Author/s:
Kroman, SL;Roos, EM;Bennell, KL;Hinman, RS;Dobson, F

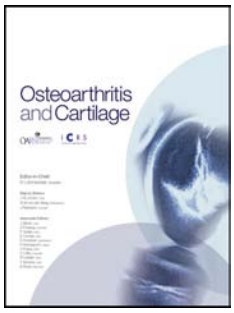
Title:
Measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee osteoarthritis: A systematic review

Date:
2014-01-01

Citation:
Kroman, S. L., Roos, E. M., Bennell, K. L., Hinman, R. S. & Dobson, F. (2014). Measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee osteoarthritis: A systematic review. Osteoarthritis and Cartilage, 22 (1), pp.26-39. https://doi.org/10.1016/j.joca.2013.10.021.

Persistent Link:
https://hdl.handle.net/11343/43853

Accepted Manuscript



Measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee osteoarthritis: A systematic review

Sarah L. Kroman, Ewa M. Roos, Kim L. Bennell, Rana S. Hinman, Fiona Dobson

PII: S1063-4584(13)01004-2

DOI: 10.1016/j.joca.2013.10.021

Reference: YJOCA 3016

To appear in: *Osteoarthritis and Cartilage*

Received Date: 12 June 2013

Revised Date: 10 October 2013

Accepted Date: 29 October 2013

Please cite this article as: Kroman SL, Roos EM, Bennell KL, Hinman RS, Dobson F, Measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee osteoarthritis: A systematic review, *Osteoarthritis and Cartilage* (2013), doi: 10.1016/j.joca.2013.10.021.

Measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee osteoarthritis: A systematic review.

Authors: Sarah L. Kroman[1], Ewa M. Roos[1], Kim L. Bennell[2], Rana S. Hinman[2], Fiona Dobson[2].

Affiliations:

[1] Institute of Sports Science and Clinical Biomechanics, The University of Southern Denmark, Denmark. Sarah_kroman@hotmail.com, eroos@health.sdu.dk.

[2]Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne, Australia. fdobson@unimelb.edu.au, ranash@unimelb.edu.au, k.bennell@unimelb.edu.au.

Corresponding author:

Fiona Dobson. Centre for Health, Exercise and Sports Medicine, Department of Physiotherapy, School of Health Sciences, The University of Melbourne, Alan Gilbert Building 161 Barry St Carlton Vic, 3010, Australia. Ph: +61 3 8344 3642 Fax: +61 3 8344 3771 fdobson@unimelb.edu.au

Running title:

Performance measures when at risk of OA.

ABSTRACT

Objectives: To systematically appraise the evidence on measurement properties of performance-based outcome measures to assess physical function in young and middle-aged people known to be at high risk of hip and/or knee OA.

Methods: Electronic searches were performed in MEDLINE, CINAHL, Scopus and SPORTDiscus in May 2013. Two reviewers independently rated the measurement properties using the 4-point COSMIN checklist. Best evidence synthesis was made using COSMIN quality, consistency and direction of findings and sample size.

Results: Twenty of 2736 papers were eligible for inclusion and 24 different performance-based outcome measures knee or obese populations were evaluated. No tests related to hip populations were included. Twenty-five measurement properties including reliability (9 studies), construct validity (hypothesis testing) (9 studies), measurement error (3 studies), structural validity (2 studies), interpretability (1 study) and responsiveness (1 study) were evaluated. A positive rating was given to 12,5% (30/240) of all possible measurement ratings. Tests were grouped into two categories based on the population characteristics. The *one-legged hop for distance,* followed by the *6-meter timed hop* and *cross over hop for distance*

were the best-rated tests for the knee-injured population. Whereas the *six-minute walk test* was the only included test for the obese population.

Conclusion: This review highlights the many gaps in knowledge about the measurement properties of performance-based outcome measures for young and middle-aged people known to be at high risk of hip and/or knee OA. There is a need for consensus on which outcome measures should be used and/or combined when assessing physical function in this population. Further good quality research is required.

Keywords: Physical function; performance-based outcome measures; measurement properties; systematic review; knee-injury; obesity.

Introduction

Robust, population-specific, health outcome measures are an integral part of clinical research and practice. For people with osteoarthritis (OA), outcomes that specifically assess pain and function are advised [1]. Although the results from self-reported questionnaires and performance tests of physical function are related they measure two distinct different constructs, and assessment of both self-reported and objectively measured physical function is recommended[2-4]. In light of this, a set of performance-based measures to assess physical function in people diagnosed with hip and/or knee OA was recently recommended as complementary to questionnaire data by an international expert advisory group and endorsed by the Osteoarthritis Research Society International (OARSI) [5]. The recommended set for older people with established OA contains three core tests (30-s chair stand test, 40m fast paced walk test, a stair test) and two additional tests (timed up and go test, six-minute walk test). The recommended set of tests was selected based on global expert opinion, feasibility and available measurement property evidence [6]. The tests were considered to require moderate levels of functional demands in that they were neither too easy nor too hard for older people with OA.

Two of the most common populations known to be at high risk of developing OA are people with a previous hip and/or knee injury and people who are obese[7-15]. A recent meta-analysis that included approximately 21,000 people found a history of knee injury to be associated with a four-fold increased risk of developing knee OA[7]. Obese or overweight persons have nearly 3 times the risk of incident knee OA

compared with those who are of normal weight[8] and this risk accumulates from exposure to a high body mass index from a young age[15]. Appropriate assessment of physical function and its response to treatment in people who present with ongoing knee pain, such as following knee injury or in obese patients, is essential to clinical practice. Since joint injury commonly is sustained during sports these patients often have a higher than average physical activity level while obese patients more commonly have a lower than average physical activity level. These potential differences in physical activity indicate other tests than for those with established OA may be preferable for people at risk of future OA. Additionally, the measurement evidence for physical performance tests for subgroups at risk of OA has yet to be examined.

The aim of this study was to systematically appraise the evidence on the measurement properties of performance-based outcome measures for assessing physical function in young and middle-aged people known to be at high risk of hip and/or knee OA. It was not intended to capture or appraise measurement-property evidence of diagnostic tests that predict the risk factors of developing OA. This review will be used to assist a planned consensus of opinion of a core set of outcome measures that are specific and relevant to younger populations known to be at risk of hip and/or knee OA. Recommendation of a core set/test will encourage clinicians and researchers to more consistently select tests to measure performance-based change in this population. Findings should be seen as complementary to our previous related review [6] which was used to recommend tests for older people with established OA [5].

Methodology

Information sources and 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-analysis (PRISMA) guidelines [16]. With the exception of the target population, a similar methodology was adopted as that described in our previous review [6]. As the present review was targeted to younger people at risk of OA, two distinct groups were pre-identified for searching criteria. These were people with previous hip and/or knee injuries/deformities and people with obesity.

Electronic searches were performed in MEDLINE via PubMed, CINAHL via EBSCO, Scopus via Elsevier and SportDISCUS via EBSCO up to end of May 2013. In addition, a manual search was performed of reference lists from all relevant articles.

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 AND injuries OR surgeries OR deformities OR obesity
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 incorporated recommendations for performing systematic reviews of measurement properties [17] and is more fully described in Appendix 1. For MEDLINE (PubMed), we adopted a measurement property search filter shown to retrieve more than 97% of publications related to measurement properties [18].

Eligibility criteria

Studies were initially screened by title and abstract to filter any articles that were clearly irrelevant. Remaining articles were screened by two independent reviewers followed by an independent full-text review of potentially eligible studies. Any disagreements on eligibility were discussed and resolved with an experienced third reviewer (FD). 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* [19]. If the test was a battery of multi-task items, then at least 80 percent of the items were required to assess activities.
2. **Target population:** the study population was people known to be at high risk of hip and/or knee OA defined as young or middle aged (≥ 80 percent of the study population should be ≥18-60 years of age) with ≥80 percent having a

history of hip and/or knee injuries/deformities (incl. concomitant/several injuries and surgeries) and/or obesity (at least 80 percent should have a BMI ≥ 30).

