Osteoarthritis and Cartilage 20 (2012) 1548–1562

Osteoarthritis and Cartilage

Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review

F. Dobson †*, R.S. Hinman †, M. Hall †, C.B. Terwee ‡, E.M. Roos §, K.L. Bennell †

† Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne, Australia
‡ VU University Medical Center, Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, The Netherlands
§ Institute of Sports Science and Clinical Biomechanics, University of Southern Denmark, Denmark

ARTICLE INFO

Article history:
Received 19 April 2012
Accepted 22 August 2012

Keywords:
Performance-based measures
Physical function
Measurement properties
Clinimetrics
Systematic review
Osteoarthritis

SUMMARY

Objectives: To systematically review the measurement properties of performance-based measures to assess physical function in people with hip and/or knee osteoarthritis (OA).
Methods: Electronic searches were performed in MEDLINE, CINAHL, Embase, and PsycINFO up to the end of June 2012. Two reviewers independently rated measurement properties using the consensus-based standards for the selection of health status measurement instrument (COSMIN). "Best evidence synthesis" was made using COSMIN outcomes and the quality of findings.
Results: Twenty-four out of 1792 publications were eligible for inclusion. Twenty-one performance-based measures were evaluated including 15 single-activity measures and six multi-activity measures. Measurement properties evaluated included internal consistency (three measures), reliability (16 measures), measurement error (14 measures), validity (nine measures), responsiveness (12 measures) and interpretability (three measures). A positive rating was given to only 16% of possible measurement ratings. Evidence for the majority of measurement properties of tests reported in the review has yet to be determined. On balance of the limited evidence, the 40 m self-paced test was the best rated walk test, the 30 s-chair stand test and timed up and go test were the best rated sit to stand tests, and the Stratford battery, Physical Activity Restrictions and Functional Assessment System were the best rated multi-activity measures.
Conclusion: Further good quality research investigating measurement properties of performance measures, including responsiveness and interpretability in people with hip and/or knee OA, is needed. Consensus on which combination of measures will best assess physical function in people with hip/and or knee OA is urgently required.

Crown Copyright © 2012 Published by Elsevier Ltd on behalf of Osteoarthritis Research Society International. All rights reserved.

Introduction

Measurement of treatment outcomes and change in health status over time is a critical component of research and clinical practice for people with osteoarthritis (OA). The Osteoarthritis Research Society International (OARSI) and Outcome Measures in Rheumatology and Clinical Trials (OMERACT) jointly advocate the use of core outcome measures for clinical trials of OA that address the domains of pain and function[1]. Currently there is no singular gold standard for the assessment of physical function. Physical function is related to "the ability to move around"[2] and "the ability to perform daily activities"[3] and can be classified as *Activities* using the World Health Organization International Classification of Functioning, Disability and Health (ICF) model[4].

Measurement of physical function is complex as it contains multi-dimensional constructs[3,5]. A range of both self-report and performance-based measures have been used to assess physical function. Performance-based measures are defined as assessor-observed measures of tasks classified as "activities" using the ICF model[4] and are usually assessed by timing, counting or distance methods. They are not specific to body structure, body function or impairments such as measures of muscle strength or range of motion. Performance-based measures assess what an individual can do rather than what the individual perceives they can do, which is determined by self-report measures[3]. Increasing evidence

* Address correspondence and reprint requests to: F. Dobson, Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne, 200 Berkeley Street, Victoria 3010, Australia. Tel: 61-3-8344-3642; Fax: 61-3-8344-3771.
E-mail addresses: fdobson@unimelb.edu.au (F. Dobson), ranash@unimelb.edu.au (R.S. Hinman), halm@unimelb.edu.au (M. Hall), cb.terwee@vumc.nl (C.B. Terwee), eroos@health.sdu.dk (E.M. Roos), k.bennell@unimelb.edu.au (K.L. Bennell).

1063-4584/$ – see front matter Crown Copyright © 2012 Published by Elsevier Ltd on behalf of Osteoarthritis Research Society International. All rights reserved.
http://dx.doi.org/10.1016/j.joca.2012.08.015

suggests that performance-based measures capture a different construct of function and are more likely to fully characterize a change in body function than self-reported measures alone[6–8]. Both types of measures are now seen as complementary rather than competing when evaluating functional outcomes in people with OA[5,9,10].

A previous systematic review of performance-based measures in OA concluded that better designed studies assessing the measurement properties of these measures in OA populations were required[3]. Also, only a small percentage (7%) of measurement properties were rated as 'positive' for the quality of the findings and the levels of evidence were generally unknown or very limited. This previous review evaluated studies published up until early 2004 and since then further studies have been published. In addition, a new quality evaluation tool, the consensus-based standards for the selection of health status measurement instruments (COSMIN)[11,12] and scoring system[13], has been developed to standardize the assessment of methodological quality of measurement studies.

The aim of this study was to systematically review the measurement properties of performance-based tests to measure physical function in people with hip and/or knee OA using a robust quality evaluation tool and scoring system (COSMIN). Such a review would be a useful and timely update for researchers and clinicians to assist them in selecting appropriate clinical performance-based measures for people with hip and knee OA.

Methodology

Literature search

The search strategy was developed, reviewed and refined by multiple authors, in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[14]. Electronic searches of entire databases up until June 2012 were performed using MEDLINE *via* PubMed, CINAHL *via* EBSCO, Embase *via* Elsevier, and PsycINFO *via* CSA. Key search terms and synonyms were searched separately in four main filters which were then combined. These filters are summarized as:

1. Construct: physical function OR physical performance OR physical activity
2. Target population: Hip OR knee OR lower-limb AND osteoarthritis OR arthritis OR OA OR replacement OR arthroplasty
3. Measurement instrument: performance test/measure/instrument/assessment/index OR objective test/measure/assessment/OR observational test/measure/assessment/index OR task performance and analysis
4. Measurement properties: instrument development OR psychometrics OR clinimetrics OR validity OR reliability OR responsiveness OR interpretability OR meaningful change.

The search strategy was based on recommendations for performing systematic reviews of measurement properties[15] and is more fully described in Appendix 1. For MEDLINE (PubMed), we adopted a measurement properties search filter shown to retrieve more than 97% of publications related to measurement properties[16]. Targeted hand-searching of reference lists was also performed.

Eligibility criteria

Studies were screened by two independent reviewers (FD and MH). This included independent screening of the titles and abstracts from all retrieved studies followed by independent full-text review of potentially eligible studies. Any disagreements were discussed and resolved with a third reviewer (CT). Studies were included if they met the following criteria:

1. **Construct**: The test was a measure of physical function, defined according to the ICF model as *Activities*, which relate to the ability to move around and perform daily activities[4]. If the test was a battery of multi-task items, then at least 80% of the items were required to assess activities.
2. **Target population**: The study population comprised at least 80% of people diagnosed with symptomatic hip or knee OA using clinical or radiographic criteria. This could include all stages of disease as well as individuals who had recently undergone a specific intervention such as joint arthroplasty or an exercise program, where measures pre-intervention were provided.
3. **Measurement instrument**: The measure under study should be a performance-based measure which is evaluated by an observer as the activity is being performed by the individual, usually by timing, counting or distance methods.
4. **Setting**: The measure was conducted within the clinic or field and required non-technical, readily available, inexpensive and portable equipment.
5. **Measurement properties**: The study aim was to evaluate one or more measurement properties (e.g., internal consistency, reliability, validity, responsiveness and/or interpretability).
6. **Full-text** studies published as original articles.

Studies were excluded if: (1) the focus was on validating self-reported measures of function; (2) the measure predominately targeted the ICF level of impairment or health related quality of life; (3) treatment effectiveness was evaluated without a specific aim to study the measurement properties of performance measures; (4) the measure required expensive sophisticated equipment such as three-dimensional gait analysis or accelerometers; (5) they were published only as 'grey literature' such as scientific meeting abstracts, dissertations or unpublished literature; and (6) they were published in languages other than English due to limited language translational ability.

