Although others claim good face validity for the core set of performance-based measures (Dobson et al. 2013, 2017), in our view this is not straightforward. For example, standing up and sitting down in rapid sequence, as measured by the 30-s CST, is not really exemplary for stand-to-sit movement in daily life. Fewer repetitions on the test does not necessarily mean the quality of a sit-to-stand move-

ment in daily living is more or less impaired. The same goes for walking speed and stair ascent, which does not directly represent more or less impairment. Merely timing the activity or counting repetitions cannot capture impairment caused by limping or joint instability, nor avoidance of an activity in daily living (Steultjens et al. 1999, Holla et al. 2014). This is a possible explanation as to why the construct validity could not be confirmed.

The responsiveness analysis showed that change in pain scores was strongly correlated to change in PROM scores, but not related to performance-based measure scores. Others have presented this low correlation with pain scores as a strength of performance-based measures, claiming this makes them more "objective" (Dobson et al. 2012, 2013). In our opinion, it seems unlikely that the degree of pain during an activity would not influence performance in daily living (Holla et al. 2014). Furthermore, it has been shown that pain during activity does affect the quality of movement, and impaired quality of movement is associated with lower perceived physical function (Steultjens et al. 1999, Rosenlund et al. 2016). Although pain reduction is not related to an increase in speed on the tested activities, the quality and manner of performance might improve (Steultjens et al. 1999), and patients might no longer avoid the activities (Holla et al. 2014). These factors of physical performance are not grasped by the performance-based measures under study. The number of repetitions or speed scored on the performance-based measures might be of interest for research purposes, but tin the authors' opinion actual change and perceived change need to be related to some degree for a test to be clinically relevant. Hypotheses in this regard were all rejected, contributing to the negative conclusion on the responsiveness of the OARSI core set of performance-based measures.

The strict adherence to the methodological criteria provided by COSMIN is a strength of the present study (Mokkink et al. 2010b). Most previous reports on the measurement properties of the performance-based measures under study reported combined groups of hip and knee OA patients, resulting in heterogeneous populations (Kennedy et al. 2005, Gill and McBurney 2008, Dobson et al. 2017). The present study reports on an unselected, consecutive group of only end-stage hip OA patients. The results can therefore be considered more accurate and representative for this population.

The group size for test–retest measurements was kept relatively small, to reduce the burden of repeated measurements for patients. As there is evidence from other studies showing similar results on reliability (Kennedy et al. 2005, Wright et al. 2011, Dobson et al. 2017), in our view it can be concluded that the performance-based measures under study have adequate test–retest reliability. The percentage of patients lost to follow-up for the responsiveness analysis was 14%. In our opinion, this can be considered acceptable, especially as the group of patients with incomplete data did not show systematic difference in baseline characteristics (Table 1).

In summary, the 30-s CST, 40 m FPWT, and 10-step SCT have good reliability in the assessment of the domains sit-to-stand movement, walking short distances, and stair negotiation in the construct physical function. Construct validity and responsiveness, using patient-reported measures and muscle strength as comparator instruments, could not be confirmed. Therefore, the present study does not justify their use for clinical practice in patients with severe hip OA.

Supplementary data

Tables 2–4 and a specification of comparator instruments used are available as supplementary data in the online version of this article, http://dx.doi.org/ 10.1080/17453674.2018.1539567

JT and MR contributed to the conception and design of the study and drafting of the article. CP provided methodological support. All authors contributed to interpretation of the data and critically revised the article.

The authors would like to sincerely thank C. van Doesburg, H. Kox, D. Latijnhouwers, and M. Mariam for their work in administrative and testing procedures.

Acta thanks Margareta Hedstrom and Anders Holsgaard-Larsenfor help with peer review of this study.

Atkinson G N A. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sport Med 1998; 26(4): 21-38.

Bolink S A A N, Lenguerrand E, Brunton L R, Wylde V, Gooberman-Hill R, Heyligers I C, Blom A W, Grimm B. Assessment of physical function following total hip arthroplasty: inertial sensor based gait analysis is supplementary to patient-reported outcome measures. Clin Biomech 2016; 32: 171-9.