3. **Measurement instrument:** the measure was a quantitative performance-based test evaluated by an observer as the individual is performing the activity e.g. by timing, counting or distance measures.
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 clinimetric properties of a performance-based measure (e.g. internal consistency, reliability, validity, responsiveness, 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 performance-based measure was only compared to non-performance-based measures e.g. self-reported measures; (3) the test predominately targeted the ICF level of impairment or health related quality of life; (4) treatment effectiveness was evaluated without a specific aim to study the measurement properties of performance measures; (5) the test required expensive sophisticated equipment such as 3-dimensional gait analysis or accelerometers not relevant to most clinical/research settings; (6) published only as 'grey literature' such as scientific meeting abstracts, dissertations or unpublished literature; and (7) 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 [20]. Two raters independently assessed the quality of included studies using the four-point scored COSMIN checklist [21]. This standardized and validated tool contains multiple items to assess the quality of individual measurement properties such as reliability, measurement error, validity, responsiveness and overall interpretability. Each measurement property is allocated a separate rating (excellent, good, fair or poor) using the "worse-score counts method" for items assessed under the property.

Evaluation of the measurement property results

In addition to a methodological quality evaluation with COSMIN, the quantitative results for each measurement property was assessed using a criteria [22]. These criteria consist of positive (+), indeterminate (?) and negative ratings (-) for each measurement property as defined in Appendix 2.

Best evidence synthesis: levels of evidence

To synthesize the evidence, "a best evidence synthesis" [17] was performed by consensus of two reviewers using the criteria outlined in Appendix 3. The possible levels of evidence for a measurement property are "strong", "moderate", "limited", "conflicting" and "unknown" (Appendix 3). These levels were derived using the methodological quality of the studies (COSMIN score), the rating and consistency of the measurement property result from Appendix 2 (+/?/-), the number of related

studies evaluating each measurement property and sample size. Multiple studies were only combined when the same variation of the performance-based measure was evaluated for the same measurement property, that is they were comparable with regards to activity and procedure, and they had comparable populations in regards to diagnosis. As recommended, measurement properties from studies that were rated "poor" on the COSMIN were not eligible to be combined in evidence synthesis[17].

Following recommendations by the developers of the COSMIN tool and similar procedures applied to performance-based measurement properties employed in our previous review [6], the sample size item was not used to determine the "worst score counts" for the COSMIN quality score, but instead was accounted for at the evidence synthesis stage. This process, similar to those employed in meta-analysis, enabled the majority of otherwise fair, good or excellent studies to be considered for evidence synthesis, but still enabled sample size to inform the final level of evidence. Evidence could only be assigned as: "strong" when consistent findings were found in multiple good or at least one excellent quality study and the total sample size of eligible combined studies was ≥100; "moderate" when consistent findings in multiple fair or one good quality study with a total sample ≥50, or at least one good or excellent quality study with a total sample of 50-99; "limited" when findings were found in at least one fair, good or excellent quality study and the total samples between 25-49; and "unknown" when findings were of indeterminate rating, in studies with poor methodological quality or with a sample of ≤ 25.

Results

Description of included studies and performance-based measures

Twenty eligible studies were identified and are described in Table 1. The selection procedures are summarized in Figure 1.

Measurement properties from 23 single-activity measures were investigated in 19 studies [23-41] and one multi-activity measure was investigated in one study[42]. Fourteen studies included participants with ACL reconstructions [24,28-31,33-37,39-42], three with ACL injuries [27,38,42], two with meniscectomy [23,27], one with patellofemoral pain syndrome (PFPS) [25] and one following knee arthroscopy [27]. These studies were grouped as they were all related to knee injuries. Two studies included obese participants [26,32] and formed a second group. No hip-related population was identified, whereas two studies included a population with more than one type of knee injury [27,42].

The knee-injured group included 23 different performance-based tests with a wide range of variations. Tests were grouped for convenience into the following categories: (1) *knee bends* (e.g. number of knee bends on one leg [23]), (2) *sitting to standing* (e.g. one-legged rising from a chair [23]), (3) *stepping* (e.g. step up and down a platform [25], (4) *hopping* (e.g. one-legged hop [23,24,27,28,30,34-37,39-41]), (5) *stairs running* (e.g. run a number of stairs [31]) or (6) *running* (e.g. run a figure-of-eight [31]). There

192 were also variations in performance within each category e.g. one- or two-legged
193 performances [23-42].
194
195 The demands of the activities varied from (1) *slow controlled movements* (e.g. single-
196 leg wall slide [23]) to (2) *fast coordinated movements* (e.g. six-meter timed hop [28-30,39]),
197 (3) *accelerative demanding movements* (e.g. shuttle run test [33,38]) or (4) *repetitive*
198 *movements* (e.g. knee bends [23])
199
200 The method of scoring included (1) *timing* (total time for completion of test,
201 difference in time between involved and uninvolved leg), (2) *counting* (maximum
202 repetitions in a defined time) and measures of (3) *lengths* (maximum distance
203 performed), (4) *heights* (lowest height from succeeded rising, highest jump
204 performed) or (5) *angles* (maximum angle of knee flexion during test) [23-42].
205
206 Required equipment varied in different lengths of walkways, numbers and heights of
207 stairs, stools or step-up boards, accessible walls, a Velcro-belt and measuring
208 instruments such as stopwatch, measuring tape, goniometer and different marks or
209 lines on floors or walls [23-42].
210
211 The *six-minute walk test* was the only test evaluated in the obese group. This test
212 included variations in length of walkway (20 and 40 meters) [26,32].
213
214 ***Measurement properties***
215

Six of the 10 measurement properties outlined in COSMIN were evaluated in the 20 included studies (Table 1). These were reliability [23-26,28-30,32,40] and hypothesis testing [27,33-39,41] in nine studies each, measurement error in three studies [25,30,32], structural validity in two studies [31,42], and interpretability [23] and responsiveness [30] in one study each. Other measurement properties described in COSMIN such as internal consistency, criterion validity, content validity or cross-cultural validity were not assessed in any of the included studies.

The inter-rater agreement of the independent COSMIN ratings was good (absolute agreement = 85%, kappa = 0.77, 95% CI = 0.51-1.0). Disagreements were mainly due to reading errors and were easily resolved. Measurement properties were rated as "excellent" in one study [42], as "good" in eight studies [24-26,28-30,32,40], as "fair" in eight studies" [23,27,31,36-39,41] and as "poor" in four studies [23,33-35].

Knee-injured group

Reliability and measurement error (Table 2)

The *one-legged hop for distance* was the most frequently evaluated test in this group [23,24,27,28,30,34-37,39-41] and was performed with variations in (1) method of measurement (best of two or three trials (cm), mean value of three trials (cm), measure from great toe to heel, and (2) performance (hands behind back or arms free, warm up or not, visible tape measure or not). It was rated positive (i.e. ICC > 0.70) for intra-rater reliability in three "good" and one "fair" quality studies on people with meniscectomy[23] and ACL reconstructions [28,30,40]. Reliability for this test

was also rated positive (ICCs from 0.88 – 0.97) in another "good" quality study [24], however the type of reliability tested was not specified.