Methodological quality evaluation of the studies

The COSMIN tool was used to evaluate the methodological quality of included studies[11,17]. Two raters (FD and MH) with prior COSMIN tool experience assessed the quality of all included studies independently using the four-point scored COSMIN checklist[13]. This standardized and validated tool consists of 10 sections, each assessing a different measurement property: internal consistency, reliability, measurement error, content validity, construct validity (structural validity and hypothesis testing), cross-cultural validity, criterion validity, responsiveness and interpretability. Each section contains between 5 and 18 items.

Each item within a section is scored using a four-point scoring system with defined response options representing excellent, good, fair or poor quality[13]. An overall quality score for each measurement property reported in a study is defined as the lowest rating of any item within that section, i.e., "worst score counts" method. Depending on the number of measurement properties assessed in a study, some studies receive one quality evaluation whereas other studies receive several.

Evaluation of the measurement property result

In addition to a methodological quality evaluation with COSMIN, an overall rating of the study findings for each measurement property was assessed using a commonly used checklist of criteria

for good measurement properties[18]. These criteria consist of positive, indeterminate and negative ratings for the study findings and are defined in Table I.

Best evidence synthesis: levels of evidence

To synthesize the results from multiple studies on the same performance test, "a best evidence synthesis"[15] was performed by the first author using the criteria outlined in Appendix 2. This best synthesis of evidence is similar to that used for synthesizing evidence from clinical trials[19]. The possible levels of evidence for a measurement property are "strong", "moderate", "limited" "conflicting" or "unknown" (Appendix 2). Best evidence synthesis was derived using the methodological quality of the studies (COSMIN score), the rating and consistency of the measurement property result (positive, indeterminate, negative – Table I), as well as the number of related studies evaluating each measurement property. For this review, studies could only be considered related when the same variation of the performance-based measure was evaluated, that is they were comparable in regards to activity and procedure. Measurement properties from studies that were rated as "poor" on the COSMIN were not eligible to contribute to best evidence synthesis[15].

The COSMIN scoring system used in this review was initially developed for assessing psychometric properties in self-reported questionnaires and defines a minimum adequate sample size as 30 (fair), and adequate sample size as 100 (excellent). It was anticipated that many studies, particularly those evaluating reliability and measurement error, were likely to contain smaller sample sizes than those recommended for self-reported questionnaires. Based on discussions with the developers of the COSMIN, it was decided that to avoid the exclusion of many small samples (which might otherwise be of excellent/good quality) from best evidence synthesis, the sample size item was removed from the COSMIN quality assessment and the "second worst score counts" method was used. Sample size was then accounted for at the evidence synthesis stage. Evidence was assigned as: "strong" when the total sample size of eligible combined studies was ≥100; "moderate" with total samples between 50 and 99; "limited" with total samples between 25 and 49, and "unknown" with samples less than 25.

Results

Description of included studies and performance-based measures

Selection procedures are summarized in Fig. 1. Twenty-four eligible studies were identified and are described in Table II. Measurement properties from 15 single-activity measures were investigated in 12 studies[6,20–30] and from six multi-activity measures investigated in 12 studies[7,8,10,31–39]. Single-activity measures could be grouped into three main activity domains: (1) walking tests, (2) sit to stand tests, and (3) stair negotiation tests.

There were two main types of walk tests, those over short distances (<100 m) and those over long distances (>100 m). There were nine different short-distance walk tests with variations in (1) set pace (self-paced, fast-paced); (2) distance walked (range 2.4–80 m); (3) functional measure (time, speed, distance, quality grading); and (4) incorporated turns (range 0–7). Short-distance

Table I
Quality criteria for rating the results of measurement properties

Property	Rating	Quality criteria
Reliability		
Internal consistency	+	Cronbach's alpha(s) ≥0.70
	?	Cronbach's alpha not determined
	–	Cronbach's alpha(s) <0.70
Reliability	+	ICC/weighted kappa ≥0.70 OR Pearson's $r \geq 0.80$
	?	Neither ICC/weighted kappa, nor Pearson's r determined
	–	ICC/weighted kappa <0.70 OR Pearson's $r < 0.80$
Measurement error	+	MIC >SDC OR MIC outside the LOA
	?	MIC not defined
	–	MIC ≤SDC OR MIC equals or inside LOA
Validity		
Content validity	+	The target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be complete
	?	No target population involvement
	–	The target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete
Structural validity	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	–	Factors explain <50% of the variance
Construct validity hypothesis testing	+	Correlation with an instrument measuring the same construct ≥0.50 OR at least 75% of the results are in accordance with the hypotheses AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	–	Correlation with an instrument measuring the same construct <0.50 OR <75% of the results are in accordance with the hypotheses OR correlation with related constructs is lower than with unrelated constructs
Cross-cultural validity	+	Original factor structure confirmed OR no important DIF between language versions
	?	Confirmatory factor analysis not applied and DIF not assessed
	–	Original factor structure not confirmed OR important DIF found between language versions
Criterion validity	+	Convincing arguments that gold standard is "gold" AND correlation with gold standard ≥0.70
	?	No convincing arguments that gold standard is "gold" OR doubtful design or method
	–	Correlation with gold standard <0.70, despite adequate design and method
Responsiveness		
Responsiveness	+	Correlation with an instrument measuring the same construct ≥0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥0.70 AND correlation with related constructs is higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	–	Correlation with an instrument measuring the same construct <0.50 OR <75% of the results are in accordance with the hypotheses OR AUC <0.70 OR correlation with related constructs is lower than with unrelated constructs

SDC, smallest detectable change; LoA, limits of agreement; DIF, differential item functioning; +, positive rating; ?, indeterminate rating; –, negative rating.
Adapted from Terwee *et al.* J Clin Epidemiol 2007;60(1):34–42.

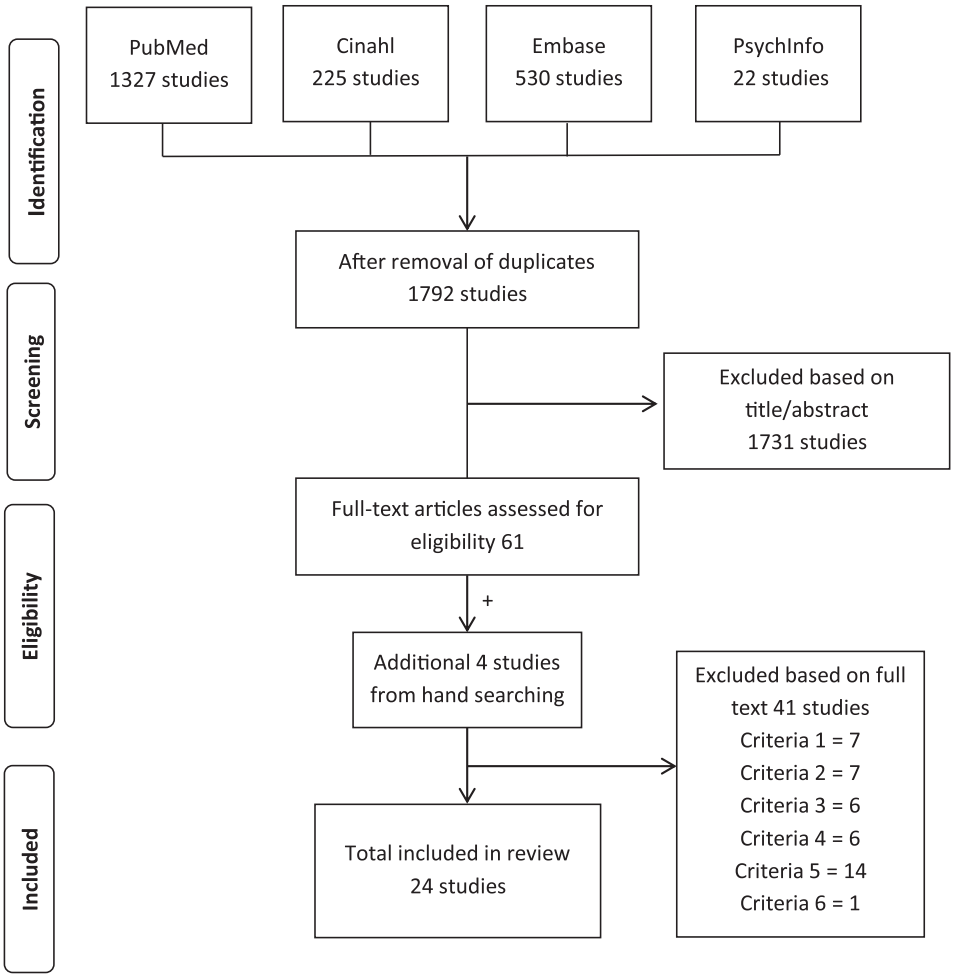

Fig. 1. Flowchart of the selection and inclusion of studies.

walk tests were included in five/six multi-activity measures[7,8,10,31–34,36–39]. The 6-min walk test was the only long-distance walk test and was investigated in four studies[6,22,26,28] and included in two multi-activity measures[8,10,35].

