What tests should be used to assess functional performance in youth and young adults following anterior cruciate ligament or meniscal injury? A systematic review of measurement properties for the OPTIKNEE consensus

Bjørnar Berg [1,2] Anouk P Urhausen [3] Britt Elin Øiestad [4] Jackie L Whittaker [5,6] Adam G Culvenor [7] Ewa M Roos [8] Kay M Crossley [7] Carsten B Juhl [8,9] May Arna Risberg [1,3]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bjsports-2022-105510).

For numbered affiliations see end of article.

Correspondence to
Dr Bjørnar Berg, Division of Orthopaedic Surgery, Oslo University Hospital, Oslo 0424, Norway;
bjornar.berg@studmed.uio.no

Accepted 24 May 2022

Check for updates

© Author(s) (or their employer(s)) 2022. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Berg B, Urhausen AP, Øiestad BE, *et al*. *Br J Sports Med* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/bjsports-2022-105510

ABSTRACT

Objectives To critically appraise and summarise measurement properties of functional performance tests in individuals following anterior cruciate ligament (ACL) or meniscal injury.

Design Systematic review.

Data sources Systematic searches were performed in Medline (Ovid), Embase (Ovid), CINAHL (EBSCO) and SPORTSDiscus (EBSCO) on 7 July 2021.

Eligibility criteria for selecting studies Studies evaluating at least one measurement property of a functional performance test including individuals following an ACL tear or meniscal injury with a mean injury age of ≤30 years. The COnsensus-based Standards for the selection of health Measurement INstruments Risk of Bias checklist was used to assess methodological quality. A modified Grading of Recommendations Assessment, Development and Evaluation assessed evidence quality.

Results Thirty studies evaluating 26 functional performance tests following ACL injury were included. No studies were found in individuals with an isolated meniscal injury. Included studies evaluated reliability (n=5), measurement error (n=3), construct validity (n=26), structural validity (n=1) and responsiveness (n=1). The Single Leg Hop and Crossover Hop tests showed sufficient intrarater reliability (high and moderate quality evidence, respectively), construct validity (low-quality and moderate-quality evidence, respectively) and responsiveness (low-quality evidence).

Conclusion Frequently used functional performance tests for individuals with ACL or meniscal injury lack evidence supporting their measurement properties. The Single Leg Hop and Crossover Hop are currently the most promising tests following ACL injury. High-quality studies are required to facilitate stronger recommendations of performance-based outcomes following ACL or meniscal injury.

INTRODUCTION

Sport-related traumatic knee injuries in adolescents and young adults are widespread.[1 2] Assessing physical function following knee injury is a key component of evaluating rehabilitation progression, treatment success and return to sport readiness.[3] Objective performance tests are recommended to comprehensively evaluate physical function as they provide complementary information to patient-reported outcomes.[4–6] Functional performance may also inform early identification of individuals at risk of post-traumatic knee osteoarthritis and facilitate development of appropriate secondary prevention strategies.[7 8] While a range of functional performance tests are available, we lack information on those with the best measurement properties (such as reliability, validity and responsiveness).[9]

Knowledge of measurement properties is essential to guide practitioners and researchers in choosing the most appropriate functional performance test for clinical practice or research.[10] To inform rehabilitation progression and evaluation of treatment success a test should be free from measurement error, measure the intended construct and be able to detect changes over time.[11] Using tests with poor or unknown measurement properties introduces the risk of imprecise or biased results, potentially leading to suboptimal treatment and outcomes for individual patients.[10 12] Furthermore, it constitutes a waste of resources if applied for research purposes.[13]

Measurement properties of functional performance tests have rarely been evaluated in systematic reviews for individuals with knee disorders. Past reviews have either: (1) assessed only a single measurement property[14]; (2) been limited to individuals with knee osteoarthritis[15 16] or (3) lacked a structured approach to evaluate measurement properties.[17 18] There is now a need to appraise the measurement properties of the large number of functional performance tests used in adolescents and young adults following knee injury.

This systematic review aimed to critically appraise and summarise the measurement properties of functional performance tests in young adults following anterior cruciate ligament (ACL) or meniscal injury, irrespective of management approach. This systematic review is one of several contributing to the development of evidence-based consensus recommendations for rehabilitation to optimise musculoskeletal health and prevent post-traumatic osteoarthritis following knee trauma (OPTIKNEE; https://bit.ly/OPTIKNEE).

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

MATERIALS AND METHODS

The systematic review was conducted using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guideline[19 20] and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement.[21]

Eligibility criteria for selecting studies

Studies were eligible for inclusion if they met the following criteria:

1. Population: the study sample included individuals with an isolated ACL tear, ACL tear with concomitant meniscal injury, or isolated meniscal injury, with a mean age of injury ≤30 years.
2. Construct: the study included a test that measured physical function, defined as *Activities* according to the International Classification of Functioning, Disability and Health (ICF) framework.[22]
3. Instrument: the instrument was a quantitative functional performance test.
4. Study type: the study evaluated at least one measurement property of a functional performance test (eg, reliability, validity, measurement error, responsiveness).

Studies of individuals with or without surgical intervention were both considered for inclusion. A mean sample age of injury ≤30 years was chosen to limit the inclusion of degenerative knee joint conditions. Studies were excluded if evaluating measurement property of a functional performance test was not the primary aim; the functional performance test was used to validate another instrument (eg, patient-reported outcomes); the study was a randomised controlled trial, systematic review or literature review; the study was not published in full-text (eg, abstracts or conference proceedings), and; the study was published prior to 2000. We restricted the review to studies published from 2000 and onward because the quality of reporting has increased substantially since 2000 and to ensure that the test is likely to be used in clinical practice.

Search strategy and study selection

The search strategy was developed in collaboration with a senior librarian (Medical library, University of Oslo) who also performed the searches. The following databases were searched from earliest available to 16 June 2020, and repeated on 7 July 2021: Medline (Ovid), Embase (Ovid), CINAHL (EBSCO) and SPORTSDiscus (EBSCO). In addition to comprehensive search terms for population and measurement instrument, we used a highly sensitive validated search filter to identify studies evaluating measurement properties (online supplemental appendix 1).[23] No constraints were set on language or date of publication (studies published prior to 2000 were excluded after the search strategy); however, systematic reviews and randomised controlled trials were excluded (using an exclusion filter). The reference lists of all included studies were hand-searched for additional relevant studies by two authors independently (BB and APU).

Identified publications from all databases were imported to EndNote (V.X9.3.3, Clarivate Analytics) and duplicates removed. Titles and abstracts were screened using the Rayyan Qatar Computing Research Institute application[24] by two independent authors (BB and APU), who also independently reviewed the full-text studies for eligibility. Two senior authors (BEØ and MAR) were consulted to resolve discrepancies.

Risk of bias assessment

Two authors (BB and APU) independently assessed included studies using the COSMIN Risk of Bias checklist[25] and the extension for studies on reliability and measurement error.[20] This recently published extension of the COSMIN standards enables transparent and systematic evaluation of the methodological quality of studies investigating performance-based tests reliability and measurement error.[20 26] Each study evaluating a measurement property was rated on a four-point scale (very good, adequate, doubtful or inadequate) based on the standards specific to that measurement property. The lowest rating of any standard was used as the overall rating ('worst score counts' principle).[20 25] For studies evaluating construct validity using several comparator instruments, we considered each result as a study comparison and only combined the methodological quality rating if each standard for all 'study comparisons' were rated the same.[27]

Data extraction

Data were extracted from the included studies by two authors independently (BB and APU). The data extracted included the following items: (1) study and patient sample characteristics; (2) description of the functional performance test, including the procedure and equipment; (3) measurement properties evaluated and (4) results on the measurement properties.

Data synthesis and analysis

The result for the measurement property of each single study was rated as either sufficient (+), insufficient (-) or indeterminate (?) according to the updated criteria for good measurement properties (online supplemental appendix 2).[19] Quantitative pooling was performed if possible, or we qualitatively summarised the results of single studies to obtain an overall result for each functional performance test by measurement property. If the results were inconsistent, we pooled the results but downgraded quality of evidence due to inconsistency.[27] The pooled or summarised results were rated against the same quality criteria,[19] resulting in a rating of either sufficient (+), insufficient (-), inconsistent (±) or indeterminate (?). At least 75% of the results needed to be sufficient (or insufficient) to rate the summarised results as sufficient (or insufficient).[19]

For construct validity, we used quadriceps and hamstring muscle strength assessed by dynamometry and four patient-reported outcomes as comparator instruments (Knee Documentation Committee Subjective Knee Form,[28] Knee injury and Osteoarthritis Outcome Score sport and recreation subscale (sport/rec) and knee-related quality of life subscale,[29] and ACL Return to Sport after Injury Scale).[30] To interpret the results, we formulated a priori hypotheses about the expected relationship (direction and magnitude) between the functional performance tests and comparator instruments based on the generic hypotheses proposed by the COSMIN initiative,[10 27] current literature and the review team's clinical experiences (online supplemental appendix 3).[27] The following reasoning was made specifically for functional performance tests involving hopping or jumping:

1. Muscle strength, particularly quadriceps strength, is an important aspect of activities requiring high force generation.[31]
2. Patient-reported outcomes and functional performance tests measure related but distinct components of physical function.[6]
3. We expected correlations ≥0.50 (quadriceps strength) and ≥0.40 (hamstrings strength), and correlations between 0.30 and 0.50 with patient-reported outcomes.

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

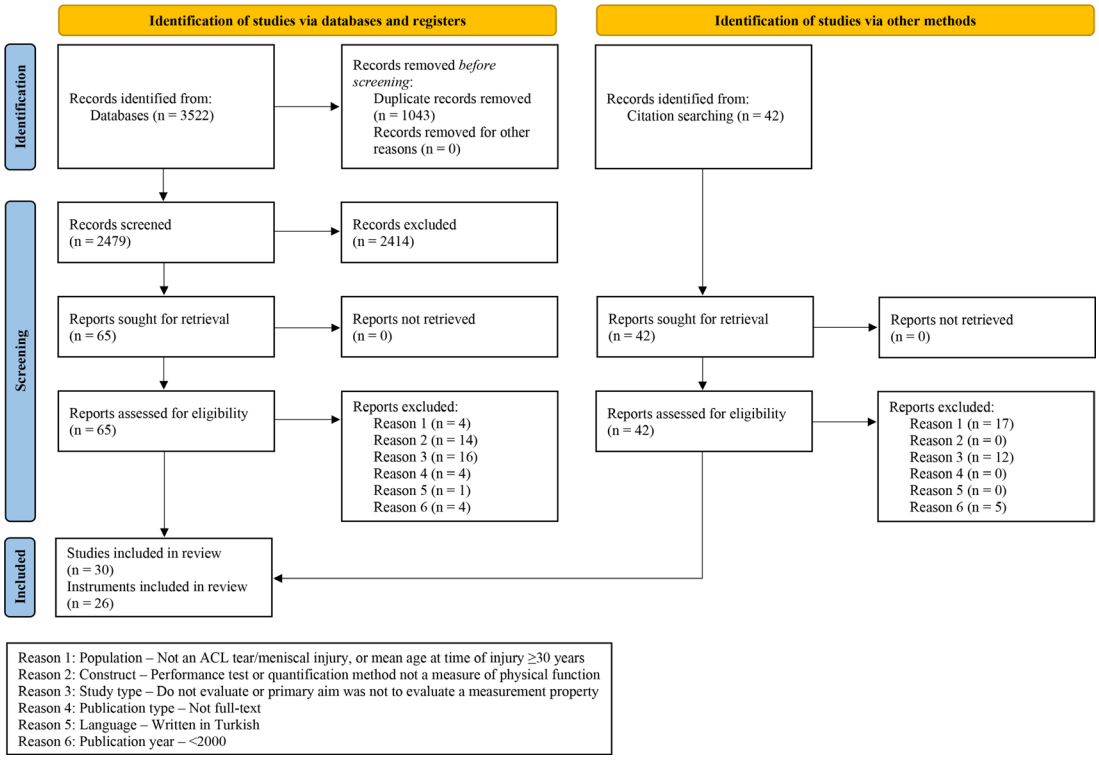


Figure 1 PRISMA flow diagram of study selection. ACL, anterior cruciate ligament; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

When pooling or summarising construct validity results, we considered correlations with quadriceps and hamstrings strength as stronger evidence than patient-reported outcomes as they are physical function based.[27] Studies reporting only p values were not included in construct validity data synthesis.[27] For responsiveness, we hypothesised that changes on the functional performance test would correlate (0.30–0.50) with changes in region-specific patient-reported measures of physical function (related but dissimilar construct).