Additionally in studies on people with ACL reconstructions, a positive result rating was reported for intra-rater reliability of the *six-meter timed hop* (ICCs from 0.82-0.97) in three "good" quality studies [28-30], the *crossover hop for distance* (ICC 0.84 and 0.98) in two "good" quality studies [29,30] and the *triple hop for distance* (ICC 0.88) in one "good" quality study [30]. In people with PFPS, a positive result rating was reported for intra-rater reliability for the *balance and reach test* (ICC 0.92), the *bilateral squat* (ICC 0.83), the *antero-medial lunge* (ICC 0.82) and *step down test* (0.94) in one "good" quality study [25]. In people with meniscectomy, a positive result rating was reported for intra-rater reliability for the *knee bend test* (ICC 0.92) and the *one-leg rise test* (ICC 0.84) in one "fair" quality study [23].

Measurement error was assessed in eight studies (Table 2). As minimal important changes have not been defined for any of these performance measures, interpretation of the findings for measurement error is undetermined.

Validity studies (Table 3)

In people with ACL reconstructions, the *one-legged hop for distance* was rated positive on hypothesis testing in two "fair" quality studies [39,41]. Two other "fair" quality studies were rated negative [27,36] and one was rated indeterminate[37]. The *six-meter timed hop test* and the *crossover hop for distance* were positive rated for hypothesis testing in one "fair" quality study [39], but the *crossover hop for distance*

was rated negative in another "fair" quality study [36]. The *triple hop for distance* was rated positive for both hypothesis testing and structural validity in two "fair" quality studies [31,41] and the *stairs hopple test* was rated positive for structural validity in one "fair" quality study [31]. The *vertical jump test* was rated positive for structural validity in one "fair" quality study [31] and rated negative for hypothesis testing in another "fair" quality study [36]. The *figure-of-eight* and *stairs-running test* were rated positive in one "fair" quality study on structural validity [31], whereas negative ratings were reported on hypothesis testing for the *co-contraction test*, *shuttle run test* and *carioca test* in one "fair" quality study [38]. For *the test-battery*, an indeterminate rating was reported for structural validity in one "excellent" quality study with ACL reconstructions and ACL injuries [42].

Responsiveness

In studies on people with ACL reconstructions, responsiveness was reported in one "good" quality study for the *one-legged hop for distance, six-meter timed hop, crossover hop for distance* and *triple hop for distance* [30]. All were rated "indeterminate" as a result of the correlations being determined only with unrelated constructs.

Interpretability

Interpretability was reported in 7/24 of the performance measures in one "poor" quality study [23]. As such, evidence could only be rated as indeterminate in evidence synthesis.

Obesity group

The *six-minute walk test* was the only test evaluated in the obese group [26,32]. It was rated positive for intra-rater reliability (ICC 0.96) in one "good" quality study[32] and for test-retest reliability (ICC 0.94) in another "good" quality study[26].

Best evidence synthesis: levels of evidence

A summary of best evidence synthesis is provided in Table 4. Given the large variety of performance-based measures, results were rarely combined. The exceptions were for the *one-legged hop for distance, the six-meter timed hop, crossover hop for distance* and the *6-minute walk test* as the populations and procedures for these tests were comparable across studies (Table 1). Although the *one-legged hop for distance test* was described with two different procedures (arms behind back [23,27,28,30,34,35,39-41] and arms free [24]), this test was collapsed into one group because of missing information about the procedure in two other studies [36,37]. A positive rating was given to 30 out of 240 possible ratings (12.5%).

Discussion

Twenty studies reported the measurement properties of 24 performance-based outcome measures of physical function for young and middle-aged people known to be at high risk of hip and/or knee OA. Overall, a great deal of measurement property evidence is unknown for many of the tests identified in the review either because of

no available information, indeterminate information or because evidence was rated as poor quality. No performance-based test in this review contained evidence on all measurement properties. Intra-rater reliability was most frequently reported and a number of tests had acceptable levels of evidence for this measurement property. This provides useful, but limited, information for clinicians and researchers about which performance-based tests are currently the best for use in the target population. Furthermore, it clarifies the need for more comprehensive studies investigating measurement properties.

For the knee-injured group, the strongest positive evidence was found for the *one-legged hop for distance test* [23,24,28,30,40]. This was followed by moderate positive evidence for the *six-meter timed hop* [28-30] and *crossover hop for distance* [29,30]. Only limited positive evidence was found for the *figure-of-eight, stairs running* and *stairs hopple tests* [31], *the knee bends* and *one-leg rise test* [23], the *six-meter timed hop* [39] and the *triple hop for distance* [30,31,41]. Limited negative evidence was found for the *co-contraction test*, *shuttle run test* and *carioca test* [38]. Remaining possible ratings were "unknown" or not rated because of no available information.

For the obese group the *six-minute walk test* was found to have limited positive evidence on intra-rater reliability[32] and unknown evidence on test-retest reliability[26]. This was the only test identified in the current review that was also a recommended performance-based test for people diagnosed with knee OA [6]. This suggests that it may be a useful candidate test to consider when the objective is to measure across age groups in obese populations.

For 15 of the 20 identified tests, measurement-property evidence was only evaluated in single studies. Given that many gaps in measurement-property evidence still remain for the performance-based tests included in this review, it is difficult to identify the "best" tests specific for people at high risk of developing OA. Thus, it is only appropriate to highlight those tests with the strongest existing evidence, which may therefore change as further evidence comes to hand.

The difference between being "at risk" or "at high risk" of developing lower limb OA is often subtle and therefore not always easy to differentiate. We also cannot preclude that some participants included in studies captured in this review were at an early stage of the disease when tested. The age cut-offs for the target population of this review reflected the intent to capture younger people than those included in the previous review and therefore a different population from one that has already been diagnosed with OA[6]. As age has an impact on development of OA [43,44], it is important to point out that 16 out of 18 studies for the knee-injured group included a population with a mean age ≤ 31 [24,25,27-31,33-40,42]. This characterizes the knee-injured population in this review as young rather than middle-aged. In contrast, the mean age for the obese group was ≥ 46 [26,32] and would be best characterized as middle-aged, whilst the mean age for the participants diagnosed with OA in the previous review [6] was ≥ 60 years and better characterized as an older population.

With the intention to capture tests relevant to younger people at risk of OA, two separate and searchable sub-populations that were known to be at high risk were

purposely targeted. Other groups, such as people who participate in strenuous running[45,46] and sports activity[11,47,48], were not specifically searched, as these more controversial risk factors have been reported to have no or inconsistent associations with OA. The functional demands of the performance-based tests identified for each included population was distinct. In the knee-injured group, the identified performance tests represented higher functional demands than the test described for obese people, and did not overlap with tests identified for people diagnosed with OA in the previous review [6]. By contrast in the obese group, the performance test identified (six-minute walk test) was one of the recommended tests for people already diagnosed with OA and represented a lower functional demand than tests identified in the current knee-injured group.

A number of limitations are acknowledged. Only tests that required non-technical, readily available, inexpensive and portable equipment were eligible for inclusion. A preference for such simple measures that are possible to perform in most settings was made because they are more likely to be implemented in both research and clinical practice. Although the literature search was not limited to those published in the English language, two non-English studies retrieved from the original search were eliminated due to limited translation ability [49,50]. It is possible that further evidence on additional tests (the six-minute walk test, shuttle test and a functional scale for resumption of sports activity (COFRAS scale)) may have been identified in these two studies. The COSMIN quality-scoring system was originally developed for self-report questionnaires. However, it has also shown to be useful for evaluation of non-patient reported measures [6,51]. Consistent with our earlier review [6], it was

modified to enable smaller studies that were otherwise of acceptable quality, to be included in best evidence synthesis. Half of the reliability studies and several validity studies would have been excluded without this modification.

This review highlights a number of areas worthy of future research. Given the shortage of measurement-property evidence for performance-based measures for people at high risk of developing lower limb OA, more quality studies, in particular those that evaluate responsiveness, measurement error and interpretability are required. Importantly, before any recommendations can be made, consensus is still required on which combination of tests best assesses physical function in people at high risk of developing lower limb OA.