There were six different sit to stand tests with variations in (1) method of measurement (count over 30 s, time for five repetitions, total time and quality grading) and (2) height of chair (standard and high) and (3) incorporated walking and/or turning components (timed up and go test, which incorporates walking 3 m, turning and returning to sit down and the get up and go test, which incorporates walking 20 m with no return). Sit to stand tests were included in three multi-activity measures[7,8,10,31–34].

There were seven different stair negotiation tests with variations in (1) number of stairs (range 4–12); (2) ascend only, descend only or both; (3) hand-rail support and (4) leading limb step pattern. Stair negotiation tests were included in five/six multi-activity measures[7,8,10,31–36].

Three studies included participants with hip OA[24,30,32], five with knee OA[6,20,22,26,27] and 16 with both hip and knee OA[7,8,10,21,23,25,28,29,31,33–39]. The majority of studies included participants in the end stage of OA or the stage of disease was not specified.

Measurement properties

The inter-rater agreement of the independent methodological quality of included studies was good [absolute agreement = 90%, kappa = 0.85, 95% confidence interval (CI) 0.72, 0.98]. Disagreement was mainly due to reading errors and was easily resolved using a consensus method between the two raters.

Internal consistency

Internal consistency was only applicable to multi-activity measures and was assessed in three measures[31,35,37] (Table III). Two studies were rated as "excellent" quality[35,37]. A positive internal consistency rating ($\alpha = 0.82$ and 0.84) was found in both studies.

Reliability and measurement error

Reliability was assessed in 16/21 of the performance measures. Measurement error was assessed in 14/21 of the performance measures (Table III).

Table II
Characteristics of included studies

Author (Year)	Mean age years ± SD (range)	OA site	OA stage	Performance measure	Activity	No. of PPMs	No. of scores	Equipment required	Measurement property assessed
Single-activity measures									
French (2011)[22]	65.3 ± 6.9	Knee	NS	TUG CST 6MWT	Stand, 3 m walk, turn, return, sit Chair-rise × five reps 6 min walking	3	3	Chair, stopwatch walking space	Responsiveness
Gill (2008)[23]	70.3 ± 9.8	Hip/knee	ES/PA	WT CST	Walk 50-feet (15.2 m) fast-paced Chair-rise over 30 s	2	5	20 m walkway Chair, stopwatch	Test–retest reliability Inter-reliability Measurement error
Mizner (2011)[6]	65.0 ± 9.0	Knee	ES/PA	TUG SCT 6MWT	Stand, 3 m walk, turn, return, sit Up and down 12 stairs 6 min walking	3	3	Chair, stopwatch Stairs, Walking space	Responsiveness Construct validity
Wright (2011)[30]	66.5 ± 9.4	Hip	NS	TUG WT CST	Stand, 3 m walk, turn, return, sit Walk 4 × 10 m self-paced Chair-rise over 30 s	4	4	Chair, stopwatch 20 cm step 10 m walkway	Interpretability Inter-reliability Measurement error
Hoeksma (2003)[24]	72.0 ± 6.0	Hip	Early-late K&L 0-IV	WT	Walk 80 m fast-paced			15 m walkway Stopwatch	Responsiveness
Borjesson (2007)[20]	63.0 ± 5.0	Knee	ES/PA	WT	Walk 5 m slow-paced Walk 5 m medium-paced Walk 5 m fast-paced	3	3	<10 m walkway Stopwatch	Responsiveness
Kennedy (2005)[28]	63.7 ± 10.7	Hip/knee	ES/PA	WT SCT TUG 6MWT	Walk 2 × 20 m fast-paced Up and down nine stairs Stand, 3 m walk, turn, return, sit 6 min walking	4	4	Chair, stopwatch >20 m walkway Nine-step stairs Walking space	Test–retest reliability Measurement error Responsiveness
Parent (2002)[26]	68.6 ± 8.7	Knee	ES/PA	6MWT	6 min walking	1	1	Walking space Stopwatch	Responsiveness
Davey (2003)[21]	69.5 ± 7.2	Hip/knee	NS	WT SCT	Walk eight feet self-paced Up and down four stairs	2	2	<5 m walkway Four-step stairs	Test–retest reliability Measurement error
Piva (2004)[27]	62.0 ± 9.0	Knee	Mid-late K&L > 2	GUG	Stand, walk 20 m, no return	1	1	Chair with arms 20 m walkway 15.2 mark Stopwatch	Intra-/inter-reliability Measurement error Construct validity
Marks (1994a)[25]	65.9 ± 8.3	Knee	NS	WT	Walk 13 m self-paced	1	1	13 m walkway Stopwatch	Test–retest reliability Measurement error
Marks (1994b)[29]	59.2 ± 11.1	Knee	NS	WT	Walk 13 m self-paced	1	1	13 m walkway Stopwatch	Test–retest reliability Measurement error Responsiveness
Multi-activity measures									
Oberg (1994)[33]	69.0 ± 9.0	Hip/knee	Early-Mid	FAS	Rise from half stand max no. Sit to stand lowest height Step (max height) Stand one leg Stair climbing (NS) Gait speed over 65 m Walking aid	7	1	Adj height chair Adj height step Stopwatch 65 m walkway Stairs	Inter-reliability Structural validity
Oberg (1997)[34]	68.9 ± 9.7	Hip/knee	Early-Mid	FAS	Rise from half stand max no. Sit to stand lowest height Step (max height) Stand one leg Stair climbing (NS) Gait speed over 65 m Walking aid	7	1	Adj height chair Adj height step Stopwatch 65 m walkway Stairs	Criterion validity
Nilsdotter (2001)[32]	72.6 (52–86)	Hip	ES/PA K&L > 2	FAS	Rise from half stand max no. Sit to stand lowest height Step (max height) Stand one leg Stair climbing (NS)	7	1	Adj height chair Adj height step Stopwatch 65 m walkway Stairs	Responsiveness