Quantitative synthesis

Quantitative pooling (ie, performing meta-analysis) was conducted on functional performance tests with two or more available results (for the same comparator) for construct validity. If a study presented results from multiple time points, we considered the shortest time point (from injury or surgery) to be most clinically relevant and analysed data only at this time.[32] We calculated weighted mean Pearson correlation coefficients and 95% CIs. Effect sizes were obtained through the Fisher's *z*-transformation and the pooled effect sizes back-transformed (z–r).[33] A random effect model was applied considering heterogeneity of studies in terms of timing of functional performance testing. The I^2 statistic was calculated to test the proportion of variation in the pooled estimates due to between-study heterogeneity.[34]

Qualitative synthesis

A qualitative summary of the results was performed when quantitative pooling was not possible. The number of confirmed hypotheses were counted across studies on construct validity and responsiveness. For reliability, the point estimate of the intraclass correlation coefficient (ICC) or Cohen's kappa (*k*) was used to conclude whether a measurement instrument had sufficient reliability.[26] Comprehensive research questions were also deducted from the design of studies on reliability and measurement error to determine how the results inform the quality of the functional performance test.[20] For reliability studies describing different operationalisation of the measurement protocol for the same functional performance test, the studies were combined if results were consistent (ie, either sufficient or insufficient).[26]

Grading the quality of evidence

The modified Grading of Recommendations Assessment, Development and Evaluation[19] approach was applied. The quality of evidence was graded separately for each measurement property for each functional performance test, resulting in a grading of either high, moderate, low or very low quality of evidence. Grading was based on four factors: (1) risk of bias (ie, methodological quality of studies using the COSMIN checklist); (2) inconsistency (ie, unexplained inconsistency of results across studies); (3) imprecision (ie, total sample size of available studies) and (4) indirectness (ie, evidence from different populations).[19] No rating of quality of evidence was given for inconsistent results with no explanation for inconsistency or when the overall rating was indeterminate.[19]

Finally, we examined whether excluding studies published prior to year 2000 impacted the results (measurement property rating and quality of evidence).

Patient and public involvement

Two individuals with lived experience of ACL tear (and ACL reconstruction) and four clinicians (ie, physiotherapists, orthopaedic surgeons) contributed to the priority theme setting of this review.

RESULTS

After searches and removal of duplicates, 2521 studies underwent title and abstract review. Of these, we screened 107 full

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Table 1 Characteristics of included studies

Author (year)	Sample size	Females, no (%)	Age, mean±SD	Population	Timing of functional performance test, mean±SD	Functional performance test	Measurement property assessed
Birchmeier (2019)[44]	52	33 (64)	23±5	ACLR	38±24 months postoperative	Single Leg Drop Jump	Construct validity
Bodkin (2017)[45]	51	36 (71)	23±5	ACLR	54 months postoperative	Single Leg Hop	Construct validity
						Triple Hop	
						Crossover Hop	
						6 m Timed Hop	
Chaput (2021)[46]	48	32 (67)	18±3	ACLR	8±2 months postoperative	Single Leg Hop	Construct validity
Clagg (2015)[47]	66	46 (70)	18±3	ACLR	7±2 months postoperative	SEBT anterior	Construct validity
						SEBT post.med	
						SEBT post.lat	
Ebert (2021)[48]	50	16 (32)	28±9	ACLR	10±1 months postoperative	Single Leg Hop	Construct validity
						Triple Hop	
						Crossover Hop	
						6 m Timed Hop	
						Vertical Hop	
						Medial Hop	
						Lateral Hop	
						Timed Speedy Hop	
Fischer (2017)[49]	169	60 (36)	22±8	ACLR	7±3 months postoperative	Vertical Hop	Construct validity
Gustavsson (2006)[73]	30	12 (40)	31±9	ACLD	48±43 weeks postinjury	Vertical Hop	Structural validity
	35	10 (29)	27±7	ACLR	28±2 weeks postoperative	Single Leg Hop	
						Drop Jump	
						Square Hop	
						Side Hop	
Harput (2016)[51]	90	0 (0)	29±8	ACLR	6 months postoperative	Single Leg Hop	Construct validity
						SEBT anterior	
						SEBT post.med.	
						SEBT post.lat.	
Harput (2018)[50]	72	0 (0)	28±8	ACLR	6 months postoperative	Single Leg Hop	Construct validity
						SEBT anterior	
						SEBT post.med.	
						SEBT post.lat.	
Hopper (2002)[52]	19	6 (32)	27±8	ACLR	12±1.5 months postoperative	6 m Timed Hop	Intrareliability
						Crossover Hop	
						Stair Hop	
						Vertical Hop	
Hunnicutt (2020)[53]	30	11 (37)	22 (14–41)*	ACLR	8 (6–23)* months postoperative	Single Leg Hop	Construct validity
Jamshidi (2005)[54]	11	?	30±8	ACLR	>6 months postoperative	Single Leg Hop	Construct validity
						Crossover Hop	
						Vertical Hop	
Keays (2003)[55]	31	9 (29)	27±6	ACLR	33±43 months postinjury	Shuttle Run	Construct validity
					6 months postoperative†	Side-Step	
						Carioca	
						Single Leg Hop	
						Triple Hop	
Kong (2012)[56]	30	0 (0)	23±3	ACLR	≥6 months postoperative	Co-Contraction	Construct validity
						Shuttle Run	
						Carioca	
Laudner (2015)[57]	26	11 (42)	19±4	ACLR	8±2 months postoperative	Vertical Hop	Construct validity
						Vertical Jump	
						4-Jump Vertical Hop	
Lee (2018)[58]	75	0 (0)	28±9	ACLR	9 months postoperative	Vertical Hop	Construct validity
						Single Leg Hop	Intrareliability
Menzer (2017)[59]	88	39 (44)	19±4	ACLR	7±2 months postoperative	Single Leg Hop	Construct validity
						Triple Hop	
						Crossover Hop	

Continued

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Table 1 Continued

Author (year)	Sample size	Females, no (%)	Age, mean±SD	Population	Timing of functional performance test, mean±SD	Functional performance test	Measurement property assessed
Myers (2018)[60]	45	16 (36)	21±6	ACLR	8±2 months postoperative	YBT anterior	Construct validity
						YBT post.med.	
						YBT post.lat.	
						YBT composite	
O'Connor (2020)[61]	452	0 (0)	26±6	ACLR	9±1 months postoperative	Vertical Hop	Construct validity
						Single Leg Drop Jump	
Pua (2015)[71]	87	14 (16)	26±6	ACLD	56* (IQR 21–168) days postinjury	Single Leg Hop	Construct validity
						6 m Timed Hop	
Read (2020 a)[62]	73	0 (0)	24±5	ACLR	≥6 months postoperative	10 s Vertical Hop	Construct validity
Read (2020b)[63]	20	0 (0)	25±4	ACLR	36±11 weeks postoperative	6 m Timed Hop	Inter-reliability
							Measurement error
Reid (2007)[64]	42	19 (45)	26±9	ACLR	16 weeks postoperative	Single Leg Hop	Intrareliability
						6 m Timed Hop	Measurement error
						Triple Hop	Responsiveness
						Crossover Hop	
Reinke (2011)[65]	69	41 (59)	21* (16–35)	ACLR	2–3 years postoperative	Single Leg Hop	Construct validity
						Triple Hop	
						Crossover Hop	
						6 m Timed Hop	
Roe (2021)[66]	66	35 (53)	21±7	ACLR	6±1 months postoperative	Step-Down	Construct validity
						YBT anterior	
Ross (2002)[67]	50	14 (28)	21±1	ACLR	31±16 months postoperative	Single Leg Hop	Construct validity
	10	3 (30)	21±1		27±13 months postoperative	Single Leg Hop	Intrareliability
							Measurement error
Sonesson (2021)[72]	68	30 (44)	31±8	ACLD	43±7 months postinjury	Single Leg Hop	Construct validity
						Crossover Hop	
						Triple Hop	
						6 m Timed Hop	
Sueyoshi (2017)[68]	29	17 (59)	16±1	ACLR	26±6 weeks postoperative	Single Leg Hop	Construct validity
						Triple Hop	
						6 m timed Hop	
						Crossover Hop	
Tunay (2008)[69]	37	0 (0)	25±7	ACLR	6 months postoperative	Single Leg Hop	Construct validity
						Vertical Hop	
Xergia (2015)[70]	22	0 (0)	29±11	ACLR	7±1 months postoperative	Single Leg Hop	Construct validity

*Median (range).
†Indicates the selected time point for timing of functional performance test when multiple time points were presented.
ACLD, anterior cruciate ligament deficient; ACLR, anterior cruciate ligament reconstruction; SEBT, Star Excursion Balance Test; YBT, Y-Balance test.

texts, and finally included 30 studies evaluating 26 unique functional performance tests (figure 1). Of the 107 full texts screened, 9 studies investigating construct validity,[35–39] reliability[40–42] or structural validity[43] were excluded based on publication prior to 2000.

Characteristics of included studies and functional performance tests

Twenty-seven of the included studies involved individuals who had an ACL reconstruction,[44–70] two evaluated ACL deficient individuals[71 72] and one included both ACL reconstructed and ACL deficient individuals (table 1).[73] No studies evaluating individuals with an isolated meniscal injury were identified.