Conclusion

A large variety of outcome measures are used for young and middle-aged people known to be at high risk of developing hip and/or knee OA and few measurement properties for these tests have been evaluated. For the knee-injured group the *one-legged hop for distance* was the best-rated test followed by *six-meter timed hop* and *crossover hop for distance*. For the obese group only the *six-minute walk test* was evaluated, with moderate positive evidence on intra-rater reliability. This review highlights current gaps in our knowledge about the measurement properties and need for consensus of opinion on which performance tests should be used when assessing physical function in people known to be at high risk of hip or knee OA.

Author Contribution

SK contributed to the collection and assembly of data, analysis and interpretation of data, and writing of the manuscript and final approval of the article.
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.
EMR, RSH and KLB contributed to conception and design of the study including obtaining of funding, analysis and interpretation of 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. E-mail last author: fdobson@unimelb.edu.au

Acknowledgements

Ms Michelle Hall contributed to collection and assembly of data, including quality assessment using COSMIN.

Role of funding Source

This project was partly funded by the Osteoarthritis Research Society International (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 osteoarthritis.

Sarah L. Kroman was partly funded by the Oticon Foundation, Knud Højgaard Foundation, the council of Faaborg-Midtfyn and the University of Southern Denmark. 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.

Competing interest statement

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 regards to the work.

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 and cartilage / OARS, Osteoarthritis Research Society 2004;12:389-99.
2. Stratford PW KD, Woodhouse LJ. . Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Physical therapy 2006;86(11):1489-96.
3. Faucher M PS, Lefevre-Colau MM, Rannou F, Fermanian J, Revel M. . Algo-functional assessment of knee osteoarthritis: Comparison of the test-retest reliability and construct validity of the Womac and Lequesne indexes. Osteoarthritis and Cartilage 2002;10(8):602-10
4. Thumboo J CL, Soh CH. . Validation of the Western Ontario and McMaster University Osteoarthritis Index in Asians with osteoarthritis in Singapore. Osteoarthritis and Cartilage 2001;9(5):440-46
5. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford PW, Davis AM, et al. OARSI recommended performance-based tests to assess physical function in

464 people diagnosed with hip or knee osteoarthritis. Osteoarthritis and Cartilage
465 Accepted May 2013.
466 6. Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL.
467 Measurement properties of performance-based measures to assess physical
468 function in hip and knee osteoarthritis: a systematic review. Osteoarthritis and
469 cartilage / OARS, Osteoarthritis Research Society 2012;20:1548-62.
470 7. Muthuri SG, McWilliams DF, Doherty M, Zhang W. History of knee injuries
471 and knee osteoarthritis: a meta-analysis of observational studies. Osteoarthritis
472 and cartilage / OARS, Osteoarthritis Research Society 2011;19:1286-93.
473 8. Blagojevic M, Jinks C, Jeffery A, Jordan KP. Risk factors for onset of
474 osteoarthritis of the knee in older adults: a systematic review and meta-analysis.
475 Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society 2010;18:24-
476 33.
477 9. Chaganti RK, Lane NE. Risk factors for incident osteoarthritis of the hip
478 and knee. Current reviews in musculoskeletal medicine 2011;4:99-104.
479 10. Maffulli N, Longo UG, Gougoulias N, Caine D, Denaro V. Sport injuries: a
480 review of outcomes. Br Med Bull 2011;97:47-80.
481 11. Richmond SA, Fukuchi RK, Ezzat A, Schneider K, Schneider G, Emery CA.
482 Are joint injury, sport activity, physical activity, obesity, or occupational
483 activities predictors for osteoarthritis? A systematic review. The Journal of
484 orthopaedic and sports physical therapy 2013;43:515-24.
485 12. Thelin N, Holmberg S, Thelin A. Knee injuries account for the sports-
486 related increased risk of knee osteoarthritis. Scandinavian journal of medicine &
487 science in sports 2006;16:329-33.
488 13. Wang Y, Wluka AE, Simpson JA, Giles GG, Graves SE, de Steiger RN, et al.
489 Body weight at early and middle adulthood, weight gain and persistent
490 overweight from early adulthood are predictors of the risk of total knee and hip
491 replacement for osteoarthritis. Rheumatology (Oxford, England) 2013;52:1033-
492 41.
493 14. Lohmander LS, Englund PM, Dahl LL, Roos EM. The long-term
494 consequence of anterior cruciate ligament and meniscus injuries: osteoarthritis.
495 American Journal of Sports Medicine 2007;35:1756-69.
496 15. Wills AK, Black S, Cooper R, Coppack RJ, Hardy R, Martin KR, et al. Life
497 course body mass index and risk of knee osteoarthritis at the age of 53 years:
498 evidence from the 1946 British birth cohort study. Annals of the rheumatic
499 diseases 2012;71:655-60.
500 16. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for
501 Systematic Reviews and Meta-Analyses: The PRISMA Statement. Journal of
502 Clinical Epidemiology 2009;62:1006-12.
503 17. De Vet CWH, Terwee, C. B., Mokkink, L. B., Knol, D. L. Measurement in
504 Medicine: A practical guide to biostatistics and epidemiology. London:
505 Cambridge University Press; 2011.
506 18. Terwee CB, Jansma EP, Riphagen, II, de Vet HC. Development of a
507 methodological PubMed search filter for finding studies on measurement
508 properties of measurement instruments. Quality of life research : an
509 international journal of quality of life aspects of treatment, care and
510 rehabilitation 2009;18:1115-23.
511 19. World Health Organization. International Classification of Functioning,
512 Disability, and Health: ICF. Geneva, Switzerland 2001.

20. 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. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation 2010;19:539-49.
21. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation 2012;21:651-7.
22. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of clinical epidemiology 2007;60:34-42.
23. 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.
24. Kramer JF, Nusca D, Fowler P, Webster-Bogaert S. Test-retest reliability of the one-leg hop test following ACL reconstruction. Clinical Journal of Sport Medicine 1992;2:240-3.
25. Loudon JK, Wiesner D, Goist-Foley L, Asjes C, Loudon KL. Intrarater reliability of functional performance tests for subjects with patellofemoral pain syndrome. Journal of athletic training 2002;37:256-61.
26. Beriault K, Carpentier AC, Gagnon C, Ménard J, Baillargeon J, Ardilouze J, et al. Reproducibility of the 6-minute walk test in obese adults. International journal of sports medicine 2009;30:725-7.
27. Borsa PA, Lephart SM, Irrgang JJ. Comparison of performance-based and patient-reported measures of function in anterior-cruciate-ligament-deficient individuals. Journal of Orthopaedic & Sports Physical Therapy 1998;28:392-9.
28. Brosky JA, Jr., Nitz AJ, Malone TR, Caborn DNM, Rayens MK. Intrarater reliability of selected clinical outcome measures following anterior cruciate ligament reconstruction. Journal of Orthopaedic & Sports Physical Therapy 1999;29:39-48.
29. Hopper DM, Goh SC, Wentworth LA, Chan DYK, Chau JHW, Wootton GJ, et al. Test-retest reliability of knee rating scales and functional hop tests one year following anterior cruciate ligament reconstruction. Physical Therapy in Sport 2002;3:10-8.
30. Reid A, Birmingham TB, Stratford PW, Alcock GK, Giffin JR. Hop testing provides a reliable and valid outcome measure during rehabilitation after anterior cruciate ligament reconstruction. Physical therapy 2007;87:337-49.
31. Risberg MA, Ekeland A. Assessment of functional tests after anterior cruciate ligament surgery. Journal of Orthopaedic & Sports Physical Therapy 1994;19:212-7.
32. Larsson UE, Reynisdottir S. The six-minute walk test in outpatients with obesity: reproducibility and known group validity. Physiotherapy Research International 2008;13:84-93.
33. Kong DH, Yang SJ, Ha JK, Jang SH, Seo JG, Kim JG. Validation of functional performance tests after anterior cruciate ligament reconstruction. Knee Surg Relat Res 2012;24:40-5.