					Gait speed over 65 m Walking aid				
McCarthy (2004)[36]	64.7 ± 9.8	Knee	NS	ALF	8 m walk test Seven step SCT up and down Sit transfer test	3	1	10 m space Seven-step stair Chair (no arms) Stopwatch	Test–retest reliability Measurement error Construct validity Responsiveness
Rejeski (1995)[35]	68.8 ± 5.6	Knee	NS	PAR	6MWT Five or nine-step SCT up and down Lift + carry timed In/out car timed	4	1	Walking space Five or nine-step stair Movable shelves 2.2 kg weight Mock up car	Internal consistency Test–retest reliability Convergent validity Concurrent validity
Lin (2001)[31]	69.4 ± 5.9	Hip/knee	NS	Lin Battery	Eight feet walk test Four-step SCT ascend Four-step SCT descend CST x5	4	1	3 m space Four-step stair Chair Stopwatch	Test–retest reliability Measurement error Floor/ceiling Internal consistency Construct validity
Steultjens (1999)[37]	68.0 ± 8.9	Hip/knee	NS	Steultjens	Walk 1 min self-paced Sitting down timed Lying down timed Bend + lift timed	4	1	8 m space Chair Bench 2 kg weight Stopwatch video Trained observer	Internal consistency Construct validity
Steultjens (2000)[38]	68.0 ± 8.9	Hip/knee	NS	Steultjens	Walk 1 min self-paced Sitting down timed Lying down timed Bend + lift timed	4	1	8 m space Chair Bench 2 kg weight Stopwatch video Trained observer	Construct validity
Steultjens (2001)[39]	67.9 ± 8.7	Hip/knee	NS	Steultjens	Walk 1 min self-paced Sitting down timed Lying down timed Bend + lift timed	4	1	8 m space Chair Bench 2 kg weight Stopwatch video Trained observer	Responsiveness
Stratford (2006a)[8]	65 (58–72) (1–3 QR)	Hip/knee	ES/PA	WT TUG SCT 6MWT	Walk 2 × 20m fast-paced Stand, 3 m walk, turn, return, sit Up and down nine stairs 6 min walking	4	1	>20 m space Chair Nine-step stair Walkway	Construct validity
Stratford (2006b)[10]	65.0 (55–77)	Hip/knee	ES/PA	WT TUG SCT 6MWT	Walk 2 × 20 m fast-paced Stand, 3 m walk, turn, return, sit Up and down nine stairs 6 min walking	4	1	>20 m space Chair Nine-step stairs Stopwatch	Construct validity
Stratford (2009)[7]	61.7 ± 10.7	Hip/knee	K&L > 2 ES/PA	WT SCT TUG	Walk 2 × 20 m fast-paced Up and down nine stairs Stand, 3 m walk, turn, return, sit	3	1	>20 m space, Nine-step stair Chair Stopwatch	Construct validity

6MWT, 6-min walk test; CST, chair stand test; ES/PA, end stage/post arthroplasty, FAS, functional assessment system; GUG, get up & go test; K&L, Kellgren and Lawrence classification; SCT, stair-climb test; TUG, timed up & go test; WT, walk test.

Table III
Measurement properties of performance-based measures (reliability and measurement error)

Performance-based measure	Internal consistency			Reliability					Measurement error		
	Result	Study n	COSMIN score	Result	Design	Time interval	Study n	COSMIN score	Result	Study n	COSMIN score
Walk tests											
50ft fast-paced[23]	N/A			$ICC_{1,1}$ 0.91–0.97 (0.86–0.98)	Intra-rater	Intra-session	35–47	Fair	SEM 1.32 s	81	Fair
				$ICC_{1,1}$ 0.94–0.97 (0.90, 0.98)	Inter-rater	Intra-session	28–31	Fair*	MDC_{90} 3.08 s		
40 m self-paced[30]	N/A			$ICC_{2,1}$ 0.95 (0.90, 0.98)	Inter-rater	<1 week	29	Good*	SEM 1.0 m/s	29	Good*
80 m fast-paced[24]	N/A			–					–		
40 m fast-paced[28]	N/A			$ICC_{2,1}$ 0.91 (0.81, 0.97)	Test–retest	Mean 25.4 weeks	21	Fair*	SEM 1.73 s (CI 1.39, 2.29)MDC_{90} 4.04 s	17	Fair*
8 ft self-paced[21]	N/A			Pearson *r* 0.92	Test–retest	<1 week	21	Fair*	SEM 0.12 s	21	Fair*
13 m self-paced[25,29]	N/A			$ICC_{1,1}$ 0.83	Test–retest	6 weeks	10	Good*	SEM 1.5 s	10	Poor
5 m multi-paced[20]	N/A			–					–		
6MWT[22]	N/A			–					–		
6MWT[28]	N/A			$ICC_{2,1}$ 0.94 (0.88, 0.98)	Test–retest	Mean 25.4 weeks	21	Fair*	SEM: 26.29 m (CI 21.14, 34.77)	17	Fair*
6MWT[6]	N/A			–					–		
6MWT[26]	N/A			–					–		
CST											
x5 chair stand[22]	N/A			–					–		
30 s-chair stand[23]	N/A			$ICC_{1,1}$ 0.97–0.98 (0.94, 0.99)	Intra-rater	Intra-session	37–47	Fair	SEM 0.7 stands	40	Fair
				$ICC_{1,1}$ 0.93–0.98 (0.87, 0.99)	Inter-rater	Intra-session	28–42	Fair*	MDC_{90} 1.64 stands		
30 s-chair stand[30]	N/A			$ICC_{2,1}$ 0.81 (0.63, 0.91)	Inter-rater	<1 week	29	Good*	SEM 1.27 stands	29	Good*
TUG[22]	N/A			–					–		
TUG[6]	N/A			–					–		
TUG[30]	N/A			$ICC_{2,1}$ 0.87 (0.74, 0.94)	Inter-rater	<1 week	29	Good*	SEM 0.84 s	29	Good*
TUG[28]	N/A			$ICC_{2,1}$ 0.75 (0.51, 0.89)	Test–retest	Mean 25.4 weeks	21	Fair*	SEM 1.07 s (0.86, 1.41)	17	Fair*
GUG[27]	N/A			ICC 0.95 (0.72–0.98)	Intra-rater	2 min	25	Poor	SEM 0.55 s, MDC 1.5 s	25	Poor
				ICC 0.98 (0.94–0.99)	Inter-rater	2 min	25	Good*	SEM 0.42 s, MDC 1.2 s	25	Good*
SCTs											
12-stair up/down[6]	N/A			–					–		
Nine-stair up/down[28]	N/A			$ICC_{2,1}$ 0.90 (0.79, 0.96)	Test–retest	Mean 25.4 weeks	21	Fair*	SEM 2.35 s (1.89, 3.10)	17	Fair*
Four-stair up/down[21]	N/A			Pearson *r* 0.92	Test–retest	<1 week	21	Fair*	SEM 0.23 s		
Multi-activity tests											
Lin battery[31]	$\alpha = 0.84$	106	Poor	ICC 0.94–0.96 (0.75–0.99)	Test–retest	N/S	10	Fair*	SEM 0.10–1.44 s	10	Good*
PAR[35]	$\alpha = 0.82$	203	Excellent	$r = 0.88$–0.93 (range of all tests)	Test–retest	2 weeks	25	Fair*	–		
				$r = 0.72$–0.86 (range of all tests)	Test–retest	3 months	148	Fair*			
ALF[36]	–			ICC 0.99 (0.98–0.99) total ALF	Test–retest	1 week	15	Good*	SEM 0.86 s	15	Good*
Steultjens battery[37–39]	$\alpha = 0.84$	198	Excellent	–					–		
Stratford battery[7,8,10]	N/A			–					–		
FAS[33]	–			$G = 0.99$–1.0 (range of all tests)	Inter-tester	?	42	Fair	–		

N/A, not applicable for single-activity tests or multi-activity tests using reflective models; FAS, functional assessment system; *G*, Goodman–Kruskal gamma; MDC, minimal detectable change.
* Denotes a change of COSMIN score after to removal of sample size item from the rating.

Single-activity measures

For walking tests, a positive rating [i.e., intraclass correlation coefficient (ICC) > 0.70] for intra-rater reliability [ICC 0.91–0.97 (CI: 0.86–0.98)] and inter-rater reliability [ICC 0.94–0.97 (CI: 0.90, 0.98)] was reported for the 50ft (15.2 m)-walk test in one "fair" quality study of hip and knee OA[23]. A positive rating for inter-rater reliability [ICC 0.95 (CI: 0.90, 0.98)] was also reported for the 40 m-walk test in one "good" quality study of hip OA[30]. For sit to stand tests, a positive rating for inter-tester reliability [ICC 0.87 (CI: 0.74, 0.94)] was reported for the timed up and go test in one "good" study of hip OA[30]. The 30 s-chair stand test was also found to have a positive rating for intra-tester [ICC 0.97–0.98 (CI: 0.94, 0.99)] and inter-tester [ICC 0.93–0.98 (CI: 0.87, 0.99)] reliability in a "fair" study of hip and knee OA[23] and inter-tester [ICC 0.81 (CI: 0.63, 0.91)] reliability in a "good" study of hip OA[30]. Evidence for stair negotiation tests and other single-activity measures was limited by small total sample sizes or inappropriate time intervals between repeat testing.