Twenty-five of the 26 tests were single-activity tests,[44–72] with the remaining test a multi-activity functional performance test or battery (table 2).[73] Tests involved a range of tasks, including: (1) hopping,[44–46 48–55 57–59 61–65 67–73] (2) jumping,[57] (3) stepping,[66] (4) running[55 56] or (5) dynamic balance.[47 50 51 60 66] Required equipment varied from common equipment such as measuring tape and stopwatch to more advanced equipment such as a force platform, accelerometer or timing gates. Test administration (eg, best vs mean score) and task constraints (eg, landing requirements or arm placement) also varied between studies for the same functional performance test.

Data synthesis

Included studies investigated five measurement properties: reliability,[52 58 63 64 67] measurement error,[63 64 67] construct validity,[44–51 53–62 65–72] structural validity[73] and responsiveness.[64]

Reliability and measurement error

Four studies[52 58 64 67] evaluated intrarater reliability in individuals who had an ACL reconstruction. The Limb Symmetry Index (LSI) score was the outcome of interest in three studies,[58 64 67] defined as the ratio of the involved limb score and the uninvolved limb score (involved/uninvolved ×100 = LSI). One study[52] included both the LSI score and absolute values. Test scoring varied slightly: the best score of three trials was used in two studies,[52 58] the mean of two trials in one study[64] and the mean of three

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Table 2 Brief description of included functional performance tests

Functional performance test	Activity	Quantification measure	Equipment
Single leg hop	Horizontal hop	Length (cm)	Measuring tape
6 m timed hop	Hop for 6 m	Time (sec)	Stopwatch/timing gates, marks on floor
Crossover hop	Three crossover hops	Length (cm)	Measuring tape, marks on floor
Triple hop	Three hops	Length (cm)	Measuring tape
Vertical hop	Vertical hop	Height (cm)	Contact mat/accelerometer
Single leg drop jump	Drop, vertical hop	Height (cm)	Force platform, box
4-Jump vertical hop	Four vertical hops	Height (cm)	Contact mat
Vertical jump	Vertical jump	Height (cm)	Contact mat
Medial hop	Sideways hop in medial direction	Length (cm)	Measuring tape
Lateral hop	Sideways hop in lateral direction	Length (cm)	Measuring tape
Timed speedy hop	Forwards, backwards, and sideways hops (Agility course)	Time (sec)	SW, hop bars
10 s vertical hop	Vertical hops for 10 s	Height (cm)	Optical measurement device
Stair Hop	Hop up then down three steps	Time (sec)	Stopwatch, 3-step platform
Step-down	Step down on one leg (60 s)	No. of reps.	Stopwatch, box (20 cm), scale
Carioca	Lateral crossover steps	Time (sec)	Stopwatch, marks on floor
Shuttle Run	Run back and forth	Time (sec)	Stopwatch, marks on floor
Side-Step	Run laterally	Time (sec)	Stopwatch, marks on floor
Co-contraction	Run with Velcro belt	Time (sec)	Stopwatch, Velcro belt
SEBT anterior	Dynamic single leg balance	Length (cm)	Measuring tape, marks on floor
SEBT posteromedial	Dynamic single leg balance	Length (cm)	Measuring tape, marks on floor
SEBT posterolateral	Dynamic single leg balance	Length (cm)	Measuring tape, marks on floor
YBT anterior	Dynamic single leg balance	Length (cm)	Measuring tape, stance/reach plates
YBT posteromedial	Dynamic single leg balance	Length (cm)	Measuring tape, stance/reach plates
YBT posterolateral	Dynamic single leg balance	Length (cm)	Measuring tape, stance/reach plates
YBT composite	Dynamic single leg balance	Length (cm)	Measuring tape, stance/reach plates
Test battery: Vertical Hop Single Leg Hop Drop Jump Square Hop Side Hop	Vertical hop Horizontal hop Drop, two horizontal hops Hop in and out of square (30 s) Lateral hops over 40 cm (30 s)	Height (cm) Length (cm) Length (cm) No. of reps. No. of reps.	Contact mat Measuring tape Measuring tape, box Stopwatch, square Stopwatch, marks on floor

SEBT, Star Excursion Balance Test; YBT, Y-Balance test.

trials in one study.[67] No study specified rater experience (online supplemental appendix 4). Sufficient intrarater reliability was found for the Single Leg Hop (ICCs from 0.92 to 0.94),[58 64 67] 6 m Timed Hop (ICCs 0.82 and 0.93),[52 64] Crossover Hop (ICCs 0.84 and 0.94),[52 64] Triple Hop (ICC 0.88),[64] Vertical Hop (ICCs 0.81 and 0.98)[52 58] and Stair Hop (ICC 0.90)[52] tests in studies using LSI as the outcome score (table 3). In one study using absolute values as the outcome score, sufficient intrarater reliability was found for the 6 m Timed Hop (ICC 0.96), Crossover Hop (ICC 0.98), Vertical Hop (ICC 0.98) and Stair Hop (ICC 0.94) tests.[52]

One study[63] evaluated inter-rater reliability of the 6 m Timed Hop Test in individuals who had an ACL reconstruction. Raters included an experienced rehabilitation specialist and a strength and conditioning coach with substantial experience (>5 years) in the test procedures (online supplemental appendix 4). The outcome of interest was the LSI score, dichotomised into pass (≥90%) or fail (<90%). No inter-rater agreement was found for rating a participant as fail (k −0.11) (table 3).

Three studies[63 64 67] evaluated measurement error in individuals who had an ACL reconstruction. The SE of Measurement (SEM) was 2.41% and 3.49% for the Single Leg Hop,[64 67] 5.59% for the 6 m Timed Hop,[64] 5.28% for the Crossover Hop[64] and 4.32% for the Triple Hop test (table 3).[64] The overall rating for all tests was indeterminate because the minimal important change (MIC) was not defined.[26]

Validity

Twenty-six studies evaluated construct validity (hypothesis testing).[44–51 53–62 65–72] Quantitative pooling was applied to the correlations of the Single Leg Hop, 6 m Timed Hop, Crossover Hop, Triple Hop and Vertical Hop tests. Moderate to strong correlations were found with quadriceps and hamstrings strength, whereas correlations with patient-reported outcomes were weak to moderate (table 4). We based the overall rating on correlations with muscle strength outcomes and downgraded the evidence quality by one level due to inconsistency. For the Single Leg Hop Test the quality of evidence was downgraded by two levels due to a lower mean correlation coefficient with hamstrings strength (due to a negative correlation in one study).[70] Individual study methodological quality and results are presented in online supplemental appendix 5.

Sufficient construct validity was found for seven of 19 functional performance tests assessed and summarised in qualitative synthesis: Medial Hop,[48] Lateral Hop,[48] Single Leg Drop Jump,[44 61] 4-jump Vertical Hop,[57] Vertical Jump,[57] Step-Down[66] and Carioca[55 56] tests. In contrast, the Timed Speedy Hop Test,[48] 10 s Vertical Hop Test[62] and Star Excursion Balance Test (anterior, posteromedial, and posterolateral)[47] were rated insufficient. The following tests had inconsistent construct validity: Shuttle Run Test,[55 56] Side-Step Test,[55 56] Co-Contraction Test[56] and Y-balance Test (anterior,[60 66] posterolateral,[60] posteromedial[60]

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Table 3 COSMIN methodological quality ratings and results for reliability and measurement error for each functional performance test

Functional performance test	Outcome score	Study	Design	Reliability			Measurement error	
				n	COSMIN score	Results (rating)	COSMIN score	Results
Single leg hop	LSI	Lee 2018[58]	Intrarater	75	Adequate	ICC: 0.93 (+)		
		Reid 2007[64]	Intrarater	35	Adequate	$ICC_{2,1}$: 0.92 (+)	Adequate	SEM: 3.49
		Ross 2002[67]	Intrarater	10	Very good	$ICC_{2,k}$: 0.94 (+)	Adequate	SEM: 2.41
6 m timed hop	LSI	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.93 (+)		
		Reid 2007[64]	Intrarater	35	Adequate	$ICC_{2,1}$: 0.82 (+)	Adequate	SEM: 5.59
		Read 2020b[63]					Adequate	95% LoA: −23.1, 26.8
	Absolute	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.96 (+)		
		Read 2020b[63]					Adequate	95% LoA: −0.259, 0.245 s.
	LSI*	Read 2020b[63]	Interrater	20	Adequate	κ: −0.11 (-)		
Crossover hop	LSI	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.94 (+)		
		Reid 2007[64]	Intrarater	35	Adequate	$ICC_{2,1}$: 0.84 (+)	Adequate	SEM: 5.28
	Absolute	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.98 (+)		
Triple hop	LSI	Reid 2007[64]	Intrarater	35	Adequate	$ICC_{2,1}$: 0.88 (+)	Adequate	SEM: 4.32
Vertical hop	LSI	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.81 (+)		
		Lee 2018[58]	Intrarater	75	Adequate	ICC: 0.98 (+)		
	Absolute	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.98 (+)		
Stair hop	LSI	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.90 (+)		
	Absolute	Hopper 2002[52]	Intrarater	19	Adequate	$ICC_{3,1}$: 0.94 (+)		

*Binary outcome: ≥90% or <90% Leg Symmetry Index.
COSMIN, COnsensus-based Standards for the selection of health Measurement INstruments; ICC, intraclass correlation coefficient; LoA, Limits of Agreement; LSI, Leg Symmetry Index; MDC, Minimal Detectable Change; SEM, SE of measurement.

and composite score).[60] Individual study methodological quality and results are presented in online supplemental appendix 6.

One study[73] of adequate methodological quality evaluated structural validity of a test battery consisting of five hop tests. The factor analysis divided the five tests in two factors: maximal hop tests (Vertical Hop, Single Leg Hop and Drop Jump) and endurance hop tests (Square Hop and Side Hop). Structural validity was rated indeterminate due to inadequate reporting.

Responsiveness

One study[64] of very good methodological quality investigated responsiveness in individuals who had an ACL reconstruction. Correlations were evaluated between change in LSI scores and change in the Lower Extremity Functional Scale[74] following a 6-week rehabilitation programme. Sufficient rating was found for the Single Leg Hop and Crossover Hop tests (r=0.37 and r=0.41, respectively). In contrast, the 6 m Timed Hop (r=0.28) and Triple Hop (r=0.26) tests were rated insufficiently responsive.[64]

Summary of findings

A summary of the measurement property quality of evidence for each functional performance test is shown in table 5 and Summary of Findings Tables are presented in online supplemental appendix 7. For 21 of the 26 included functional performance tests, only 1 measurement property was evaluated. Sufficient rating for three measurement properties was only found for the Single Leg Hop and Crossover Hop tests, namely intrarater reliability (high and moderate quality of evidence, respectively), construct validity (low and moderate quality of evidence, respectively) and responsiveness (low quality of evidence).