Davis A M, Perruccio A V, Canizares M, Hawker G A, Roos E M, Maillefert J-F, Lohmander L S. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthritis Cartilage 2009; 17(7): 843-7.

Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. Bone Joint Surg (Br) 1996; 78(2): 185-90.

de Vet H C W, Terwee C B, Mokkink L B, Knol D L. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.

Dobson F, Hinman R S, Hall M, Terwee C B, Roos E M, Bennell K L. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthritis Cartilage 2012; 20(12): 1548-62.

Dobson F, Hinman R S, Roos E M, Abbott J H, Stratford P, Davis A M, Buchbinder R, Snyder-Mackler L, Henrotin Y, Thumboo J, Hansen P, Bennell K L. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthritis Cartilage 2013; 21(8): 1042-52.

Dobson F, Hinman R S, Hall M, Marshall C J, Sayer T, Anderson C, Newcomb N, Stratford P W, Bennell K L. Reliability and measurement error of the Osteoarthritis Research Society International (OARSI) recommended performance-based tests of physical function in people with hip and knee osteoarthritis. Osteoarthritis Cartilage 2017; 6-10.

Gill S, McBurney H. Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. Physiother Res Int 2008; 13(3): 141-52.

Holla J F M, Sanchez-Ramirez D C, van der Leeden M, Ket J C F, Roorda L D, Lems W F, Steultjens M P M, Dekker J. The avoidance model in knee and hip osteoarthritis: a systematic review of the evidence. J Behav Med 2014; 37(6): 1226-41.

Holstege M S, Lindeboom R, Lucas C. Preoperative quadriceps strength as a predictor for short-term functional outcome after total hip replacement. Arch Phys Med Rehabil 2011; 92(2): 236-41.

Kennedy D M, Stratford P W, Wessel J, Gollish J D, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord 2005; 6: 3.

Kolk S, Minten M J M, Van Bon G E A, Rijnen W H, Geurts A C H, Verdonschot N, Weerdesteyn V. Gait and gait-related activities of daily living after total hip arthroplasty: a systematic review. Clin Biomech 2014; 29(6): 705-18.

Mokkink L B, Terwee C B, Knol D L, Stratford P W, Alonso J, Patrick D L, Bouter L M, de Vet H C. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010a; 10: 22.

Mokkink L B, Terwee C B, Patrick D L, Alonso J, Stratford P W, Knol D L, Bouter L M, De Vet H C W. 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 2010b; 19(4): 539-49.

Prinsen C A C, Vohra S, Rose M R, Boers M, Tugwell P, Clarke M, Williamson P R, Terwee C B, 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. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set": a practical guideline. Trials 2016; 17(1): 449.

Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001; 33(5): 337-43.

Reiman M P, Manske R C. The assessment of function: how is it measured? A clinical perspective. J Man Manip Ther 2011; 19(2): 91-9.

Rosenlund S, Holsgaard-Larsen A, Overgaard S, Jensen C. The Gait Deviation Index is associated with hip muscle strength and patient-reported outcome in patients with severe hip osteoarthritis: a cross-sectional study. PLoS One 2016; 11(4): 1-13.

Ruyssen-Witrand A, Fernandez-Lopez C J, Gossec L, Anract P, Courpied J P, Dougados M. Psychometric properties of the OARSI/OMERACT osteoarthritis pain and functional impairment scales: ICOAP, KOOS-PS and HOOS-PS. Clin Exp Rheumatol 2011; 29(2): 231-7.

Steultjens M P, Roorda L D, Dekker J, Bijlsma J W. Responsiveness of observational and self-report methods for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 2001; 45(15): 56-61.

Steultjens M P M, Dekker J, van Baar M E, Oostendorp R B, Bijlsma J W J. Internal consistency and validity of an observational method for assessing disability in mobility in patients with osteoarthritis. Arthritis Rheum 1999; 12(1): 19-25.

Stratford P W, Kennedy D M. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 2006; 59(2): 160-7.

Terwee C B, Mokkink L B, Steultjens M P M, Dekker J. 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) 2006; 45(7): 890-902.