The standard error of measurement (SEM), along with minimum important change (MIC) was reported in only three of the 12 single-activity measures (40 m-walk test, timed and 30 s-chair stand test)[30]. Measurement error and MIC was defined in one "good" quality study for the 40 m-walk test (SEM 1.0 m/s; MIC 2.0 m/s), timed up and go test (SEM 0.84 s; MIC 0.8–1.4) and the 30 s-chair stand test (SEM 1.27 stands; MIC 2.0–2.6 stands)[30]. As MIC was not calculated for the remaining single-activities, quality ratings were indeterminate for these measures.

Multi-activity measures

Reliability of multi-activity measures was reported in three "fair" quality studies[31,33,35] and one "good" quality study[36]. A positive rating for test–rest reliability was reported for the Physical Activity Restrictions (PAR) (ICC 0.72–0.86)[35]. A positive rating for inter-tester rating (Goodman–Kruskal Gamma 0.99–1.0) was found for the Functional Assessment System (FAS)[33]. Evidence of reliability for other test batteries was limited due to inadequate total sample size.

Measurement error was reported in two test batteries[31,36] however as MIC has not been calculated for either battery, quality ratings were indeterminate.

Validity studies

Validity was assessed in 9/21 (43%) of performance tests (Table IV).

Single-activity measures

Construct validity was investigated for three single-activity performance measures[6,27]. In one "good" quality study, a positive rating of construct validity was found for the timed up and go test and the 12-step stair-climb test as more than 75% of the results were in accordance with the hypotheses[6]. In another "good" quality study a negative rating of construct validity was found for the get up and go test as less than 75% of the results were in accordance with the hypotheses[27].

Multi-activity measures

Validity was investigated in all six multi-activity batteries and four were rated as "good" quality for construct validity[7,8,10,35,37,38] and one was rated as "fair" quality for criterion and structural validity[34]. The PAR[35] demonstrated mostly positive convergent validity with treadmill time, VO_2 peak and strength and divergent validity with self-reported dysfunction as predicted. The Steultjens battery[38] demonstrated a negative convergent validity with self-reported mobility and joint range of motion. The Stratford battery demonstrated positive construct validity in two "good" quality studies and one "fair" study[7,8,10]. The FAS demonstrated positive structural validity in one "fair" quality study[33] and positive criterion validity with good sensitivity (0.70–0.89) and specificity (0.57–1.0)[34].

Responsiveness

Single-activity measures

Responsiveness was reported in 12/15 single-activity measures (Table IV). Responsiveness of walking tests was reported in four "fair" quality studies following either physiotherapy/exercise[24,30] or joint arthroplasty[20,28]. A positive rating [i.e., area under the curve (AUC) > 0.70] was reported for the 40 m-walk test (AUC = 0.89)[30] and the 80 m-walk test (AUC = 0.71)[24]. Responsiveness of other walk tests was reported using standard response means (SRM) or effect sizes (ES) (see Table IV) and results were therefore indeterminate. Responsiveness of sit to stand tests was reported in three "fair" quality studies following either physiotherapy[30] or joint arthroplasty[6,28]. A positive rating was reported for the 30 s-chair stand test (AUC = 0.73) and a negative rating (AUC < 0.70) was reported for the timed up and go test (AUC = 0.69) following physiotherapy/exercise[30]. Responsiveness of other sit to stand tests following joint arthroplasty[6,28] and all stair negotiation tests[6,28] was reported using ES and/or SRM and therefore results were indeterminate.

Multi-activity measures

Responsiveness was reported in three/six multi-activity measures following either exercise[36,39] or hip arthroplasty[32]. One study was "good" quality[39] and the others were "fair"[32,36]. A negative rating of responsiveness of the Steultjens battery[39] was found as <75% of the results were in accordance with the hypotheses. Other batteries provided SRM and results were indeterminate.

Interpretability

Evidence of interpretability was reported in one "good" quality study that evaluated three single-activity measures[30]. Major clinically important improvement (MCII) of the 40 m self-paced walk test (0.2–0.3 m/s), 30 s-chair stand test (2.0–2.6 stands) and the timed up and go test (0.8–1.4 s), were reported[30].

Best evidence synthesis: levels of evidence

A summary of best evidence synthesis for each of the 21 performance tests is provided in Table V. This synthesis was derived from information found in Tables III and IV including (1) the methodological quality (COSMIN), (2) the findings (result), and (3) the sample size. Given the large variety of performance-based measures, results were rarely combined. The exceptions were for the Steultjens battery and the Stratford battery. A positive rating (limited, moderate or strong evidence) was given to only 25/153 (16%) of all possible ratings.

Discussion

In this systematic review we identified 24 eligible studies that reported the measurement properties of 21 different performance-based measures of physical function in individuals with hip and/or knee OA. The majority of studies were rated as "fair" quality using the modified COSMIN tool. Evidence for most measurement properties is yet to be determined either because there was no information available, information was indeterminate or because evidence was only available from poor quality studies. Studies were mostly rated as poor quality due to unclear hypotheses and/or non-optimal analyses. Although none of the measures included in the

Table IV
Measurement properties of performance-based measures (validity, responsiveness and interpretability)

Performance-based measure	Validity (hypothesis testing)				Responsiveness			Interpretability	
	Design	Result	Study *n*	COSMIN score	Treatment	Result	COSMIN score	Result	COSMIN score
Walk tests	–								
50ft fast-paced[23]	–				–				
40 m self-paced[30]	–				PT x9 sessions	AUC 0.89 (0.76, 1.00)	Fair	MCII 0.2–0.3 m/s	Good
80 m fast-paced[24]	–				PT x9 sessions	AUC 0.71 (0.58, 0.83) GRI 0.45	Fair		
40 m fast-paced[28]	–				Hip/knee arthroplasty	SRM −0.89 (−1.42, −0.68) pre-first post; SRM 0.79 (0.66, 1.45) first-second post	Fair		
8ft self-paced[21]	–				–				
13 m self-paced (29)	–				Quads exercise (6 weeks)	$r = 0.9$ with quads strength	Poor		
5 m multi-paced[20]	–				Knee arthroplasty	ES/SRM/RE at slow speed: 0.58/0.71/1.62	Fair		
6MWT[22]	–				PT mean 5.8 sessions	ES/ES med/SRM 0.39/0.43/0.54	Poor		
6MWT[28]	–				Hip/knee arthroplasty	SRM pre-post1: −1.74 (1.60, 1.97) SRM post1-post2: 1.90 (1.46, 2.39)	Fair		
6MWT[6]	–								
6MWT[26]	–				Knee arthroplasty ± PT	SRM/ES: pre-2 mth post 0.63/0.41 2–4 mth post 1.51/0.82 pre-4 mth post 0.58/0.35	Fair		
CST									
x5 chair stand[22]	–				PT mean 5.8 sessions	ES/Es med/SRM 0.36, 0.33, 0.39	Poor		
30 s-chair stand[23]	–				–				
30 s-chair stand[30]	–				PT x9 sessions	AUC 0.73 (0.55, 0.91)	Fair	MCII 2.0–2.6 stands	Good
TUG[22]	–				PT mean 5.8 sessions	ES/ES med/SRM 0.33/0.17/0.35	Poor		
TUG[6]	Construct	Low correlations with PROs as predicted; $r = -0.40$ to -0.48 with quads strength as predicted	100	Good	Knee arthroplasty	ES pre-1 mth/pre-12 mth /1-12 mth: −0.43, 0.79, 1.17	Fair		
TUG[30]	–				PT x9 sessions	AUC 0.69 (0.48, 0.90)	Fair	MCII 0.8–1.4 s	Good
TUG[28]	–				Hip/knee arthroplasty	SRM pre-post1: −1.08 (−1.38, −0.92) SRM post1–post2: 1.04 (0.84, 1.61)	Fair		
GUG[27]	Construct Divergent	Sig diff b/w patients and controls $P < 0.001$	50	Fair		–			
	Convergent	$r = 0.39$; −0.44; −0.34 with WOMAC/ SF-36 PF/ADLS correlation with related constructs higher than unrelated <75% of results in accordance with hypothesis	105	Good	–				
SCTs									
12-stair up/down[6]	Construct	Poor correlation with PROs as predicted; $r = -0.36$ to -0.46 with quads strength as predicted	100	Good	Knee arthroplasty	ES pre-1 mth/pre-12 mth /1-12 mth:−0.71, 0.84, 1.26	Fair		
Nine-stair up/down[28]	–				Hip/knee arthroplasty	SRM pre-post1: −1.74 (−2.13, −1.45) SRM post1–post2: 1.98 (1.68, 2.42)	Fair		
Four-stair up/down[21]	–				–				