Synthesising data from studies published prior to 2000 otherwise meeting our inclusion criteria did not change our results (online supplemental appendix 8).[35–43] The only exception was the quality of evidence for intrarater reliability of the Vertical Hop Test, which would not have been downgraded for imprecision due to the inclusion of one study with 15 individuals. However, this study used a vertical jump apparatus for measuring jump height,[40] in contrast to the included studies which used a force plate or contact mat.[52 58]

DISCUSSION

This review systematically summarised 30 studies evaluating the measurement properties of 26 functional performance tests according to recently updated COSMIN methodology. All identified studies included individuals who had torn their ACL. Our data synthesis showed a general lack of high-quality evidence for the measurement properties of functional performance tests for use in an ACL injured population. Only the Single Leg Hop and Crossover Hop tests demonstrated sufficient intrarater reliability, construct validity and responsiveness.

Reliability and measurement error

High-quality evidence for sufficient intrarater reliability was found for the Single Leg Hop Test,[58 64 67] meaning we are very confident that the true measurement property lies close to the summarised estimate (ICCs 0.92–0.94). For the five other hop tests investigated, the quality of evidence was downgraded to moderate (6 m Timed Hop,[52 64] Crossover Hop[52 64] and Vertical Hop)[52 58] or very low (Triple Hop[64] and Stair Hop)[52] due to only one or two small studies of adequate methodological quality evaluating the performance test. While these tests are reliable for use in groups (ICC values>0.70), ICC values in excess of 0.90 are recommended for application in individuals.[10] For the Single Leg Hop Test, ICCs above the 0.90 cut-off were consistently found across three studies, indicating sufficient intrarater reliability for use in individuals. Slightly lower ICC values (from 0.82 to 0.88) were reported for the 6 m Timed Hop, Crossover Hop and Triple Hop tests in one study.[64] Recommended strategies to improve these measures' reliability include training raters in the measurement procedures, patient familiarisation, and averaging repeated measures (eg, three measurements per rater).[10]

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Table 4 Meta-analyses of hop tests for construct validity

Functional performance test	Weighted mean correlation coefficients (95% CIs)											
	Quadriceps strength	I^2	Hamstring strength	I^2	IKDC	I^2	KOOS sport/rec	I^2	KOOS QoL	I^2	ACL-RSI	I^2
Single leg hop	0.54 (0.43, 0.64)$^{10;395}$	49	0.36 (0.07, 0.59)$^{4;169}$	68	0.32 (0.21, 0.42)$^{7;369}$	15	0.27 (0.16, 0.38)$^{4;280}$	0	0.29 (0.11, 0.46)$^{2;192}$	43	0.23 (0.06, 0.38)$^{2;140}$	0
6 m timed hop	0.53 (0.24, 0.74)$^{2;137}$	74	0.51 (0.36, 0.63)$^{2;116}$	0	0.24 (0.11, 0.36)$^{4;238}$	0	0.21 (0.03, 0.38)$^{2;120}$	0	0.22 (0.04, 0.39)$^{2;120}$	0	0.32 (-)$^{1;50}$	–
Crossover hop	0.65 (0.21, 0.87)$^{3;90}$	81	0.46 (-)$^{1;29}$	–	0.25 (0.10, 0.39)$^{5;249}$	27	0.26 (0.12, 0.38)$^{3;208}$	0	0.18 (-0.01, 0.36)$^{2;120}$	8	0.22 (-)$^{1;50}$	–
Triple hop	0.53 (0.37, 0.66)$^{3;110}$	0	0.42 (0.18, 0.61)$^{2;60}$	0	0.25 (0.13, 0.43)$^{4;238}$	0	0.26 (0.12, 0.38)$^{3;208}$	0	0.22 (0.02, 0.40)$^{2;120}$	18	0.27 (-)$^{1;50}$	–
Vertical hop	0.56 (0.34, 0.73)$^{6;368}$	82	0.33 (-0.07, 0.64)$^{2;101}$	69	0.50 (0.36, 0.62)$^{3;136}$	0	–	–	–	–	0.40 (0.08, 0.65)$^{3;577}$	89

Numbers in superscript represent number of studies; total number of individuals. Negative values (6 m Timed Hop) have been removed from table for consistent presentation of result
ACL-RSI, Anterior Cruciate Ligament Return to Sport after Injury Scale; IKDC, International Knee Documentation Committee Subjective Knee Form; KOOS, Knee injury and Osteoarthritis Outcome Score; QoL, quality of life.

Ratings for intrarater reliability were based on studies using the LSI score as the outcome.[52 58 64 67] The LSI provide an important measure of performance of the involved limb in relation to the uninvolved limb and is easy to understand for clinicians and patients. LSI values may also be more consistent than absolute values as it accounts for the learning effect in both limbs.[64] However, caution is required when interpreting changes in hop performance using LSI scores as bilateral deficits may be present.[75]

One study[63] compared consistency of pass (≥90% LSI) and fail (<90% LSI) decisions between two raters for the 6 m Timed Hop Test. Although agreement was suitable (80%), they never agreed on rating participants as fail (inter-rater reliability: k −0.11). While this may question the commonly used criteria to complete rehabilitation and subsequently return to sport, we interpret the result with caution as only four players were rated as fail.[63]

Measurement error was rated indeterminate for the Single Leg Hop,[64 67] 6 m Timed Hop,[63 64] Crossover Hop[64] and Triple Hop[64] tests because MIC thresholds have not been established. Information on the MIC is required to apply the criteria for good measurement error,[19] that is, if the smallest detectable change is lower than changes perceived as important by patients (or clinicians).[10] Although the four tests were rated indeterminate, the results provide important information to clinicians and researchers on the ability to detect true changes. For the Single Leg Hop Test, the smallest detectable change was 6.7% and 9.7% (calculated based on the SEM: 1.96*√2*SEM),[64 67] indicating that changes in LSI scores above 10% in individual patients represents a real change. The smallest detectable change was slightly higher for the 6 m Timed Hop (15.5%), Crossover Hop (14.6%) and Triple Hop (12.0%) tests.

Validity

To provide an estimate for construct validity, we compared functional performance tests with knee muscle strength and patient-reported outcomes. In line with our expectations, tests involving hopping demonstrated moderate to strong associations with quadriceps strength. However, these tests should not be considered a measure of knee muscle strength following ACL injury as muscle compensation strategies may be adopted.[76] Further, in ACL reconstructed athletes it has been demonstrated that the knee only contributes 12% to the propulsive phase of the Single Leg Hop Test, and that asymmetries in knee work and kinematics (propulsion and landing phases) can exists despite between-limb hop distance symmetry.[77] Hop performance tests, therefore, measure total lower extremity function rather than solely performance of the knee joint or single muscles.[77 78]

Pooled effect sizes showed weak to moderate correlations with patient-reported outcomes for the Single Leg Hop, 6 m Timed Hop, Crossover Hop, Triple Hop and Vertical Hop tests. These results are consistent with a previous review[14] and support a clear distinction between these two types of constructs. Functional performance tests may better distinguish between pain and function and evaluate what individuals can do rather than what they perceive they can do, as assessed by patient-reported outcomes.[16 79]

Responsiveness

All ratings on responsiveness were based on one study[64] assessing relationships between changes in functional performance tests following a 6-week intervention period with changes in the Lower Extremity Functional Scale,[74] a region-specific

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Table 5 Quality of measurement property evidence by functional performance test

	Reliability			Validity		
Functional performance test	Intra	Inter	Measurement error	Structural	Construct	Responsiveness
Single leg hop	+ (High)	0	?	0	+ (Low)	+ (Low)
6 m timed hop	+ (Moderate)*	- (Very low)	?	0	+ (Moderate)	- (Low)
Crossover hop	+ (Moderate)*	0	?	0	+ (Moderate)	+ (Low)
Triple hop	+ (Very low)	0	?	0	+ (Moderate)	- (Low)
Vertical hop	+ (Moderate)*	0	0	0	+ (Moderate)	0
Medial hop	0	0	0	0	+ (Low)	0
Lateral hop	0	0	0	0	+ (Low)	0
Timed speedy hop	0	0	0	0	- (Moderate)	0
10 s. vertical hop	0	0	0	0	- (Moderate)	0
Single leg drop jump	0	0	0	0	+ (Low)	0
4-Jump vertical hop	0	0	0	0	+ (Very low)	0
Vertical jump	0	0	0	0	+ (Low)	0
Stair hop	+ (Very low)	0	0	0	0	0
Step-down	0	0	0	0	+ (Moderate	0
Carioca	0	0	0	0	+ (Moderate)	0
Shuttle run	0	0	0	0	±	0
Side-step	0	0	0	0	±	0
Co-contraction	0	0	0	0	±	0
SEBT anterior	0	0	0	0	- (High)	0
SEBT posteromedial	0	0	0	0	- (Moderate)	0
SEBT posterolateral	0	0	0	0	- (High)	0
YBT anterior	0	0	0	0	±	0
YBT posteromedial	0	0	0	0	±	0
YBT posterolateral	0	0	0	0	±	0
YBT composite	0	0	0	0	±	0
Test battery	0	0	0	?	0	0

+=sufficient rating; ?=indeterminate rating: minimal important change not defined (measurement error) or not all information reported for + (structural); 0=no information; ±=inconsistent rating; -=insufficient rating
*Consistent result for leg symmetry index and absolute values as the outcome score, but very low quality of evidence for absolute scores.
SEBT, Star Excursion Balance Test; YBT, Y-Balance test.

self-reported functional status measure. Although the evidence quality was low, the Single Leg Hop and Crossover Hop tests appear to be able to detect changes in physical function.[64] We noted somewhat stronger correlations with a generic Global Rating of Change questionnaire in the same study.[64] However, responsiveness concerns the ability to measure the right amount of change in the purported construct.[10]

Limitations

The present review used the COSMIN methodology to evaluate measurement properties of functional performance tests, including the extended version of the COSMIN Risk of Bias checklist for studies on reliability and measurement error of performance-based outcome measures.[20] Although the COSMIN methodology has been used in several reviews of functional performance tests,[15 16 18 80] it was originally developed for reviews of patient-reported outcome measures.[19] Considering the differences in instrument characteristics between performance-based tests and patient-reported outcomes, this may have affected our conclusions. As an example, if the number of hypotheses tested and the cut-off for good measurement properties varied, the results may change. The current criteria for good measurement properties for construct validity and responsiveness (75% of hypotheses confirmed) is based on consensus but arbitrary to a certain extent.[81] Furthermore, hypotheses about correlations are to some extent a best guess. A different team of reviewers would likely construct slightly different hypotheses, possibly leading to different conclusions.[82]

Tests using kinematic quantification methods, for example, gait analysis, were outside of the present review scope as they were considered measures of Body function in the ICF model.[22] Neither did we intend to capture instruments used for predictive purposes. The COSMIN methodology was developed for instruments used for evaluative purposes and would require adaptions for predictive applications.[27] The predictive validity of functional performance tests has been summarised in a recent systematic review.[14] We did not include test batteries of multitask items not solely assessing Activities but acknowledge that tests combining aspects of function with impairments (eg, muscle strength) provide complementary information regarding physical function. However, they are challenging to interpret because they measure different underlying constructs.