Terwee C B, Bot S D M, de Boer M R, van der Windt D A, Knol D L, Dekker J, Bouter L M, de Vet H C W. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007; 60: 34-42.

Tolk J J, Janssen R P A, Prinsen C A C, Latijnhouwers D A J M, van der Steen M C, Bierma-Zeinstra S M A, Reijman M. The OARSI core set of performance-based measures for knee osteoarthritis is reliable but not valid and responsive. Knee Surgery, Sport Traumatol Arthrosc 2017; epub ahead of print.

World Health Organization. International Classification of Functioning, Disability and Health. Geneva: WHO; 2001.

Wright A A, Cook C E, Baxter G D, Dockerty J D, Abbott J H. A comparison of 3 methodological approaches to defining major clinically important improvement of 4 performance measures in patients with hip osteoarthritis. J Orthop Sports Phys Ther 2011; 41(5): 319-27.

Zeni J, Abujaber S, Pozzi F, Raisis L. Relationship between strength, pain, and different measures of functional ability in patients with end-stage hip osteoarthritis. Arthritis Care Res (Hoboken) 2014; 66(10): 1506-12.

Supplementary data

Supplementary file – specification of comparator instruments used

Comparator instruments

HOOS-PS

The Hip injury and Osteoarthritis Outcome Score - Physical Function Short Form (HOOS-PS) is a 5-item PROM for measurement of the construct physical function. The HOOS-PS is scored on a 0 to 100 scale, 0 indicating no symptoms and 100 indicating extreme symptoms (Davis et al. 2009). The HOOS-PS has good construct validity and responsiveness in hip OA patients (Davis et al. 2009).

OHS

The Oxford Hip Score (OHS) is a 12-item disease specific PROM for measurement of pain and function of the hip in relation to different activities of daily life. The total score ranges from 12 indicating no difficulties symptoms to 60 indicating most difficulties (Dawson et al. 1996). The OHS has shown to be consistent, reliable, valid and sensitive to clinical change (Dawson et al. 1996, Gosens et al. 2005).

EQ-5D

EuroQol 5D-3L (EQ-5D) is a standardized instrument developed as a measure of health-related quality of life (Rabin and de Charro 2001). This PROM consists of a 5-question descriptive part and a visual analogue scale score (EQ-VAS) ranging from 0 to 100 (Rabin and de Charro 2001). From the 5-question part a sum score can be calculated, where 1 represents the best possible health state and lower scores represent worse health state (Rabin and de Charro 2001). The EQ-5D has shown to be valid and reliable in hip OA patients (Conner-Spady et al. 2015).

NRS pain

Pain during activity was scores using a numerical rating scale (NRS pain). Patients were asked to score pain during activity in the past week on an 11-point scale, the patients rate their pain during activity from 0 to 10. A score of 0 represented 'no pain' and a score of 10 represented 'worst imaginable pain'. Good reliability and responsiveness are reported for this NRS pain scale (Ruyssen-Witrand et al. 2011).

Anchor question

At 12-months follow-up a 7-point Likert scale anchor question was scored for change in activities of daily living. The question 'how has your general daily functioning changed since the operation on your knee?' was scored from 1 (a lot worse) to 7 (very much improved).

Muscle strength

Strength of the knee extensors and hip abductors of the affected leg were tested for all subjects in the study. Maximal isometric knee extensor strength was measured in Newton (N) using a handheld dynamometer (HHD). In an upright sitting position, the HHD was positioned on the anterior aspect of the tibia, five cm proximal to the medial malleolus. A protective shin guard was used for patient comfort as well as standardization of HHD placement. Hip abductor strength was measured with subjects in supine position and with 5° of hip abduction. The HHD was positioned on the lateral femoral condyle and its position was held constant between trials to avoid changes in the resistance moment arm. For both muscle groups three consecutive measurements were obtained, the highest value was used for analysis. The HHD is a widely used instrument to measure knee extensor and hip abductor strength, with good reliability in OA patients. An ICC of 0.94–0.97 is reported (Holstege et al. 2011, Zeni et al. 2014).