Multi-activity tests						
Lin battery[31]	Construct	$r = 0.48-0.54$ with WOMAC-PF	106	Poor	–	
PAR[35]	Construct	0.30–0.60 Treadmill time, VO_2 peak	104–437	Good	–	
	Convergent	quads strength				
	Divergent	0.03–0.93 self-reported dysfunction	104–437			
ALF[36]	Construct	$r = 0.59/-0.53$ with WOMAC/SF-36PF	214	Poor	Exercise program	SRM 0.49 at 12 months f/u
Steultjens battery[37–39]	Construct	$r = 0.29-0.55$ with self-rated mobility	198	Fair	Exercise program	No differential responsiveness of observed vs self-report
		$r = 0.25-0.35$ with ROM	198	Good		Different factor structure than expected
Stratford battery[7,8,10]	Construct	SPWT, TUG, 6MWT best combination to evaluate Pain and performance	177	Fair	–	
	Construct	Change in pain rather than performance (time/distance) is principal determinant of change in self-reported function	85	Good	–	
	Construct	ANOVA $P < 0.001$: PB was more sensitive to change than SR measures	73	Good	–	
FAS[32–34]	Structural	PCA-5 factors loading with physical disability primarily 1 factor explaining 51–82% of variance	105	Fair	Hip arthroplasty	SRM of mean score = 0.4 at 3 months post-op SRM of mean score = 0.7 at 6 months post-op
	Construct	PPMs were better able to discriminate btw healthy and OA and btw hip and knee OA $P < 0.001$ delta 0.67–0.93				
	Criterion	Sensitivity 0.70–0.89 Specificity 0.57–1.0 (SPWT and SCT had best sensitivity and specificity)	Controls 42 Hip OA 302 Knee OA 258	Fair		

ADLS, activities of daily living; ANOVA, analysis of variance; ES, effect size index; ES med, effect size median; FAS, functional assessment system; GRI, Gyatts responsiveness index; PCA, principal component analysis; PB, performance battery; PPM, physical performance measure; PRO, patient-reported outcome; PT, physiotherapy; ROM, range of movement; SF-36 PF, short-form health survey physical function; SPWT, self-paced walk test; WOMAC, Western Ontario and McMaster Universities Arthritis Index.

Table V
Levels of evidence of performance-based measures

Performance-based measure	Internal consistency	Reliability			Measurement error	Validity	Responsiveness	Interpretability
		Intra	Inter	Retest				
Single-activity measures								
Walk tests								
50ft fast-paced[23]	N/A	+(HK)	+(HK)	0	?	0	0	0
40 m self-paced[30]	N/A	0	+(H)	0	+(H)	0	+(H)*	++(H)
80 m fast-paced[24]	N/A	0	0	0	0	0	+(H)*	0
13 m self-paced[25,29]	N/A	0	0	?	?	0	0	0
8ft self-paced[21]	N/A	0	0	?	?	0	0	0
40 m fast-paced[28]	N/A	0	0	?	?	0	?	0
5 m-slow/medium/fast[20]	N/A	0	0	0	0	0	?	0
6-min[6,22,26,28]	N/A	0	0	?	?	0	?	0
Sit to stand tests								
30 s-chair stand[23,30]	N/A	+(HK)	+(HK)	0	+(H)	0	+(H)*	++(H)
X5 chair stand[22]	N/A	0	0	?	?	0	?	0
Timed up and go[6,22,30]	N/A	0	+(H)	?	+(H)	++(K)	−(H)*	++(H)
Get up and go[27]	N/A	?	0	?	?	−−(K)	0	0
Stair negotiation tests								
12-stair up and down[6]	N/A	0	0	0	0	++(K)	?	0
Nine-stair up and down[28]	N/A	0	0	?	?	0	?	0
Four-stair up and down[21]	N/A	0	0	?	?	0	0	0
Multi-activity measures								
Lin31	?	0	0	?	?	?	0	0
PAR[35]	+++(K)	0	0	+(K)	0	++(K)	0	0
ALF[36]	0	0	0	?	?	?	?	0
Steultjens[37–39]	+++(HK)	0	0	0	0	−−(HK)	−−(HK)	0
Stratford[7,8,10]	0	0	0	0	0	+++(HK)	0	0
FAS[32–34]	0	0	+(HK)	0	0	+(HK)† +(HK)‡	?	0

+++ or −−− strong evidence, ++ or −− moderate evidence, + or − limited evidence, ± conflicting evidence, ? unknown, 0 no information [+ = positive, − negative rating (results)], (H) = hip, (K) = Knee, (HK) = Hip and Knee.

* Physiotherapy/exercise.

† Structural validity.

‡ Criterion validity.

review reported evidence for all measurement properties, positive evidence for a selected few measures was established across multiple measurement properties. This provides useful information for clinicians and researchers about which performance-based measures are currently the most suitable for assessing people with hip and/or knee OA.

Similar to a previous review[3], the current review identified a variety of performance-based measures that represented several different activity domains. For example, in this review, 10 different variations of the walking test were identified. As such, we found it useful to group the measures under three main activity themes: (1) walking tests; (2) sit to stand tests; and (3) stair negotiation tests. An additional group, multi-activity measures, contains different variations and combinations of the three activity domains as well as some additional domains such as getting in/out of a car[35] and lift and carrying tasks[35,37–39].

Walking tests

Walking tests with the best measurement evidence included the 40 m self-paced walk test for hip OA[30] and the 50ft (15.2 m) fast-paced walk test for hip/knee OA[23]. Evidence for other walk tests such as the 6-min walk test has yet to be determined in people with hip and/or knee OA.

Sit to stand tests

Sit to stand tests with the best measurement evidence included the 30 s-chair stand test and the timed up and go test for hip/knee OA[6,23,30]. Evidence for the five-repetition chair stand test has yet to be determined. Based on current levels of evidence, the get up and go test[27] is not recommended for use in people with either hip or knee OA.

Stair negotiation tests

Evidence for most variations of stair tests has yet to be determined. Only evidence of construct validity was reported for the 12-step stair test for knee OA[6]. Given the current limited evidence of stair negotiation tests, recommendations about which tests might be more useful cannot be made.

Multi-activity measures

Multi-activity measures with the best measurement evidence were the PAR[35], the Stratford battery[7,8,10] and the FAS[32–34]. In addition, the PAR provided a good justification for the choice of included activities which consisted of a walking test (6-min walk test), a stair negotiation test (five or nine-stair ascent/decent), a lift and carry test and a car test. Based on current levels of evidence, the Steultjens battery is not recommended for hip and knee OA[38,39]. Evidence for the aggregated locomotor function (ALF) and Lin test is yet to be determined.