One small study presented results on construct validity from multiple time points,[55] and only data from 6 months postsurgery were included in the data analysis. Analysing data from the same participants pre-surgery would not have changed our results. Further, we only included studies with a mean sample age of injury ≤30 years to limit the inclusion of tests evaluated in middle-aged and older people with a supposedly lower physical activity level. The age criterion precluded the inclusion of one study evaluating individuals with a meniscal tear.[83] Finally, studies published before year 2000 were excluded. However, our results remained essentially unchanged when synthesising data from studies published prior to year 2000 otherwise meeting our inclusion criteria.[35–43]

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Clinical implications and future research

With the currently available evidence, it is difficult to formulate strong recommendations for functional performance tests to choose for clinical practice. Based on the identified measurement property evidence, we recommend the Single Leg Hop and Crossover Hop tests in individuals following ACL reconstruction as they have excellent intrarater reliability, support for validity and are potentially responsive to detect changes. However, recent studies have questioned whether horizontal hop distance is the best metric of knee function.[77 84] Given that the knee contributes greater work during a vertical hop, and that vertical hop tests are better at identifying between-limb asymmetries compared with horizontal hop tests,[84] it may also be important to assess vertical hop performance (eg, with the Vertical Hop Test) for rehabilitation progression and return to sport readiness. Finally, it is important to point out that the other functional performance tests identified in the current review are not necessarily of 'poor' quality. Most measurement properties are unknown and robust studies are needed to determine their suitability for use in clinical practice and research.

This review identified important areas for future research. First, better reporting standards are needed to enhance the quality and reproducibility of future research and facilitate correct administration of these tests in clinical practice. Even though slight alterations in test administration may not influence intrarater reliability, it may lead to different scores and make comparison or synthesis of study findings difficult when used as an outcome for interventional studies or return to sport criteria.[85] Second, studies evaluating reliability often lack essential information on sources of variability in the measurement setting (eg, rater characteristics) and the statistical approach. Future studies should adhere to a reporting guideline to improve interpretation of their findings, such as the Guidelines for Reporting Reliability and Agreement Studies.[86] Third, more knowledge on interpretability and responsiveness are needed in particular. To establish MIC values for patient-reported outcomes the widely accepted optimal approach is the use of an external patient-reported anchor.[10] However, patient-reported outcomes and functional performance tests are distinctly different concepts and establishing a credible anchor-based MIC value is less straightforward for performance-based outcomes. Therefore, changes that are measurable (smallest detectable change) may be the most appropriate benchmark for functional performance tests. Such information can further inform an opinion-based approach with a Delphi process to determine changes considered to be minimally important for different functional performance tests.[10 87]

What is already known?

⇒ Functional performance tests are frequently used in research and clinical practice to assess physical function following knee injury.
⇒ Functional performance tests complement patient-reported outcomes, but consensus on which tests have the best measurement properties and clinical relevance in individuals who have had an anterior cruciate ligament (ACL) tear or meniscal injury is lacking.

What are the new findings?

⇒ A wide variety of functional performance tests has been used following ACL injury, but there is a paucity of evidence about their measurement properties.
⇒ The Single Leg Hop Test and Crossover Hop Test are the highest rated tests for use with individuals that have had an ACL injury and reconstruction, displaying excellent intrarater reliability, and support for construct validity and responsiveness.
⇒ The 6 m Timed Hop Test and Triple Hop Test demonstrate good intrarater reliability and support for construct validity, but insufficient responsiveness.
⇒ There is no information published about the measurement properties of functional performance tests for use after isolated meniscal injury

CONCLUSION

This systematic review highlighted the paucity of evidence for measurement properties of functional performance tests following ACL or meniscal injury. The Single Leg Hop was the most studied and highest rated test, followed by the Crossover Hop Test. Still, high-quality studies on responsiveness and interpretability are needed to make strong recommendations for which functional performance tests should be used in youth and young adults following knee injury.

Author affiliations
[1]Division of Orthopaedic Surgery, Oslo University Hospital, Oslo, Norway
[2]Centre for Intelligent Musculoskeletal Health, Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway
[3]Department of Sports Medicine, Norwegian School of Sport Sciences, Oslo, Norway
[4]Department of Physiotherapy, Oslo Metropolitan University, Oslo, Norway
[5]Department of Physical Therapy, Faculty of Medicine, The University of British Columbia, Vancouver, British Columbia, Canada
[6]Arthritis Research Centre, Vancouver, Vancouver, Canada
[7]La Trobe Sport and Exercise Medicine Research Centre, La Trobe University School of Allied Health Human Services and Sport, Bundoora, Victoria, Australia
[8]Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark
[9]Department of Physiotherapy and Occupational Therapy, Copenhagen University Hospital, Herlev and Gentofte, Kobenhavn, Denmark

Twitter Anouk P Urhausen @AnoukUrhausen, Jackie L Whittaker @jwhittak_physio, Adam G Culvenor @agculvenor and Ewa M Roos @ewa_roos

Acknowledgements The authors greatly acknowledge senior librarian Marte Ødegaard (Medical library, University of Oslo) for assistance with developing search strategies and performing the searches. We further acknowledge Pætur Holm, Marienke van Middelkoop, Stephanie Filbay, and Erin Macri for contributing valuable methodological input.

Contributors JLW, AGC, EMR, KMC, CBJ and MAR contributed to the conception of the study. BB, APU, BEØ and MAR designed the study. BB and APU screened studies for inclusion, performed risk of bias assessment and data extraction. All authors contributed to the interpretation of data. BB wrote the initial draft. All authors revised the draft critically for important intellectual content and approved the final version.

Funding This review is part of the OPTIKNEE consensus (https://bit.ly/OPTIKNEE) which has received funding from the Canadian Institutes of Health Research (OPTIKNEE principal investigator JLW #161821).

Disclaimer The funders had no role in any part of the study or in any decision about publication.

Competing interests JLW and AGC are Associate Editors of the British Journal of Sports Medicine (BJSM). JLW is an Editor with the Journal of Orthopaedic and Sports Physical Therapy. KMC is a senior advisor of BJSM, project leader of the Good Life with Osteoarthritis from Denmark (GLA:D)-Australia a not-for profit initiative to implement clinical guidelines in primary care, and holds a research grant from Levin Health outside the submitted work. CBJ an Associate Editor of Osteoarthritis and Cartilage. ER is Deputy Editor of Osteoarthritis and Cartilage, developer of Knee injury and Osteoarthritis Outcome Score (KOOS) and several other freely available patient-reported outcome measures, and founder of the GLA:D). All other authors declare no competing interests.

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Bjørnar Berg http://orcid.org/0000-0002-9017-5562
Anouk P Urhausen http://orcid.org/0000-0002-4409-1501
Britt Elin Øiestad http://orcid.org/0000-0002-0547-9781
Jackie L Whittaker http://orcid.org/0000-0002-6591-4976
Adam G Culvenor http://orcid.org/0000-0001-9491-0264
Ewa M Roos http://orcid.org/0000-0001-5425-2199
Kay M Crossley http://orcid.org/0000-0001-5892-129X
Carsten B Juhl http://orcid.org/0000-0001-8456-5364
May Arna Risberg http://orcid.org/0000-0002-9511-1314

REFERENCES

1 Sanders TL, Maradit Kremers H, Bryan AJ, *et al*. Incidence of anterior cruciate ligament tears and reconstruction: a 21-year population-based study. *Am J Sports Med* 2016;44:1502–7.
2 Louw QA, Manilall J, Grimmer KA. Epidemiology of knee injuries among adolescents: a systematic review. *Br J Sports Med* 2008;42:2–10.
3 Filbay SR, Grindem H. Evidence-Based recommendations for the management of anterior cruciate ligament (ACL) rupture. *Best Pract Res Clin Rheumatol* 2019;33:33–47.
4 Bent NP, Wright CC, Rushton AB, *et al*. Selecting outcome measures in sports medicine: a guide for practitioners using the example of anterior cruciate ligament rehabilitation. *Br J Sports Med* 2009;43:1006–12.
5 Neeb TB, Aufdemkampe G, Wagener JH, *et al*. Assessing anterior cruciate ligament injuries: the association and differential value of questionnaires, clinical tests, and functional tests. *J Orthop Sports Phys Ther* 1997;26:324–31.
6 Reiman MP, Manske RC. The assessment of function: how is it measured? A clinical perspective. *J Man Manip Ther* 2011;19:91–9.
7 Whittaker JL, Woodhouse LJ, Nettel-Aguirre A, *et al*. Outcomes associated with early post-traumatic osteoarthritis and other negative health consequences 3-10 years following knee joint injury in youth sport. *Osteoarthritis Cartilage* 2015;23:1122–9.
8 Patterson B, Culvenor AG, Barton CJ, *et al*. Poor functional performance 1 year after ACL reconstruction increases the risk of early osteoarthritis progression. *Br J Sports Med* 2020;54:546–53.
9 Emery CA, Whittaker JL, Mahmoudian A, *et al*. Establishing outcome measures in early knee osteoarthritis. *Nat Rev Rheumatol* 2019;15:438–48.
10 de Vet HCW, Terwee CB, Mokkink LB. *Measurement in medicine: a practical guide*. Cambridge: Cambridge University Press, 2011.
11 Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010;63:737–45.
12 Robertson S, Kremer P, Aisbett B, *et al*. Consensus on measurement properties and feasibility of performance tests for the exercise and sport sciences: a Delphi study. *Sports Med Open* 2017;3:2–3.
13 Ioannidis JPA, Greenland S, Hlatky MA, *et al*. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014;383:166–75.
14 Losciale JM, Bullock G, Cromwell C, *et al*. Hop testing lacks strong association with key outcome variables after primary anterior cruciate ligament reconstruction: a systematic review. *Am J Sports Med* 2020;48:511–22.
15 Dobson F, Hinman RS, Hall M, *et al*. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. *Osteoarthritis Cartilage* 2012;20:1548–62.
16 Terwee CB, Mokkink LB, Steultjens MPM, *et al*. Performance-Based methods for measuring the physical function of patients with osteoarthritis of the hip or knee: a systematic review of measurement properties. *Rheumatology* 2006;45:890–902.
17 Clark NC. Functional performance testing following knee ligament injury. *Phys Ther Sport* 2001;2:91–105.
18 Kroman SL, Roos EM, Bennell KL, *et al*. 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 Cartilage* 2014;22:26–39.
19 Prinsen CAC, Mokkink LB, Bouter LM, *et al*. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1147–57.
20 Mokkink LB, Boers M, van der Vleuten CPM, *et al*. COSMIN risk of bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol* 2020;20:293.
21 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71.
22 World Health Organization. *International classification of functioning, disability and health : ICF*. 2nd ed. Geneva, Switzerland: World Health Organization, 2001.
23 Terwee CB, Jansma EP, Riphagen II, *et al*. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res* 2009;18:1115–23.
24 Ouzzani M, Hammady H, Fedorowicz Z, *et al*. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016;5:210.
25 Mokkink LB, de Vet HCW, Prinsen CAC, *et al*. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1171–9.
26 Mokkink L, Boers M, van der Vleuten C. COSMIN risk of bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument – user manual, 2021. Available: http://www.cosmin.nl/ [Accessed Feb 2021].
27 Mokkink L, Prinsen C, Patrick D. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs) – user manual, 2018. Available: http://www.cosmin.nl/ [Accessed Jan 2021].
28 Irrgang JJ, Anderson AF, Boland AL, *et al*. Development and validation of the International knee documentation Committee subjective knee form. *Am J Sports Med* 2001;29:600–13.
29 Roos EM, Roos HP, Lohmander LS, *et al*. Knee Injury and Osteoarthritis Outcome Score (KOOS)-development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998;28:88–96.
30 Webster KE, Feller JA, Lambros C. Development and preliminary validation of a scale to measure the psychological impact of returning to sport following anterior cruciate ligament reconstruction surgery. *Phys Ther Sport* 2008;9:9–15.
31 Burland JP, Lepley AS, Frechette L, *et al*. Protracted alterations in muscle activation strategies and knee mechanics in patients after anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2020;28:3766–72.
32 Higgins JPT, Thomas J, Chandler J. Cochrane Handbook for systematic reviews of interventions version 6.2, 2021. Available: www.training.cochrane.org/handbook
33 Borenstein M, Hedges L, Higgins J. *An introduction to meta-analysis*. Wiley, 2009.
34 Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539–58.
35 Borsa PA, Lephart SM, Irrgang JJ. Comparison of performance-based and patient-reported measures of function in anterior-cruciate-ligament-deficient individuals. *J Orthop Sports Phys Ther* 1998;28:392–9.
36 Lephart SM, Perrin DH, Fu FH, *et al*. Relationship between selected physical characteristics and functional capacity in the anterior cruciate Ligament-Insufficient athlete. *J Orthop Sports Phys Ther* 1992;16:174–81.
37 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. *Eur J Phys Med Rehabil* 1997;7:81–6.
38 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. *J Orthop Sports Phys Ther* 1998;28:23–31.
39 Wilk KE, Romaniello WT, Soscia SM, *et al*. The relationship between subjective knee scores, isokinetic testing, and functional testing in the ACL-reconstructed knee. *J Orthop Sports Phys Ther* 1994;20:60–73.
40 Brosky JA, Nitz AJ, Malone TR, *et al*. Intrarater reliability of selected clinical outcome measures following anterior cruciate ligament reconstruction. *J Orthop Sports Phys Ther* 1999;29:39–48.
41 Kramer JF, Nusca D, Fowler P, *et al*. Test-Retest reliability of the One-Leg hop test following ACL reconstruction. *Clin J Sport Med* 1992;2:240–3.
42 Paterno MV, Greenberger HB. The test-retest reliability of a one legged hop for distance in young adults with and without ACL reconstruction. *Isokinet Exerc Sci* 1996;6:1–6.
43 Risberg MA, Ekeland A. Assessment of functional tests after anterior cruciate ligament surgery. *J Orthop Sports Phys Ther* 1994;19:212–7.
44 Birchmeier T, Lisee C, Geers B, *et al*. Reactive strength index and knee extension strength characteristics are predictive of Single-Leg hop performance after anterior cruciate ligament reconstruction. *J Strength Cond Res* 2019;33:1201–7.
45 Bodkin S, Goetschius J, Hertel J, *et al*. Relationships of muscle function and subjective knee function in patients after ACL reconstruction. *Orthop J Sports Med* 2017;5:2325967117719041.
46 Chaput M, Palimenio M, Farmer B, *et al*. Quadriceps strength influences patient function more than single leg forward hop during late-stage ACL rehabilitation. *Int J Sports Phys Ther* 2021;16:145–55.
47 Clagg S, Paterno MV, Hewett TE, *et al*. Performance on the modified star excursion balance test at the time of return to sport following anterior cruciate ligament reconstruction. *J Orthop Sports Phys Ther* 2015;45:444–52.
48 Ebert JR, Edwards P, Preez LD, *et al*. Knee extensor strength, hop performance, patient-reported outcome and inter-test correlation in patients 9-12 months after anterior cruciate ligament reconstruction. *Knee* 2021;30:176–84.