34. Petschnig R, Baron R. Assessment of quadriceps strength and functional limitations determined by hop tests for distance and a newly designed vertical jump test after anterior cruciate ligament reconstruction. European Journal of Physical Medicine & Rehabilitation 1997;7:81-6.
35. Petschnig R, Baron R, Albrecht M. The relationship between isokinetic quadriceps strength test and hop tests for distance and one-legged vertical jump test following anterior cruciate ligament reconstruction. The Journal of orthopaedic and sports physical therapy 1998;28:23-31.
36. Jamshidi AA, Olyaei GR, Heydarian K, Talebian S. Isokinetic and functional parameters in patients following reconstruction of the anterior cruciate ligament. Isokinetics & Exercise Science 2005;13:267-72.
37. Järvelä T, Kannus P, Latvala K, Järvinen M. Simple measurements in assessing muscle performance after an ACL reconstruction. International journal of sports medicine 2002;23:196-201.
38. Lephart SM, Perrin DH, Fu FH, Minger K. Functional performance tests for the anterior cruciate ligament insufficient athlete. Athletic Training 1991;26:44-5;8-50.
39. Wilk Ke Fau - Romaniello WT, Romaniello Wt Fau - Soscia SM, Soscia Sm Fau - Arrigo CA, Arrigo Ca Fau - Andrews JR, Andrews JR. The relationship between subjective knee scores, isokinetic testing, and functional testing in the ACL-reconstructed knee. The Journal of orthopaedic and sports physical therapy 1994;20(2):60-73.
40. Paterno MV, Greenberger HB. The test-retest reliability of a one legged hop for distance in young adults with and without ACL reconstruction. Isokinetics and Exercise Science 1996;6:1-6.
41. Andrade MS, Cohen M, Piçarro IC, Silva AC. Knee performance after anterior cruciate ligament reconstruction. Isokinetics and Exercise Science 2002;10:81-6.
42. Gustavsson A, Neeter C, Thomeé P, Grävare Silbernagel K, Augustsson J, Thomeé R, et al. A test battery for evaluating hop performance in patients with an ACL injury and patients who have undergone ACL reconstruction. Knee Surgery, Sports Traumatology, Arthroscopy 2006;14:778-88.
43. Loeser RF. Age-related changes in the musculoskeletal system and the development of osteoarthritis. Clinics in geriatric medicine 2010;26:371-86.
44. Loeser RF. Aging and osteoarthritis. Curr Opin Rheumatol 2011;23:492-6.
45. Hansen P, English M, Willick SE. Does running cause osteoarthritis in the hip or knee? PM & R : the journal of injury, function, and rehabilitation 2012;4:S117-21.
46. Chakravarty EF, Hubert HB, Lingala VB, Zatarain E, Fries JF. Long distance running and knee osteoarthritis. A prospective study. American journal of preventive medicine 2008;35:133-8.
47. McAlindon TE, Wilson PW, Aliabadi P, Weissman B, Felson DT. Level of physical activity and the risk of radiographic and symptomatic knee osteoarthritis in the elderly: the Framingham study. The American journal of medicine 1999;106:151-7.
48. Papavasiliou KA, Kenanidis EI, Potoupnis ME, Kapetanou A, Sayegh FE. Participation in athletic activities may be associated with later development of hip and knee osteoarthritis. The Physician and sportsmedicine 2011;39:51-9.

609 49. Oria S. The Cofras score: assessment of knee function before resuming
610 sports activities, study of the score's internal structure [French]. Kinesitherapie
611 Revue 2010:27-31.
612 50. Damasceno, Freitas Dantas L, Júnior AB, Vicente Franco L, Malosá, Costa
613 D. Evaluation of physical performance and functional breathing in obese people
614 [Portuguese]. Fisioterapia em Movimento 2011;24:697-704.
615 51. Granger CL, McDonald CF, Parry SM, Oliveira CC, Denehy L. Functional
616 capacity, physical activity and muscle strength assessment of individuals with
617 non-small cell lung cancer: a systematic review of instruments and their
618 measurement properties. BMC cancer 2013;13:135.
619
620

Appendix 1. Search strategy

Filter 1: Construct Terms
("motor activity"[MH] OR "physical activity"[tw] OR "physical activities"[tw] OR "physical function*"[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
(((((((((hip injur*[Title/Abstract])) OR (hip injuries[MeSH Terms]))) OR ((((knee injur*[Title/Abstract])) OR (knee injur*[MeSH Terms]))) OR ((((medial* AND collateral ligament* AND knee[Title/Abstract])) OR (tibial* AND collateral ligament*[Title/Abstract])) OR (Medial collateral ligament, knee[MeSH Terms])) OR (((Patellofemoral pain syndrome[Title/Abstract])) OR (Patellofemoral pain syndrome[MeSH Terms])) OR ((((anterior* AND cruciate ligament*[Title/Abstract])) OR (ACL[Title/Abstract])) OR (Anterior cruciate ligament[MeSH Terms])) OR (((Posterior* AND cruciate ligament*[Title/Abstract])) OR (Posterior cruciate ligament[MeSH Terms])) OR (((((semilunar cartilage*[Title/Abstract])) OR (lateral* AND menisc*[Title/Abstract])) OR (Menisc* AND tibial*[Title/Abstract])) OR (Menisci, tibial[MeSH Terms])) OR ((((genu varu*[Title/Abstract])) OR (bow leg*[Title/Abstract])) OR (genu varum[MeSH Terms])) OR (((femoracetabular impingement*[Title/Abstract])) OR (femoracetabular impingement[MeSH Terms])) OR ((((genu valgum[MeSH Terms])) OR (genu valg*[Title/Abstract])) OR (knock knee*[Title/Abstract])) OR ((((anterior* AND cruciate ligament* AND reconstruction[Title/Abstract])) OR (ACL reconstruction[Title/Abstract])) OR (Anterior cruciate ligament reconstruction[MeSH Terms])))))) OR (obesity[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 intertechnician[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 interparticipant[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 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 ≥ 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

MIC minimal important change, SDC smallest detectable change, LoA limits of agreement, ICC intraclass correlation coefficient, DIF differential item functioning, AUC area under the curve.
+ positive rating, ? indeterminate rating,- negative rating

Adapted from Terwee et al. J Clin Epidemiol. 2007;60(1):34-42.

Appendix 3 Levels of evidence for the quality of the measurement property

Level	Rating†	Criteria
Strong	**+++ or ---**	Consistent findings in multiple good or at least one excellent quality study and a total sample of ≥ 100.
Moderate	**++ or --**	Consistent findings in multiple fair or one good quality study with a total sample of ≥ 50 or at least one good or excellent quality study with a total sample of 50-99.
Limited	**+ or -**	Findings in studies of fair, good or excellent methodological quality with a total sample of 25-49.
Conflicting	**+/-**	Conflicting findings
Unknown	**?**	Only findings of indeterminate rating, studies of poor methodological quality or with a total sample of ≥ 25.

† + = positive rating, - = negative rating

Adapted from Terwee et al. J Clin Epidemiol. 2007;60(1):34-42.