A number of factors influenced the evidence found in the review. The COSMIN quality scoring system developed for self-reported questionnaires was modified to enable smaller studies that were otherwise of acceptable quality, to be included in best evidence synthesis. This change influenced the findings of the majority of the reliability studies. Without this change, there would have been no evidence for reliability for any of the measures

included in the review. Best evidence synthesis was mostly obtained from a single study as the majority of results could not be combined because of the large variations in the testing procedures. Further, for most multi-activity tests included in this review, there was no information about the measurement model (reflective or formative) in the development of the tests, nor in the validation studies. Therefore it is difficult to tell how important internal consistency is for these tests. For some of the included tests, that were based on a formative model, where the activities define the construct (causal indicators) internal consistency may not be relevant[15].

There were some limitations to this review. Publication bias from unpublished studies may threaten the internal validity as unpublished studies are more likely to report negative or unfavourable results. The decision to exclude measures that used sophisticated equipment or measured constructs other than those defined as 'Activities' according to the ICF[4] (i.e., balance measures) meant that evidence for these types of measures was not included in the review. In addition, further evidence may have been found from some potentially good studies that fell short of the 80% OA sample criteria[40–46]. We found considerable variations in the performance-based measures which meant most evidence from multiple studies of a measure could not be combined. Stronger evidence may have been found if a larger number of more similar studies were available.

This review highlights a number of areas worthy of future research. More studies of the responsiveness and clinically MIC of performance-based measures for people with hip and knee OA are required. Although there is growing evidence for some of the performance measures included in this review, no test has been evaluated with respect to all measurement properties. On balance of the evidence, the 40 m self-paced test[30] was the best rated walk test, the 30 s-chair stand test[30] and timed up and go test[30] were the best rated sit to stand tests, and the PAR[35], Stratford battery[7,8,10], and FAS[32–34] were the best rated multi-activity measures. Additionally, before strong recommendations can be made, consensus is still required on which variation of an activity theme is best and what combination of tests would best assess physical function in people with hip and/or knee OA. Extensive variation in types of outcomes measures has been found across trials[5,47], making comparisons across studies and synthesis of results difficult[9]. We agree with recommendations that future work should be directed at whether consensus can be achieved towards a standardised set of performance-based outcome measures[3,5,9].

Conclusion

This systematic review highlighted current gaps in our knowledge of evidence about the measurement properties of performance-based measures of physical function in people with hip and/or knee OA. Further good quality research investigating the measurement properties, and in particular the responsiveness and interpretability of performance-based measures, in people with hip and/or knee OA is needed. Consensus on which combination of measures will best assess physical function in hip/and or knee OA is urgently required.

Author contributions

FD contributed to the conception and design of the study including obtaining of funding, collection and assembly of data, analysis and interpretation of data, writing of the manuscript and final approval of the article. MH contributed to collection and assembly of data, drafting and final approval of the article. RSH, KLB and EMR contributed to conception and design of the study including obtaining of funding, analysis and interpretation of the data, critical revision of the article for important intellectual content and final approval of the article. CBT contributed to the conception and design, analysis and interpretation of the data, critical revision of the article for important intellectual content and final approval of the article. First and last authors take responsibility for the integrity of the work as a whole, from inception to finished article.

Role of the funding source

This project was partly funded by the OARSI, NHMRC Program Grant #631717 and the Arthritis Australia and States & Territory Affiliates Grant and forms part of an OARSI initiative to develop a recommended set of physical performance measures for hip and knee OA. Kim Bennell is partly funded by an Australian Research Council Future Fellowship. The study sponsor did not play any role in the study design, collection, analysis or interpretation of data; nor in the writing of the manuscript or decision to submit the manuscript for publication.

Conflict of interest

There are no other financial interests that any of the authors may have, which could create a potential conflict of interest or the appearance of a conflict of interest with regard to the work.

Appendix 1. Search strategy

Filter 1: Construct terms

("physical function*"[tw] OR "motor activity"[MH] OR "physical activity"[tw] OR "physical activities"[tw] OR "physical performance*"[tw] OR "functional activity"[tw] OR "functional activities"[tw] OR "functional performance*"[tw] OR "activity limitation*"[tw] OR "functional limitation*"[tw] OR disability[Title/Abstract] OR disabilities[Title/Abstract] OR "Activities of daily living"[MH]).

Filter 2: Target population

("osteoarthritis"[MH]) OR osteoarthritis[Title/Abstract] OR "arthritis"[MH]) OR arthritis[Title/Abstract]) OR (replacement [Title/Abstract] OR arthroplasty[Title/Abstract]) AND (hip[Title/Abstract] OR knee[Title/Abstract] OR "lower limb"[Title/Abstract]).

Filter 3: Instrument terms

("physical performance measure*"[tw] OR "performance test*"[tw] OR "performance-based test"[tw] OR "performance-based tests"[tw] OR "performance based test*"[tw] OR "performance measure*"[tw] OR "performance-based measure"[tw] OR "performance-based measures"[tw] OR "performance instrument*"[Title/Abstract] OR "performance-based instrument"[Title/Abstract] OR "performance-based instruments"[Title/Abstract] OR "performance-based method"[Title/Abstract] OR "performance-based methods"[Title/Abstract] OR "performance based method*"[Title/Abstract] OR "performance index"[Title/Abstract] OR "performance indices"[Title/Abstract] OR "performance-based index"[Title/Abstract] OR "performance-based indices"[Title/Abstract] OR "performance-based assessment"[Title/Abstract] OR "performance-based assessments"[Title/Abstract] OR "objective test*"[Title/Abstract] OR "objective instrument*"[Title/Abstract] OR "objective method*"[Title/Abstract] OR "objective measure*"[Title/Abstract] OR "objective evaluation*"[Title/Abstract] OR "objective function*"[Title/Abstract] OR "objective disability"[Title/Abstract] OR "objective assessment*"[Title/Abstract] OR "observational

test*"[Title/Abstract] OR "observational-based test"[Title/Abstract] OR "observational-based tests"[Title/Abstract] OR "observational testing"[Title/Abstract] OR "observational instrument*"[Title/Abstract] OR "observational-based instrument"[Title/Abstract] OR "observational-based instruments"[Title/Abstract] OR "observational method*"[Title/Abstract] OR "observational-based method"[Title/Abstract] OR "observational-based methods"[Title/Abstract] OR "observational measure*"[Title/Abstract] OR "observational-based measure"[Title/Abstract] OR "observational-based measures"[Title/Abstract] OR "observational index"[Title/Abstract] OR "observational indices"[Title/Abstract] OR "observation-based index"[Title/Abstract] OR "observation-based indices"[Title/Abstract] OR "observed disability"[Title/Abstract] OR "observed function"[Title/Abstract] OR "gait analysis"[Title/Abstract] OR "gait evaluation"[Title/Abstract] OR "walk* test"[Title/Abstract] OR "task performance and analysis"[MH] OR Outcome Assessment[MH]).

Filter 4: Sensitive search filter for measurement properties

(instrumentation[sh] OR methods[sh] OR validation studies[pt] OR Comparative Study[pt] OR psychometrics[MH] OR psychometr* [tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MH] OR "outcome assessment"[tiab] OR "outcome measure*"[tw] OR "observer variation"[MH] OR "observer variation"[tiab] OR "Health Status Indicators"[MH] OR "reproducibility of results"[MH] OR reproducib*[tiab] OR "discriminant analysis"[MH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR coefficient[tiab] OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement [tiab] OR precision[tiab] OR imprecision[tiab] OR "precise values"[tiab] OR test–retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester [tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer [tiab] OR intraobserver[tiab] OR intraobserver[tiab] OR inter-technician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR inter-participant[tiab] OR inter-participant[tiab] OR intraparticipant [tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tiab] OR ((replicab*[tiab] OR repeated [tiab]) AND (measure[tiab] OR measures[tiab] OR findings[tiab] OR result[tiab] OR results[tiab] OR test[tiab] OR tests[tiab])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR factor analysis[tiab] OR factor analyses [tiab] OR dimension*[tiab] OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR item discriminant[tiab] OR interscale correlation*[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically [tiab]) AND (important[tiab] OR significant[tiab] OR detectable [tiab])AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference [tiab])) OR meaningful change[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab]).