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

49 Fischer F, Blank C, Dünnwald T, *et al*. Isokinetic extension strength is associated with Single-Leg vertical jump height. *Orthop J Sports Med* 2017;5:2325967117736766.
50 Harput G, Ozer H, Baltaci G, *et al*. Self-Reported outcomes are associated with knee strength and functional symmetry in individuals who have undergone anterior cruciate ligament reconstruction with hamstring tendon autograft. *Knee* 2018;25:757–64.
51 Harput G, Tok D, B Tunay V. Relation of self-reported knee function and physical performance with psychological responses in anterior cruciate ligament reconstructed individuals. *Turk J Physiother Rehabil* 2016;27:102–7.
52 Hopper DM, Goh SC, Wentworth LA, *et al*. Test–retest reliability of knee rating scales and functional hop tests one year following anterior cruciate ligament reconstruction. *Phys Ther Sport* 2002;3:10–18.
53 Hunnicutt JL, McLeod MM, Slone HS, *et al*. Quadriceps neuromuscular and physical function after anterior cruciate ligament reconstruction. *J Athl Train* 2020;55:238–45.
54 Jamshidi AA, Olyaei GR, Heydarian K, *et al*. Isokinetic and functional parameters in patients following reconstruction of the anterior cruciate ligament. *Isokinet Exerc Sci* 2005;13:267–72.
55 Keays SL, Bullock-Saxton JE, Newcombe P, *et al*. The relationship between knee strength and functional stability before and after anterior cruciate ligament reconstruction. *J Orthop Res* 2003;21:231–7.
56 Kong DH, Yang SJ, Ha JK, *et al*. Validation of functional performance tests after anterior cruciate ligament reconstruction. *Knee Surg Relat Res* 2012;24:40–5.
57 Laudner K, Evans D, Wong R, *et al*. Relationship between isokinetic knee strength and jump characteristics following anterior cruciate ligament reconstruction. *Int J Sports Phys Ther* 2015;10:272–80.
58 Lee DW, Yang SJ, Cho SI, *et al*. Single-leg vertical jump test as a functional test after anterior cruciate ligament reconstruction. *Knee* 2018;25:1016–26.
59 Menzer H, Slater LV, Diduch D, *et al*. The utility of objective strength and functional performance to predict subjective outcomes after anterior cruciate ligament reconstruction. *Orthop J Sports Med* 2017;5:2325967117744758.
60 Myers H, Christopherson Z, Butler RJ. Relationship between the lower quarter Y-BALANCE test scores and isokinetic strength testing in patients status post ACL reconstruction. *Int J Sports Phys Ther* 2018;13:152–9.
61 O'Connor RF, King E, Richter C, *et al*. No relationship between strength and power scores and anterior cruciate ligament return to sport after injury scale 9 months after anterior cruciate ligament reconstruction. *Am J Sports Med* 2020;48:78–84.
62 Read PJ, Davies WT, Bishop C, *et al*. Residual deficits in reactive strength indicate incomplete restoration of athletic qualities following anterior cruciate ligament reconstruction in professional soccer players. *J Athl Train* 2020a. doi:10.4085/169-20. [Epub ahead of print: 05 Nov 2020].
63 Read PJ, Palli T, Oliver JL. The 6-m timed hop is not a suitable clinical assessment tool for use following anterior cruciate ligament reconstruction. *J Sport Rehabil* 2020b;30:595–600.
64 Reid A, Birmingham TB, Stratford PW, *et al*. Hop testing provides a reliable and valid outcome measure during rehabilitation after anterior cruciate ligament reconstruction. *Phys Ther* 2007;87:337–49.
65 Reinke EK, Spindler KP, Lorring D, *et al*. Hop tests correlate with IKDC and KOOS at minimum of 2 years after primary ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2011;19:1806–16.
66 Roe C, Jacobs C, Kline P, *et al*. Correlations of Single-Leg performance tests to patient-reported outcomes after primary anterior cruciate ligament reconstruction. *Clin J Sport Med* 2021;31:e265–70.
67 Ross MD, Irrgang JJ, Denegar CR, *et al*. The relationship between participation restrictions and selected clinical measures following anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2002;10:10–19.
68 Sueyoshi T, Nakahata A, Emoto G, *et al*. Single-Leg hop test performance and isokinetic knee strength after anterior cruciate ligament reconstruction in athletes. *Orthop J Sports Med* 2017;5:2325967117739811.
69 Tunay VB, Baltaci G, Ergun N. Quadriceps femoris strength and knee functions in soccer players after anterior cruciate ligament reconstruction: six month follow-up. *Turk J Physiother Rehabil* 2008;19:10–14.
70 Xergia SA, Pappas E, Georgoulis AD. Association of the Single-Limb hop test with isokinetic, kinematic, and kinetic asymmetries in patients after anterior cruciate ligament reconstruction. *Sports Health* 2015;7:217–23.
71 Pua Y-H, Ong P-H, Ho J-Y, *et al*. Associations of isokinetic knee steadiness with hop performance in patients with ACL deficiency. *Knee Surg Sports Traumatol Arthrosc* 2015;23:2185–95.
72 Sonesson S, Österberg A, Gauffin H, *et al*. Low correlation between functional performance and patient reported outcome measures in individuals with non-surgically treated ACL injury. *Phys Ther Sport* 2021;47:185–92.
73 Gustavsson A, Neeter C, Thomeé P, *et al*. A test battery for evaluating hop performance in patients with an ACL injury and patients who have undergone ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2006;14:778–88.
74 Binkley JM, Stratford PW, Lott SA, *et al*. The lower extremity functional scale (LEFS): scale development, measurement properties, and clinical application. North American orthopaedic rehabilitation research network. *Phys Ther* 1999;79:371–83.
75 Gokeler A, Welling W, Benjaminse A, *et al*. A critical analysis of limb symmetry indices of hop tests in athletes after anterior cruciate ligament reconstruction: a case control study. *Orthop Traumatol Surg Res* 2017;103:947–51.
76 Barfod KW, Feller JA, Hartwig T, *et al*. Knee extensor strength and hop test performance following anterior cruciate ligament reconstruction. *Knee* 2019;26:149–54.
77 Kotsifaki A, Whiteley R, Van Rossom S, *et al*. Single leg hop for distance symmetry masks lower limb biomechanics: time to discuss hop distance as decision criterion for return to sport after ACL reconstruction? *Br J Sports Med* 2022;56:249–56.
78 Kotsifaki A, Korakakis V, Graham-Smith P, *et al*. Vertical and horizontal hop performance: contributions of the hip, knee, and ankle. *Sports Health* 2021;13:128–35.
79 Stratford PW, Kennedy DM. Performance measures were necessary to obtain a complete picture of osteoarthritic patients. *J Clin Epidemiol* 2006;59:160–7.
80 Jakobsson M, Gutke A, Mokkink LB, *et al*. Level of evidence for reliability, validity, and responsiveness of physical capacity tasks designed to assess functioning in patients with low back pain: a systematic review using the COSMIN standards. *Phys Ther* 2019;99:457–77.
81 Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Med Res Methodol* 2011;11:152.
82 Abma IL, Rovers M, van der Wees PJ. Appraising convergent validity of patient-reported outcome measures in systematic reviews: constructing hypotheses and interpreting outcomes. *BMC Res Notes* 2016;9:226.
83 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.
84 Kotsifaki A, Van Rossom S, Whiteley R, *et al*. Single leg vertical jump performance identifies knee function deficits at return to sport after ACL reconstruction in male athletes. *Br J Sports Med* 2022;56:490–8.
85 Read P, Mc Auliffe S, Wilson MG, *et al*. Better reporting standards are needed to enhance the quality of hop testing in the setting of ACL return to sport decisions: a narrative review. *Br J Sports Med* 2021;55:23.
86 Kottner J, Audigé L, Brorson S, *et al*. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
87 Benaim C, Blaser S, Léger B, *et al*. "Minimal clinically important difference" estimates of 6 commonly-used performance tests in patients with chronic musculoskeletal pain completing a work-related multidisciplinary rehabilitation program. *BMC Musculoskelet Disord* 2019;20:16.