Conner-Spady B L, Marshall D A, Bohm E, Dunbar M J, Loucks L, Khudairy A Al, Noseworthy T W. 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 2015; 24(7): 1775-84.

Davis A M, Perruccio A V, Canizares M, Hawker G A, Roos E M, Maillefert J-F, Lohmander L S. Comparative, validity and responsiveness of the HOOS-PS and KOOS-PS to the WOMAC physical function subscale in total joint replacement for osteoarthritis. Osteoarthritis Cartilage 2009; 17(7): 843-7.

Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. J Bone Joint Surg Br 1996; 78(2): 185-90.

Gosens T, Hoefnagels N, de Vet R, Dhert W, van Langelaan E, Bulstra S, Geesink R. The 'Oxford Heup Score': the translation and validation of a questionnaire into Dutch to evaluate the results of total hip arthroplasty. Acta Orthop 2005; 76(2): 204-11.

Holstege M S, Lindeboom R, Lucas C. Preoperative quadriceps strength as a predictor for short-term functional outcome after total hip replacement. Arch Phys Med Rehabil 2011; 92(2): 236-41.

Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001; 33(5): 337-43.

Ruyssen-Witrand A, Fernandez-Lopez C J, Gossec L, Anract P, Courpied J P, Dougados M. Psychometric properties of the OARSI/OMERACT osteoarthritis pain and functional impairment scales: ICOAP, KOOS-PS and HOOS-PS. Clin Exp Rheumatol 2011; 29(2): 231-7.

Zeni J, Abujaber S, Pozzi F, Raisis L. Relationship between strength, pain, and different measures of functional ability in patients with end-stage hip osteoarthritis. Arthritis Care Res (Hoboken) 2014; 66(10): 1506–12.

Table 2. Reliability analysis (n = 30)

	Score baseline mean (95% CI)	Retest score mean (95% CI)	Difference baseline–retest score mean (95% CI)	ICC (95% CI)	SEM	SDC
30-s CST (stands)	10.1 (9.0–11.2)	10.9 (9.7–12.1)	–0.8 (–0.3 to –1.4)	0.86 (0.66–0.94)	0.99	2.7
40 m FPWT (m/s)	1.32 (1.22–1.43)	1.33 (1.20–1.46)	–0.01 (–0.05 to 0.04)	0.94 (0.88–0.97)	0.08	0.22
10-step SCT (s)	14.2 (12.3–16.0)	14.1 (12.3–15.9)	–0.1 (–0.5 to 0.6)	0.96 (0.91–0.98)	1.06	2.9