LIST OF FIGURES

Table 1 Characteristics of included studies

Author (Year)	Sample size	Mean age ± SD (range)	Population type	Performance measure	Activity	Quantification measure	Equipment required	Measurement property assessed
Knee-injured group								
Andrade 2002[38]	14 f/0	32 ± 18	ACLR 4-8 mo. post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Hypothesis testing
				Triple hop	Three hops on one leg	Length (cm)	MT	Hypothesis testing
Borsa 1998[24]	29 f/14	28,7 ± 1,7	ACLI, A-SCOPY, MEN. 41,7 ± 11,7 mo. Post-inj.	One-legged hop	Hop on one leg	Length (cm)	MT	Hypothesis testing
Bremander 2007[20]	285 f/59	54 ± 11,2	MEN. 15-21 y post-op.	One-leg rising		Height (cm)	A bench	Intra-reliability Interpretability
				One-legged hop	Hop on one leg	Length (cm)	A stool	Intra-reliability Interpretability
				Single-leg wall slide	Slide down against a wall	Angles (˚)	A free wall	Interpretability
				Heel raising	Number of heel raises	Repetitions	Step-up board	Interpretability
				Lateral step-Up	Step up and down	Repetitions	MT Step-up board	Interpretability
				Climbing a stool	Climbing up/down on stool.	Height (cm)	Goniometer, Stool	Interpretability
				Knee bends	Knee bends	Repetitions	SW	Intra-reliability Interpretability
Brosky 1999[25]	15 f/0	26 ± 7,3	ACLR 15-39 mo. Post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Intra-reliability
				6 meter timed hop	Hop on one leg for 6 m.	Time (sec)	SW	Intra-reliability
Gustavsson 2006*[39]	Injury: 30 f/12	31 ± 9	ACLI 4-44 mo. Post-inj.	One-legged hop	Hop on one leg	Length (cm)	MT	Structural validity
				Vertical jump	Vertical jump on both legs	Heigth (cm)	MT	
	Surgery: 35 f/10	27 ± 7	ACLR 6 mo. Post-op.	Drop jump	Jump down, then double hop	Length (cm)	MT	
				Square hop		Repetitions	Square	
				Side hop	Jump in and out of square Side to side jump	Repetitions	Marks on floor	
Hopper 2002[26]	19 f/6	26,8 ± 8,4	ACLR 6-18 mo. Post-op.	6 meter timed hop	Hop on one leg for 6 m.	Length (cm)	MT	Intra-reliability
				Crossover hop	3 hops on one leg across a line	Length (cm)	MT, line	Intra-reliability
				Stair hop	Hop three steps up and down	Time (sec)	Three steps, SW	Intra-reliability
Jamshidi 2005[33]	11 f/?	30 ± 8	ACLR > 6 mo. Post-op.	One-legged hop	NS	(cm)	MT	Hypothesis testing
				Crossover hop	NS	(cm)	MT	Hypothesis testing
				Vertical jump	NS	(cm)	MT	Hypothesis testing
Järvelä 2002[34]	86 f/21	30,4 ± 9,4	ACLR 5-9 yrs. post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Hypothesis testing
Kong 2012[30]	30 f/0	23,4 ± 3,1	ACLR ≥ 6 mo. post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Hypothesis testing
				Cocontraction test	Run with Velcrobelt	Time (sec)	Velcrobelt, SW	Hypothesis testing
				Shuttle run test	Run back and forth (6,1 m)	Time (sec)	Marked way, SW	Hypothesis testing
				Carioca test	Run laterally (2 x 12 m)	Time (sec)	Marked way, SW	Hypothesis testing

Kramer 1992[21]	38 f/16	25 ± 7	ACLR 15-29 yrs. post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Reliability (NS)
Lephart 1991[35]	41 f/9	22,7 (16-32)	ACLI 10-36 mo. Post-inj.	Concontraction test	Run with Velcrobelt	Time (sec)	Velcrobelt, SW	Hypothesis testing
				Shuttle run test	Run back and forth (6,1 m)	Time (sec)	Marked way, SW	Hypothesis testing
				Carioca test	Run laterally (2 x 12 m)	Time (sec)	Marked way, SW	Hypothesis testing
Loudon 2002[22]	15 f/max. 15	27,6 ± 5,3	PFPS Average 5,2 mo.	Anteromedial lunge	Lunge forward to 90° flex	Length (cm)	MT, SW	Intra-reliability Measurement error
				Step-down	Step forward and down	Repetitions	Step-up board, SW	Intra-reliability Measurement error
				Bilateral squat	Lower body to 90° knee flex	Repetitions	MT, SW	Intra-reliability Measurement error
				Balance and reach	One step forward, only heel	Repetitions	MT, SW	Intra-reliability Measurement error
Paterno 1996[37]	13 f/3	22,4 ± 3,2	ACLR 20-52 wks. Post-op	One-legged hop	Hop on one leg	Length (cm)	MT	Intra-reliability
Petschnig 1997*[31]	62 f/0	Gr. B 28,4 ± 1,6 Gr. C 29,3 ± 1,1	ACLR 12-14 wks. Post-op. 51-59 wks. Post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Hypothesis testing
				Triple hop	Triple hop on one leg	Length (cm)	MT	Hypothesis testing
Petschnig 1998*[32]	55 f/0	Gr B. 27,8 ± 9 Gr. C 29,9 ± 5,8	ACLR 10-16 wks. Post-op. 41-67 wks. Post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Hypothesis testing
				Triple hop	Triple hop on one leg	Length (cm)	MT	Hypothesis testing
Reid 2007[27]	42 f/19	25,6 ± 9,2	ACLR ≤ 16 wks. Post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Intra-reliability
				6 meter timed hop	Hop on one leg for 6 m.	Time (sec)	6 m walk way, SW	Measurement error
				Triple hop	Triple hop on one leg	Length (cm)	MT	Responsiveness
				Crossover hop	3 hops on one leg across a line	Length (cm)	A marked line, MT	
Risberg 1994[28]	35 f/max. 18	26 ± 8,2	ACLR 14-22 mo. Post-op.	Vertical jump	Vertical jump on both legs	Height (cm)	MT, wall	Structural validity
				Figure-of-eight	Run in a figure of eight	Time (sec)	SW, two circles	Structural validity
				Stairs running	Up/down 55 steps	Time (sec)	Staircase, SW	Structural validity
				Triple hop	Triple hop on one leg	Length (cm)	MT	Structural validity
				Stairs hopple	Up/down 22 steps	Time (sec)	Staircase, SW	Structural validity
Wilk Ke Fau 1994[36]	50 f/16	24,5 (15-52)	ACLR 21-30 wks. Post-op.	One-legged hop	Hop on one leg	Length (cm)	MT	Hypothesis testing
				6 meter timed hop	Hop on one leg for 6 m.	Time (sec)	6 m walk way, SW	Hypothesis testing
Obese group								
Larsson 2008[29]	43 f/27	46,8 (21-62)	Obesity	6 min. walk test	6 min. walk	Distance (m)	SW, walk way	Intra-reliability Measurement error
Beriault 2009[23]	21 f/15	55,7 ± 11,2	Obesity	6 min. walk test	6 min. walk	Distance (m)	Marked walk way SW	Test-retest reliability

(ACLI) ACL injury, (ACLR) ACL reconstruction, (A-SCOPY) arthroscopy, (MEN.) meniscectomy, (PFPS) patellofemoral pain syndrome, (F) Number of females, (NS) not specified, (post-op.): post-operated, (post-inj.): post-injured, (wks.) weeks, (Mo.) Months, (yrs.) years, (*) both groups performed all tests, (MT) measuring tape, (SW) stopwatch.