Br J Sports Med: first published as 10.1136/bjsports-2022-105510 on 13 June 2022. Downloaded from http://bjsm.bmj.com/ on August 20, 2022 at Univ of the Free State Consortia. Protected by copyright.

Supplementary Appendix 1. Complete search strategy, exemplified for Ovid MEDLINE(R)

The full search strategy for all databases can be found in the Open Science Framework registration (https://osf.io/bkhr5/)

Field labels	
/	After an index term indicates a subject heading were selected.
.tw.	Indicates a search for a term in title or abstract
.kw.	= keyword heading
.kf.	= keyword heading word
*	At the end of a term indicates that this term has been truncated.
Adj3	Indicates a search for two terms next to each other, in any order, up to 3 words in between.

Searches are not limited by year or language.

Ovid MEDLINE(R) ALL 1946 to July 06, 2021
Advanced search
Date of search: 2021-07-07

#	**Searches**
1	Athletic Injuries/ or Rupture/ or "Sprains and Strains"/
2	Anterior Cruciate Ligament/ or Menisci, Tibial/
3	1 and 2
4	Anterior Cruciate Ligament Injuries/ or Tibial Meniscus Injuries/
5	3 or 4
6	(menisc* adj3 (resect* or injur* or tear* or rupture* or repair* or reconstruct* or shav* or surg*)).tw,kf.
7	((ACL* or anterior cruciate ligament*) adj4 (injur* or tear* or sprain* or rupture* or reconstruct* or surg* or repair* or rupture*)).tw,kf.
8	Meniscectomy/
9	exp Anterior Cruciate Ligament Reconstruction/
10	meniscectom*.tw,kf.
11	or/6-10
12	Arthroscopy/
13	arthroscop*.tw,kf.
14	or/12-13
15	(ACL* or anterior cruciate ligament* or menisc*).tw,kf.
16	2 or 15
17	14 and 16
18	5 or 11 or 17
19	"Task Performance and Analysis"/ or exp patient outcome assessment/
20	(((Performance or performance-based or performancebased or objective or observational or observational-based) adj3 (Measure* or test* or instrument* or method* or index or indices or assessment or assessments or evaluation* or function* or disability or analys*)) or gait analys* or gait evaluation or walk* test).tw,kw,kf.
21	or/19-20
22	(instrumentation or methods).fs.
23	(Validation Studies or Comparative Study).pt.
24	exp Psychometrics/
25	psychometr*.ti,ab.
26	(clinimetr* or clinometr*).tw.
27	exp Outcome Assessment, Health Care/
28	outcome assessment.ti,ab.
29	outcome measure*.tw.
30	exp Observer Variation/
31	observer variation.ti,ab.
32	exp Health Status Indicators/

33	exp Reproducibility of Results/
34	reproducib*.ti,ab.
35	exp Discriminant Analysis/
36	(reliab* or unreliab* or valid* or coefficient or homogeneity or homogeneous or internal consistency).ti,ab.
37	(cronbach* adj3 (alpha or alphas)).ti,ab.
38	(item adj3 (correlation* or selection* or reduction*)).ti,ab.
39	(agreement or precision or imprecision or precise values or test-retest).ti,ab.
40	(test adj3 retest).ti,ab.
41	(reliab* adj3 (test or retest)).ti,ab.
42	(stability or interrater or inter-rater or intrarater or intra-rater or intertester or inter-tester or intratester or intra-tester or interobserver or inter-observer or intraobserver or intraobserver or intertechnician or inter-technician or intratechnician or intra-technician or interexaminer or inter-examiner or intraexaminer or intra-examiner or interassay or interassay or intraassay or intra-assay or interindividual or inter-individual or intraindividual or intra-individual or interparticipant or inter-participant or intraparticipant or intra-participant or kappa or kappas or repeatab*).ti,ab.
43	((replicab* or repeated) adj3 (measure or measures or findings or result or results or test or tests)).ti,ab.
44	(generaliza* or generalisa* or concordance).ti,ab.
45	(intraclass adj3 correlation*).ti,ab.
46	(discriminative or known group or factor analysis or factor analyses or dimension* or subscale*).ti,ab.
47	(multitrait scaling adj3 (analysis or analyses)).ti,ab.
48	(item discriminant or interscale correlation* or error or errors or individual variability).ti,ab.
49	(variability adj3 (analysis or values)).ti,ab.
50	(uncertainty adj3 (measurement or measuring)).ti,ab.
51	(standard error of measurement or sensitiv* or responsive*).ti,ab.
52	(((minimal or minimally or clinical or clinically) adj3 (important or significant or detectable)) and (change or difference)).ti,ab.
53	(small* adj3 (real or detectable) adj3 (change or difference)).ti,ab.
54	(meaningful change or ceiling effect or floor effect or Item response model or IRT or Rasch or Differential item functioning or DIF or computer adaptive testing or item bank or cross-cultural equivalence).ti,ab.
55	or/22-54
56	(addresses or biography or case reports or comment or directory or editorial or festschrift or interview or lectures or legal cases or legislation or letter or news or newspaper article or patient education handout or popular works or congresses or consensus development conference or consensus development conference, nih or practice guideline or randomized controlled trial or randomized controlled trial, veterinary or "systematic review").pt.
57	((veterinar* or animal or animals or rabbit or rabbits or rodent or rodents or rat or rats or mouse or mice or hamster or hamsters or pig or pigs or piglet or piglets or porcine or pigeon* or horse* or equine or cow or cows or bovine or goat or goats or sheep or lamb or lambs or monkey or monkeys or murine or ovine or dog or dogs or canine or cat or cats or feline or dolphin*) not (patient or patients or human or humans)).ti.
58	(Animal Experimentation/ or exp Animals/ or exp Models, Animal/) not Humans/
59	systematic review/ or exp randomized controlled trial/ or (systematic review or randomi?ed controlled trial).ti.
60	or/56-59
61	18 and 21 and 55
62	61 not 60

Supplementary Appendix 2. Updated criteria for good measurement properties

Measurement property	Rating	Criteria
Reliability	+	ICC or (weighted) Kappa ≥ 0.70
	?	ICC or (weighted) Kappa not reported
	-	ICC or (weighted) Kappa < 0.70
Measurement error	+	SDC or LoA or CV*√2*1.96 < M(C)IC*; % specific agreement > 80%
	?	MIC not defined
	-	SDC or LoA or CV*√2*1.96 > M(C)IC†; % specific agreement <80%
Hypothesis testing for construct validity	+	The result is in accordance with the hypothesis**
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis**
Responsiveness	+	The result is in accordance with the hypothesis OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis OR AUC < 0.70
Structural validity	+	**CTT** CFA: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08 **IRT/Rasch** No violation of unidimensionality: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08 *AND* no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 *AND* no violation of monotonicity: adequate looking graphs OR item scalability > 0.30 *AND* adequate model fit IRT: $\chi 2 > 0.001$ Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z-standardized values > −2 and < 2
	?	CTT: not all information for '+' reported IRT/Rasch: model fit not reported
	-	Criteria for '+' not met

Based on Prinsen et al. *Qual Life Res* 2018 and the manual accompanying the extended COSMIN Risk of Bias checklist (Mokkink et al. *BMC Med Res Methodol* 2020).
+ = sufficient; ? = indeterminate; - = insufficient. * The M(C)IC value may come from another study. ** The results of all studies should be taken together and it should then be decided if 75% of the results are in accordance with the hypotheses.
ICC=Intraclass Correlation Coefficient; SDC=Smallest Detectable Change; LoA=Limits of Agreement; M(C)IC=Minimal (Clinical) Important Change; AUC=Area Under the Curve; CTT=Classical Test Theory; CFA=Confirmatory Factor Analysis; CFI=Comparative fit Index; TLI=Tucker-Lewis index; RMSEA=Root Mean Square Error of Approximation; SRMR=Standardized Root Mean Square Residual; IRT=Item Response Theory.

Supplementary Appendix 3. Hypotheses used in data synthesis for construct validity

Type of functional performance test	Data management	Hypotheses	Interpretation
Hopping/Jumping	Quantitative synthesis: Weighted mean correlation coefficients (95% CI) between the functional performance test and comparator instruments. Qualitative synthesis: Individual study results of correlation between the functional performance test and comparator instruments.	Correlation ≥0.50 with quadriceps strength. Correlation ≥0.40 with hamstrings strength. Correlation 0.30-0.50 with patient-reported outcomes (related, but dissimilar constructs)	Considered sufficient if at least 75% of the results are in accordance with the hypotheses.
Running	Qualitative synthesis: Individual study results of correlation between the functional performance test and comparator instruments.	Correlation 0.30-0.60 with quadriceps and hamstrings strength, Correlation 0.30-0.50 with patient-reported outcomes (related, but dissimilar constructs)	Considered sufficient if at least 75% of the results are in accordance with the hypotheses.
Stepping, and Dynamic balance	Qualitative synthesis: Individual study results of correlation between the functional performance test and comparator instruments.	Correlation 0.30-0.50 with quadriceps strength, hamstrings strength, and patient-reported outcomes (related, but dissimilar constructs).	Considered sufficient if at least 75% of the results are in accordance with the hypotheses.