Filter 5: Exclusion filter

("addresses"[PT] OR "biography"[PT] OR "case reports"[PT] OR "comment"[PT] OR "directory"[PT] OR "editorial"[PT] OR "festschrift"[PT] OR "interview"[PT] OR "lectures"[PT] OR "legal cases"[PT] OR "legislation"[PT] OR "letter"[PT] OR "news"[PT] OR "newspaper article"[PT] OR "patient education handout"[PT] OR "popular works"[PT] OR "congresses"[PT] OR "consensus development conference"[PT] OR "consensus development conference, nih"[PT] OR "practice guideline"[Publication Type]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms]).

Appendix 2. Levels of evidence for the quality of the measurement property

Level	Rating*	Criteria
Strong	+++ or −−−	Consistent findings in multiple studies of good Methodological quality OR in one study of excellent Methodological quality
Moderate	++ or −−	Consistent findings in multiple studies of fair Methodological quality OR in one study of good Methodological quality
Limited	+ or −	One study of fair methodological quality
Conflicting	±	Conflicting findings
Unknown	?	Only studies of poor methodological quality

Adapted from Terwee *et al.* J Clin Epidemiol 2007;60(1):34–42.
* + = positive rating, ? = indeterminate rating, − = negative rating.

References

1. Pham T, van der Heijde D, Altman RD, Anderson JJ, Bellamy N, Hochberg M, *et al.* OMERACT-OARSI initiative: Osteoarthritis Research Society International set of responder criteria for osteoarthritis clinical trials revisited. Osteoarthritis Cartilage 2004;12:389–99.
2. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, *et al.* Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol 1997;24:799–802.
3. Terwee CB, Mokkink LB, Steultjens MP, 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: 890–902.
4. World Health Organization. International Classification of Functioning, Disability, and Health. Geneva, Switzerland: ICF; 2001.
5. Wright AA, Hegedus EJ, David Baxter G, Abbott JH. Measurement of function in hip osteoarthritis: developing a standardized approach for physical performance measures. Physiother Theor Pract 2011;27:253–62.
6. Mizner RL, Petterson SC, Clements KE, Zeni Jr JA, Irrgang JJ, Snyder-Mackler L. Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments. A longitudinal analysis of outcomes. J Arthroplasty 2011;26:728–37.

7. Stratford PW, Kennedy DM, Riddle DL. New study design evaluated the validity of measures to assess change after hip or knee arthroplasty. J Clin Epidemiol 2009;62:347–52.
8. Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther 2006;86:1489–96.
9. Jordan KP, Wilkie R, Muller S, Myers H, Nicholls E. Measurement of change in function and disability in osteoarthritis: current approaches and future challenges. Curr Opin Rheumatol 2009;21:525–30.
10. Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. J Clin Epidemiol 2006;59:160–7.
11. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, *et al*. 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 2010;19:539–49.
12. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, *et al*. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res 2009;18:313–33.
13. Terwee C, Mokkink L, Knol D, Ostelo R, Bouter L, de Vet H. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651–7.
14. Moher D, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264–9.
15. de Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in Medicine: A Practical Guide to Biostatistics and Epidemiology. London: Cambridge University Press; 2011.
16. Terwee CB, Jansma EP, Riphagen II , de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res 2009;18:1115–23.
17. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, *et al*. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
18. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34–42.
19. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.
20. Borjesson M, Weidenhielm L, Elfving B, Olsson E. Tests of walking ability at different speeds in patients with knee osteoarthritis. Physiother Res Int 2007;12:115–21.
21. Davey RC, Edwards SM, Cochrane T. Test–retest reliability of lower extremity functional and self-reported measures in elderly with osteoarthritis. Adv Physiother 2003;5:155–60.
22. French HP, Fitzpatrick M, FitzGerald O. Responsiveness of physical function outcomes following physiotherapy intervention for osteoarthritis of the knee: an outcome comparison study. Physiotherapy 2011;97:302–8.
23. Gill S, McBurney H. Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. Physiother Res Int 2008;13:141–52.
24. Hoeksma HL, Van Den Ende CHM, Ronday HK, Heering A, Breedveld FC, Dekker J. Comparison of the responsiveness of the Harris Hip Score with generic measures for hip function in osteoarthritis of the hip. Ann Rheum Dis 2003;62:935–8.
25. Marks R. Walking time measures for evaluating OA of the knee. S Afr J Physiother 1994;50:5+7–8.
26. 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.
27. 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.
28. Kennedy DM, Stratford PW, Wessel J, Gollish JD, 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.
29. Marks R. Reliability and validity of self-paced walking time measures for knee osteoarthritis. Arthritis Care Res 1994;7:50–3.
30. Wright AA, Cook CE, Baxter GD, Dockerty JD, Abbott JH. 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;41:319–27.
31. Lin YC, Davey RC, Cochrane T. Tests for physical function of the elderly with knee and hip osteoarthritis. Scand J Med Sci Sports 2001;11:280–6.
32. Nilsdotter A, Roos EM, Westerlund JP, Roos HP, Lohmander LS. Comparative responsiveness of measures of pain and function after total hip replacement. Arthritis Care Res 2001;45:258–62.
33. Oberg U, Oberg B, Oberg T. Validity and reliability of a new assessment of lower-extremity dysfunction. Phys Ther 1994;74:861–71.
34. Oberg U, Oberg T. Discriminatory power, sensitivity and specificity of a new assessment system (FAS). Physiother Can 1997;49:40–7.
35. Rejeski WJ, Ettinger Jr WH, Schumaker S, James P, Burns R, Elam JT. Assessing performance-related disability in patients with knee osteoarthritis. Osteoarthritis Cartilage 1995;3:157–67.
36. McCarthy CJ, Oldham JA. The reliability, validity and responsiveness of an aggregated locomotor function (ALF) score in patients with osteoarthritis of the knee. Rheumatology (Oxford) 2004;43:514–7.
37. Steultjens MP, Dekker J, van Baar ME, Oostendorp RA, Bijlsma JW. Internal consistency and validity of an observational method for assessing disability in mobility in patients with osteoarthritis. Arthritis Care Res 1999;12:19–25.
38. Steultjens MP, Dekker J, van Baar ME, Oostendorp RA, Bijlsma JW. Range of joint motion and disability in patients with osteoarthritis of the knee or hip. Rheumatology (Oxford) 2000;39:955–61.
39. 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.
40. Almeida GJ, Schroeder CA, Gil AB, Fitzgerald GK, Piva SR. Interrater reliability and validity of the stair ascend/descend test in subjects with total knee arthroplasty. Arch Phys Med Rehabil 2010;91:932–8.
41. Bremander AB, Dahl LL, Roos EM. Validity and reliability of functional performance tests in meniscectomized patients with or without knee osteoarthritis. Scand J Med Sci Sports 2007;17:120–7.

42. Cecchi F, Molino-Lova R, Di Iorio A, Conti AA, Mannoni A, Lauretani F, *et al*. Measures of physical performance capture the excess disability associated with hip pain or knee pain in older persons. J Gerontol A Biol Sci Med Sci 2009;64:1316–24.
43. Crosbie J, Naylor JM, Harmer AR. Six minute walk distance or stair negotiation? Choice of activity assessment following total knee replacement. Physiother Res Int 2010;15:35–41.
44. Kwoh CK, Petrick MA, Munin MC. Inter-rater reliability for function and strength measurements in the acute care hospital after elective hip and knee arthroplasty. Arthritis Care Res 1997;10:128–34.
45. Jakobsen TL, Kehlet H, Bandholm T. Reliability of the 6-min walk test after total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc, in press.
46. Stevens-Lapsley JE, Schenkman ML, Dayton MR. Comparison of self-reported knee injury and osteoarthritis outcome score to performance measures in patients after total knee arthroplasty. PM R 2011;3:541–9.
47. Riddle DL, Stratford PW, Bowman DH. Findings of extensive variation in the types of outcome measures used in hip and knee replacement clinical trials: a systematic review. Arthritis Rheum 2008;59:876–83.