Table 2 Results of measurement properties of performance-based measures (reliability and measurement error)

	Reliability					Measurement error	
Performance-based measure	Study n	Results	Design	Time interval	COSMIN score	Results	COSMIN score
Knee-injured Group							
Knee bends[20]	50	ICC $_{2.1}$: 0.92 (0.86-0.96)	Intra-rater	45 min	Fair	-	
Single-leg wall slide[20]		-				-	
Heel raise[20]		-				-	
Balance and reach[22]	15	ICC $_{3.1}$: 0.83	Intra-rater	48-72 hours	Good*	SEM 0.68	Good*
Bilateral squat[22]	15	ICC $_{3.1}$: 0.79	Intra-rater	48-72 hours	Good*	SEM 0.47	Good*
Lateral step-up[20]		-				-	
One-leg rise[20]	30	ICC $_{2.1}$: 0.84 (0.69-0.92)	Intra-rater	45 min.	Fair	-	
Climbing a stool[20]		-				-	
Antero-medial lunge[22]	15	ICC $_{3.1}$: 0.82	Intra-rater	48-72 hours	Good*	SEM 0.38	Good*
Step-down[22]	15	ICC $_{3.1}$: 0.94	Intra-rater	48-72 hours	Good*	SEM 0.53	Good*
One-legged hop for distance (**)[20, 21, 25, 27, 37]	30	ICC $_{2.1}$: 0.93 (0.87-0.97)	Intra-rater	45 min.	Fair	-	
	38	ICC $_{2.1}$: 0.93	Not specified	< 5 days	Good*	-	
	15	ICC: 0.88-0.97	Intra-rater	Initial test, 1 day, 1 week and 2 weeks	Good*	-	
	13	ICC $_{2.1}$: 0.89	Intra-rater	> 24 hours	Good*	-	
	42	ICC $_{2.1}$: 0.92 (low end 0.87)	Intra-rater	>24 hours	Good*	SEM: 3.49% (4.37 upper 95%CI) MDC 90: 8.09%	Good*
6 meter timed hop[25-27]	15	Between and within session reliability (0.88-0.97)	Intra-rater	Initial test, 1 day, 1 week, 2 weeks	Good*	-	
	19	ICC $_{3.1}$: 0.96	Intra-rater	7 days	Good*		
	42	ICC $_{2.1}$: 0.82 (low end 0.70)	Intra-rater	>24 hours	Good*	SEM: 5.59% (7.01 upper 95%CI) MDC 90%: 12.96%	Good*
Crossover hop for distance[26, 27]	42	ICC $_{2.1}$: 0.84 (low end 0.74)	Intra-rater	>24 hours	Good*	SEM: 5.28% (6.62 upper 95%CI) MDC 90%: 12.25%	Good*
	19	ICC $_{3.1}$: 0.98	Intra-rater	7 days	Good*	-	
Triple hop for distance[27]	42	ICC $_{2.1}$: 0.88 (low end 0.80)	Intra-rater	>24 hours	Good*	SEM: 4.32% (5.41 upper 95%CI) MDC 90: 10.02%	Good*
Vertical jump[28, 33]		-				-	
		-				-	

Stair hop[26]	19	ICC $_{3.1}$: 0.96	Intra-rater	7 days	Good*	-	
Stairs hopple test[28]		-				-	
Co-contraction test[30, 35]		-				-	
		-				-	
Shuttle run test[33,34]		-				-	
		-				-	
Carioca test[33,34]		-				-	
		-				-	
Figure-of-eight[28]		-				-	
Stairs-running test[28]		-				-	
Test battery[39]		-				-	
6 min. walk test[23, 29]	21	ICC: 0.94 (0.87-0.97)	Test-retest	>2 hours	Good*	-	
	43	ICC $_{1.1}$: 0.96	Intra-rater	5 (2-14) days	Good*	Coefficient of variation (CV) 4.7	Good*

* Denotes a change of COSMIN score after removal of sample size item from the rating.
** More than one reported procedure of the test (arms behind back/arms free when performing the test).

Table 3 Results of measurement properties of performance-based measures (validity, responsiveness and interpretability)

		Validity			Responsiveness			Interpretability	
Performance-based measure	**Study n**	**Design**	**Result**	**COSMIN score**	**Treatment**	**Result**	**COSMIN score**	**Result (%)**	**COSMIN score**
Knee bends[20]	277		-			-		f/c:3/0	Poor
Single-leg wall slide[20]	160		-			-		f/c: 4/33	Poor
Heel raise[20]	153		-			-		f/c:5/0	Poor
Balance and reach[22]			-			-		-	
Bilateral squat[22]			-			-		-	
Lateral step-up[20]	153		-			-		f/c: 5/0	Poor
One-leg rise[20]	160		-			-		f/c: 2/1	Poor
Climbing a stool[20]	157		-			-		f/c: 8/69	Poor
Antero-medial lunge[22]			-			-		-	
Step-down[22]			-			-		-	
One-legged hop for distance[20,24,27,31-34,36,38]	50	Hypo. Test.	Positive correlation with knee extensor peak torque at 180°/sec (r=0.62 P<0.003)	Fair		-		-	
	86	Hypo. Test.	A correlation was seen between extension deficit and hop performance (p= <0.001-0.003)	Fair		-		-	
	11	Hypo. Test.	No significant correlation with isokinetic strength.	Fair*		-		-	
	62	Hypo. Test.	Correlation with peak torque (r=0.3785-0.5019)	Poor		-		-	
	39		-		Rehab. Program	Corr. with GRC 0,48 (range 0.26-0.58)	Good*	-	
	29	Hypo. Test.	Negative correlation with strength index (r=0.06)	Fair*		-		-	
	202					-		f/c: 9/0	Poor
	55	Hypo. Test.	Positive (significant) correlation with peak torque (r=0.45-0.51)	Poor		-		-	
	14	Hypo. Test.	Positive correlation with quadriceps performance (r=0.51-0.63)	Fair*		-		-	
6 meter timed hop[27,36]	50	Hypo. Test.	Positive correlation with knee extensor peak torque at 180°/sec (r=0.60 P<0.001)	Fair		-		-	
	39		-		Rehab. Program	Corr. with GRC 0.46 (0.26-0.58)	Good*	-	
Crossover hop for	50	Hypo. Test.	Positive correlation with knee	Fair		-		-	