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

Table 3. Construct validity

	30-s chair stand test		40 m fast-paced walk test		10-step stair climb test	
Predefined hypotheses	Spearman correlation coefficient	Hypothesis confirmed	Spearman correlation coefficient	Hypothesis confirmed	Spearman correlation coefficient [a]	Hypothesis confirmed
1. Moderate correlation with HOOS-PS (≤ –0.4) *	0.21	No	0.21	No	–0.24	No
2. Moderate correlation with OHS (≥ 0.4) *	0.45	Yes	0.34	No	–0.27	No
3. Moderate correlation with hip abductor strength (≥ 0.4) *	0.21	No	0.48	Yes	–0.44	Yes
4. Moderate correlation with quadriceps strength (≥ 0.4) *	0.35	No	0.46	Yes	–0.53	Yes
5. Unrelated to EQ-5D (–0.39; 0.39)	0.38	Yes	0.31	Yes	0.34	Yes
6. Correlation with HOOS-PS is minimal 0.1 stronger than with EQ-5D	0.21/0.38	No	0.21/0.31	No	–0.24/0.34	No
7. Correlation with OHS is minimal 0.1 stronger than with EQ-5D	0.45/0.38	No	0.34/0.21	Yes	–0.27/0.34	No
8. "Absolute" correlation between NRS pain and HOOS-PS is minimal 0.1 higher than between performance-based measure and NRS pain	–0.53/–0.19	Yes	–0.53/–0.12	Yes	–0.53/0.02	Yes
9. "Absolute" correlation between NRS pain and OHS is minimal 0.1 higher than perfor-mance-based measure and NRS pain	–0.63/–0.19	Yes	–0.63/–0.12	Yes	–0.63/0.02	Yes
10. "Absolute" correlation 40 m FPWT with HOOS-PS Question 4 is minimal 0.1 stronger than with HOOS-PS	NA		–0.12/0.21	No	NA	
11. "Absolute" 40 m FPWT with HOOS-PS Question 4 is minimal 0.1 stronger than with OHS	NA		–0.12/0.34	No	NA	
12. "Absolute" correlation 40 m FPWT with HOOS-PS Question 4 is minimal 0.1 higher than with EQ-5D Score	NA		–0.12/0.31	No	NA	
13. "Absolute" correlation 40 m FPWT with OHS Question 4 is minimal 0.1 stronger than with HOOS-PS	NA		–0.12/0.21	No	NA	
14. "Absolute" correlation 40 m FPWT with OHS Question 4 is minimal 0.1 stronger than with OHS	NA		–0.12/0.34	No	NA	
15. "Absolute" correlation 40 m FPWT with OHS Question 4 is minimal 0.1 stronger than with EQ-5D Score	NA		–0.12/0.31	No	NA	
16. Moderate correlation 40 m FPWT with EQ-5D Question 1 (≤ –0.4)	NA		–0.36	No	NA	
17. Moderate correlation 40 m FPWT with OHS Question 4 (≤ –0.4)	NA		–0.12	No	NA	
18. "Absolute" correlation 10-step SCT with OHS Question 6 is minimal 0.1 stronger than with HOOS-PS	NA		NA		0.31/–0.24	No
19. "Absolute" correlation 10-step SCT with OHS Question 6 is minimal 0.1 stronger than with OHS	NA		NA		0.31/–0.27	No
20. "Absolute" correlation 10-step SCT with OHS Question 6 is minimal 0.1 stronger than with EQ-5D	NA		NA		0.31/–0.31	No
21. Moderate correlation 10-step SCT with OHS Question 6 (≤ –0.4)	NA		NA		0.31	No
22. "Absolute" correlation 10-step SCT with HOOS-PS Question 1 is minimal 0.1 stronger than with HOOS-PS	NA		NA		0.34/–0.24	Yes
23. "Absolute" correlation 10-step SCT with HOOS-PS question 1 is minimal 0.1 stronger than with OHS	NA		NA		0.34/–0.27	No
24. "Absolute" correlation 10-step SCT with HOOS-PS question 1 is minimal 0.1 stronger than with EQ-5D	NA		NA		0.34/–0.31	No
25. Moderate correlation 10-step SCT with HOOS-PS question 1 (≤ –0.4)	NA		NA		0.34	No
Hypothesis confirmed	4/9		6/17		6/17	

NA = not applicable.
[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 hypothesized correlations are in the opposite direction.

Table 4. Performance-based measures and PROM scores before and after THA. Data are mean (SD) unless otherwise stated

Item	Baseline	12-month follow-up	p-value
30-s CST (stands)	9.3 (8.5–10.2)	12.0 (11.2–12.9)	< 0.001
40 m FPWT (m/s)	1.26 (1.17–1.34)	1.34 (1.26–1.42)	< 0.001
Use of assistive device during 40m FPWT (patients, n)	8	2	0.057
10-step SCT (seconds)	17.9 (15.3–20.4)	14.5 (12.9–16.2)	< 0.001
Use of handrail during 10-step SCT (patients, n)	41	28	0.047
HOOS-PS score	48.0 (44.3–51.9)	21.7 (19.8–26.2)	< 0.001
OHS	23.6 (21.9–25.7)	41.8 (40.5–43.2)	< 0.001
EQ-5D	0.51 (0.43–0.57)	0.83 (0.79–0.86)	< 0.001
EQ-VAS	64.8 (59.6–70.0)	76.1 (71.5–80.7)	0.001
NRS pain	6.8 (6.5–7.3)	1.5 (1.7)	< 0.001
Anchor question (patients, n)			
Very much improvement		34	
Much improvement		33	
A little improvement		5	
Unchanged		1	
A little worse		0	
Much worse		4	
Very much worse		0	