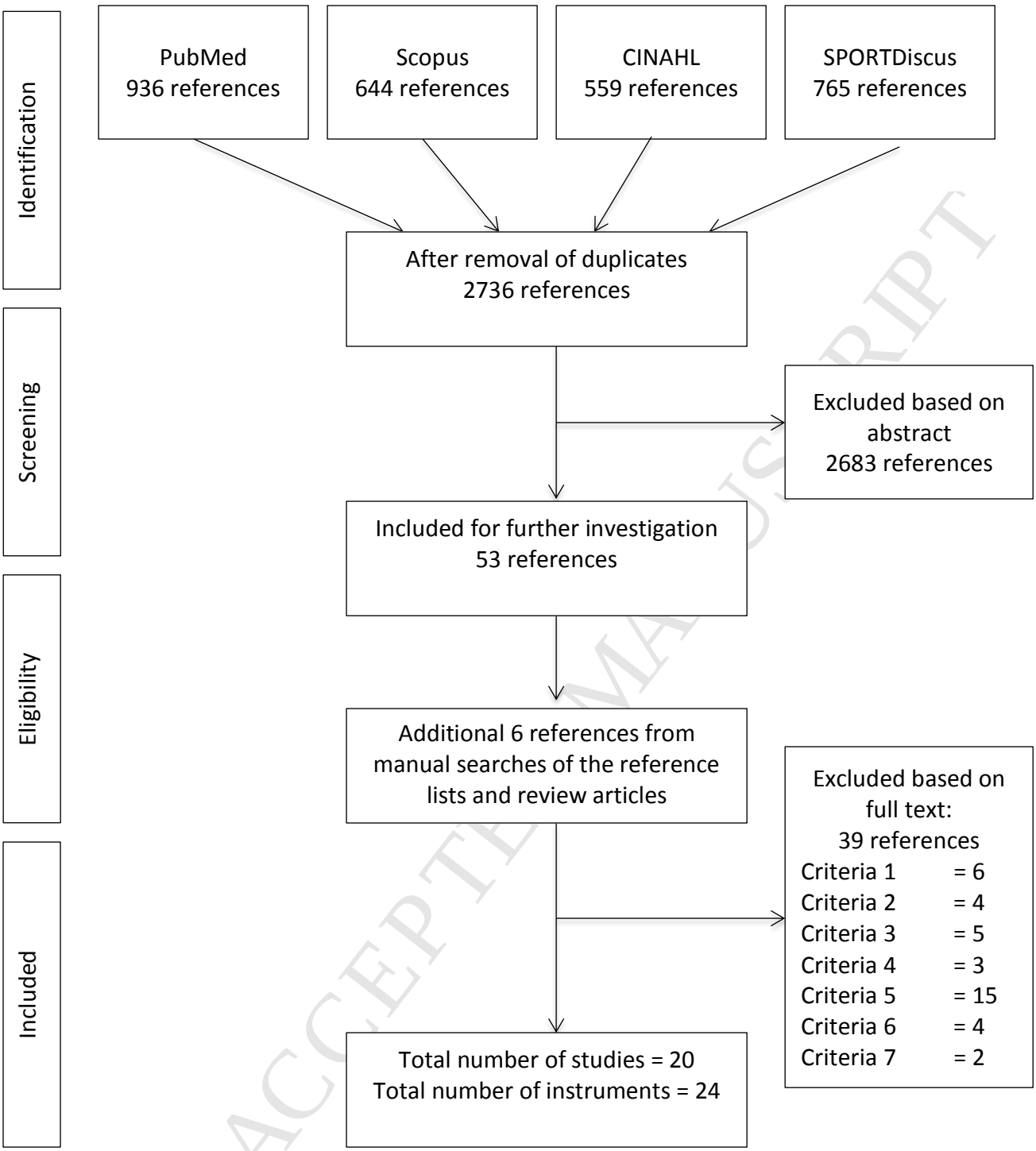
distance[27,33,36]			extensor peak torque at 180°/sec (r=0.69 P<0.0001)					
	39		-		Rehab. Program	Corr. with GRC 0.45 (0.26-0.58)	Good*	-
	11	Hypo. Test.	No significant correlation with isokinetic strength	Fair*		-		-
Triple hop for distance[27,28,31,32,38]	39		-		Rehab. Program	Corr. with GRC 0.44 (0.26-0.58)	Good*	-
	62	Hypo. Test.	Correlation with peak torque (r=0.3565-0.3937)	Poor		-		-
	55	Hypo. Test.	Positive (significant) correlation with peak torque (r=0.48-0.55).	Poor		-		-
	37	Structural	Factor 1: 0.33 Factor 2: 0.86	Fair		-		-
	14	Hypo. Test	Positive correlation with quadriceps performance (r=0.51-0.63)	Fair*		-		-
Vertical jump[28,33]	11	Hypo. Test	No significant correlation with isokinetic strength	Fair*		-		-
	37	Structural	Factor 1: -0.79 Factor2: 0.02	Fair		-		-
Stair hop[26]	-		-			-		-
Stairs Hopple test[28]	37	Structural	Factor 1: 0.03 Factor 2: 0.88	Fair		-		-
6 min. walk test[23,29]	-		-			-		-
Co-contraction test[30,35]	41	Hypo. Test	Lack of any strong correlation with isokinetic strength (r=-0.31-0.31)	Fair		-		-
	30	Hypo. Test	Positive correlation with isokinetic strength(r=0.499, p<0.008)	Poor		-		-
Shuttle run test[30,35]	41	Hypo. Test	Lack of any strong correlation with isokinetic strength (r=-0.32-0.08)	Fair		-		-
	30	Hypo. Test	Positive correlations with isokinetic strength(r=0.268, p<0.177)	Poor		-		-
Carioca test[30,35]	41	Hypo. Test	Lack of any strong correlation with isokinetic strength (r=-0.20-0.05)	Fair		-		-
	30	Hypo. Test	Positive correlation with isokinetic strength(r=0.297, p<0.132)	Poor		-		-
Figure-of-eight[28]	37	Structural	Factor 1: 0.83 Factor 2: 0.37	Fair		-		-
Stairs-running test[28]	37	Structural	Factor 1: 0.86 Factor 2: 0.22	Fair		-		-
Test battery[39]	65	Structural	Factor 1: Max. hop tests (vertical, hop for dist., drop jump with double hop) Factor 2: Endurance hop tests (square hop and side hop)	Excellent		-		-

* Denotes a change of COSMIN score after removal of sample size item from the rating.
(Hypo. Test.) Hypothesis testing, (GRC) global rating of change, (f/c) floor and ceiling effects.

Table 4 Levels of evidence of included performance-based measures

Performance-based measure	Reliability			Measurement Error		Validity	Responsiveness	Interpretability
	Intra	Inter	Retest		Structural	Hypothesis testing		
Strong evidence								
One-legged hop for distance[20,21,24,25,27,31-34,36-38]	+++ (k)	0	0	?	0	+/- (k)	?	0
Moderate evidence								
6 meter timed hop[18,19,36,30]	++ (k)	0	0	?	0	+ (k)	?	0
Crossover hop for distance[19,21,23,30]	++ (k)	0	0	?	0	+/- (k)	?	0
Limited positive evidence								
6 min walk test[23,29]	+ (o)	0	?	**?**	0	0	0	0
Triple hop for distance[27,28,31,32,38]	+ (k)	0		?	+ (k)	+ (k)	?	0
Knee bends[20]	+ (k)	0	0	0	0	0	0	0
One-leg rise[20]	+ (k)	0	0	0	0	0	0	0
Vertical jump[28,33]	0	0	0	0	+ (k)	?	0	0
Stairs hopple test[28]	0	0	0	0	+ (k)	0	0	0
Figure-of-eight[28]	0	0	0	0	+ (k)	0	0	0
Stairs-running test[28]	0	0	0	0	+ (k)	0	0	0
Limited negative evidence								
Co-contraction test[30,35]	0	0	0	0	0	- (k)	0	0
Shuttle run test[33,34]	0	0	0	0	0	- (k)	0	0
Carioca test[33,34]	0	0	0	0	0	- (k)	0	0
Unknown evidence								
Balance and reach[22]	?	0	0	?	0	0	0	0
Bilateral squat[22]	?	0	0	?	0	0	0	0
Antero-medial lunge[22]	?	0	0	?	0	0	0	0
Step down[22]	?	0	0	?	0	0	0	0
Stair hop[26]	?	0	0	0	0	0	0	0
Test Battery[39]	0	0	0	0	?	0	0	0
Heel raise*[20]	0	0	0	0	0	0	0	?
Single-leg wall slide*[20]	0	0	0	0	0	0	0	?
Climbing a stool*[20]	0	0	0	0	0	0	0	?
Lateral step-up*[20]	0	0	0	0	0	0	0	?

+ = positive rating, - = negative rating; +++ or --- strong evidence, ++ or – moderate evidence, + or – limited evidence, +/- conflicting evidence, ? unknown, 0 no information, (k) = knee, (o) = obesity.
Unknown ratings due to: sample size ≤25 (reliability/hypothesis testing), indeterminate result rating (measurement error/structural validity/responsiveness), poor COSMIN ratings (interpretability).



Criteria 1: Construct: not physical function measure
Criteria 2: Population: not 80% hip/knee injury/surgery and/or obesity and being between the age of 18 and 60
Criteria 3: Instrument: not performance-based
Criteria 4: Clinical test: not a field/clinic test (incl. equipment)
Criteria 5: Measurement study: aim was not to measure a measurement property
Criteria 6: Publication type: not a full article
Criteria 7: Language: not written in English