Supplementary Appendix 4. Elements of a comprehensive research questions for included studies on reliability and measurement error

Hopper 2002

Element	Study 1	Study 2	Study 3	Study 4
1. Name of the instrument	6-meter Timed Hop Test	Crossover Hop Test	Stair Hop Test	Vertical Hop Test
2. Version or way of operationalization	**Equipment:** Timing gates (Speedlight Sports Timing System), measuring tape.	**Equipment:** Measuring tape, marks on floor.	**Equipment:** Timing gates (Speedlight Sports Timing System), 3-step platform.	**Equipment:** Force plate (Advanced Metal Technologies Incorporated).
	Preparatory action: One rater (experience not specified). Five-minute warm-up on a stationary bicycle, and static stretches of the lower extremities. One practice trial performed for each limb. The limb to be tested first and the order of testing of the hop tests were randomly determined. Three measured trials. Adequate rest between trials to minimize fatigue effects.			
	Procedure: Hop as fast as possible between timing gates placed 6 meters apart. The time taken to hop 6 meter was recorded in seconds.	**Procedure:** Three consecutive hops for maximal distance, while crossing a 15-cm wide line. Testing was repeated if subject touched the line or grounded the other leg. Distance from the start line to the toe of the test leg was measured.	**Procedure:** Hop up, then down three steps. Turn around a fixed marker 1 m from the platform. The time taken to complete the test was recorded in seconds.	**Procedure:** Perform a deep squat and hop maximally landing on the same foot. Elbows clasped behind the back throughout. The time off the force plate was measured in milliseconds using the strain gauge based force plate.
	Assignment of the score: The best score of three recorded trials was utilized for data analysis (absolute values and leg symmetry index).			
3. Construct	Physical function.			
4. Measurement property	Reliability (rater n=1; intra-rater reliability).			
5. Components that will be repeated	The whole measurement was repeated.			
6. Source of variation	Occasion (7 days).			
7. Patient population	ACL reconstruction (12 mo. post-op.), n=19 (32% females).			
Comprehensive research question: What is the intra-rater reliability of the 6-m Timed Hop, Crossover Hop, Stair Hop and Vertical Hop tests based on the best of three trials (absolute and leg symmetry index scores) in ACL reconstructed individuals.				

Lee 2018

Element	**Study 1**	**Study 2**
1. Name of the instrument	Vertical Hop Test	Single Leg Hop Test
2. Version or way of operationalization	**Equipment:** Contact mat (Sargent jump analyzer).	**Equipment:** Measuring tape.
	Preparation action: One rater (experience not specified). Five-minute warm-up, two minutes stretching. Three practice trials for each test. Three measured trials. Adequate rest between trials to minimize fatigue effects.	
	Procedure: Perform a single deep squat with a pause, followed by a vertical hop on the contact mat. Trials were repeated if patients did not jump to maximum height upon landing or showed landing error. Jump height measured in centimetres by the contact mat.	**Procedure:** Hop forward as far as possible, jumping and landing with the same foot. Measured in centimetres using a ruler on the ground.
	Assignment of the score: The best score of three recorded trials was utilized for data analysis (leg symmetry index).	
3. Construct	Physical function.	
4. Measurement property	Reliability (rater n=1; intra-rater reliability).	
5. Components that will be repeated	The whole test was repeated.	
6. Source of variation	Occasion (14 days).	
7. Patient population	ACL reconstruction (9 mo. post-op.), n=75 (0% females).	
Comprehensive research question: What is the intra-rater reliability of the Vertical Hop Test and Single Leg Hop Test based on the best of three trials (leg symmetry index scores) in ACL reconstructed individuals.		

Read 2020b

Element	Study 1
1. Name of the instrument	6-meter Timed Hop
2. Version or way of operationalization	**Equipment:** Stop-watch, measuring tape.
	Preparatory action: Two raters (one experienced rehabilitation specialist and one strength and conditioning coach with >5 years' experience). Five-minute warm-up consisting of light jogging, and 10 repetitions each of body weight squats and lunges, followed by 3 submaximal hops. Three practice trial. Two measured trials on both the uninvolved and involved leg in that order, with 30 seconds of rest between each.
	Procedure: Hop as far as possible in a linear direction for 6-meters. A 30 second rest period between each trial. The stopwatch was started when the patient`s heel lifted and stopped when the patient crossed the finish line.
	Assignment of the score: Best of the two recorded trials was utilized for data analysis. Leg symmetry index (above or below 90%) was utilized for inter-rater reliability. Absolute scores and leg symmetry index (%) were utilized for measurement error.
3. Construct	Physical function.
4. Measurement property	Reliability (rater n=2; inter-rater reliability) and measurement error.
5. Components that will be repeated	The whole test was repeated.
6. Source of variation	Rater
7. Patient population	ACL reconstruction (27 mo. post-op.), n=10 (30% females).
Comprehensive research question: What is the inter-rater reliability (leg symmetry index values [above or below 90%]) and 95% Limits of Agreement (absolute values and leg symmetry index scores) of the 6-meter Timed Hop Test based on the mean of two trials in ACL reconstructed individuals.	

Reid 2007

Element	Study 1	Study 2	Study 3	Study 4
1. Name of the instrument	Single Leg Hop Test	6-meter Timed Hop Test	Triple Hop Test	Crossover Hop Test
2. Version or way of operationalization	**Equipment:** Measuring tape.	**Equipment:** Stop-watch, measuring tape.	**Equipment:** Measuring tape.	**Equipment:** Measuring tape, marks on floor.
	Preparation action: One rater (experience not specified). No warm-up activity. One practice trial. Two measured trials, beginning with the nonoperative limb. Rest period to minimize fatigue between types of hop tests (up to 2 minutes) and between individual hop test trials if needed (typically less than 30 seconds).			
	Procedure: Hop and land on the same limb. No restriction for arm movement. Test repeated if the subjected grounded the other leg, lost balance, touched down upper extremity, did not maintain landing for 2 seconds. Measured to the nearest centimetre at the level of the great toe.	**Procedure:** Perform large one-legged hops in series over 6 meters. No restriction for arm movement. The stopwatch was started when a subject`s heel lifted from the starting position and was stopped the moment the tested foot passed the finish line.	**Procedure:** Stand on one leg and perform three consecutive hops as far as possible. No restriction for arm movement. Test repeated if the subjected grounded the other leg, lost balance, touched down upper extremity, did not maintain landing for 2 seconds.	**Procedure:** Hop forward 3 times while alternately crossing over a marking (15-cm). No restriction for arm movement. Test repeated if the subjected grounded the other leg, lost balance, touched down upper extremity, did not maintain landing for 2 seconds.
	Assignment of the score: Mean of the two recorded trials was utilized for data analysis (leg symmetry index).			
3. Construct	Physical function.			
4. Measurement property	Reliability (rater n=1; intra-rater reliability) and measurement error.			
5. Components that will be repeated	The whole test was repeated.			
6. Source of variation	Occasion (24-48 hours).			
7. Patient population	ACL reconstruction (16 weeks post-op.), n=42 (45% females)			
Comprehensive research question: What is the intra-rater reliability/standard error of measurement of the Single Leg Hop, 6-meter Timed Hop, Triple Hop, and Crossover Hop tests based on the mean of two trials (leg symmetry index scores) in ACL reconstructed individuals.				

Ross 2002

Element	Study 1
1. Name of the instrument	Single Leg Hop Test
2. Version or way of operationalization	**Equipment:** Measuring tape.
	Preparatory action: One rater (experience not specified). Five-minute warm-up on a stationary bicycle ergometer and stretches of the lower extremities. One practice trial. Three measured trials, beginning with the non-involved leg. A 30 second rest period separated each trial.
	Procedure: Hop as far as possible forward and land on the tape measure. A 30 second rest period between each trial. The distance from the zero mark of the tape measure to the point where the subject`s heel hit the ground was measured.
	Assignment of the score: Mean of the three recorded trials was utilized for data analysis (leg symmetry index).
3. Construct	Physical function.
4. Measurement property	Reliability (rater n=1; intra-rater reliability) and measurement error.
5. Components that will be repeated	The whole test was repeated.
6. Source of variation	Occasion (5 days).
7. Patient population	ACL reconstruction (27 mo. post-op.), n=10 (30% females).
Comprehensive research question: What is the intra-rater reliability/standard error of measurement of the Single Leg Hop Test based the mean of three trials (leg symmetry index scores) in ACL reconstructed individuals.	

Supplementary Appendix 5. COSMIN methodological quality ratings and correlation coefficients of studies included in construct validity meta-analyses

Functional performance test	Study	n	COSMIN score	Quad.	Ham.	IKDC	KOOS sport/rec	KOOS QoL	ACL-RSI
Single Leg Hop	Bodkin 2017	51	Very good	-	-	0.17	0.22	0.19	-
	Chaput 2017	48	Very good	0.55	-	0.34	-	-	-
	Ebert 2021	50	Very good	0.47	-	0.15	-	-	0.17
	Harput 2016	90	Very good	-	-	-	-	-	0.26
	Harput 2018	72	Very good	-	-	0.50	0.30	0.45	-
	Hunnicutt 2020	30	Very good	0.76	-	-	-	-	-
	Jamshidi 2005	11	Adequate	0.85	-	0.60	-	-	-
	Keays 2003	31	Very good	0.66	0.25	-	-	-	-
	Menzer 2017	88	Very good	-	-	-	0.31	-	-
	Pua 2015	87	Very good	0.63	0.51	-	-	-	-
	Reinke 2011	69	Very good	-	-	0.32	0.24	0.20	-
	Ross 2002	50	Very good	0.30	-	-	-	-	-
	Sonesson 2021	68	Very good	-	-	0.31	-	-	-
	Sueyoshi	29	Very good	0.42	0.58	-	-	-	-
	Tunay 2008	37	Very good	0.38	-	-	-	-	-
	Xergia 2015	22	Very good	0.37	-0.11	-	-	-	-
6-m Timed Hop	Bodkin 2017	51	Very good	-	-	0.10	0.22	0.25	-
	Ebert 2021	50	Very good	0.38	-	0.38	-	-	0.32
	Pua 2015	87	Very good	0.64	0.53	-	-	-	-
	Reinke 2011	69	Very good	-	-	0.30	0.20	0.20	-
	Sonesson 2021	68	Very good	-	-	0.17	-	-	-
	Sueyoshi 2017	29	Doubtful	ns.	-	-	-	-	-
			Very good	-	0.43	-	-	-	-
Crossover Hop	Bodkin 2017	51	Very good	-	-	0.21	0.28	0.29	-
	Ebert 2021	50	Very good	0.46	-	0.18	-	-	0.22
	Jamshidi 2005	11	Adequate	0.93	-	0.78	-	-	-
	Menzer 2017	88	Very good	-	-	-	0.29	-	-
	Reinke 2011	69	Very good	-	-	0.20	0.20	0.10	-
	Sonesson 2021	68	Very good	-	-	0.25	-	-	-
	Sueyoshi 2017	29	Very good	0.40	0.46	-	-	-	-
Triple Hop	Bodkin 2017	51	Very good	-	-	0.17	0.18	0.10	-
	Ebert 2021	50	Very good	0.48	-	0.18	-	-	0.27
	Keays 2003	31	Very good	0.62	0.34	-	-	-	-
	Menzer 2017	88	Very good	-	-	-	0.31	-	-
	Reinke 2011	69	Very good	-	-	0.40	0.24	0.30	-
	Sonesson 2021	68	Very good	-	-	0.21	-	-	-
	Sueyoshi 2017	29	Very good	0.51	0.50	-	-	-	-
Vertical Hop	Ebert 2021	50	Very good	0.82	-	0.53	-	-	0.40
	Fischer 2017	169	Very good	0.52	-	-	-	-	-
	Jamshidi 2005	11	Adequate	0.66	-	0.70	-	-	-
	Laudner 2015	26	Very good	0.71	0.53	-	-	-	-
	Lee 2018	75	Adequate	0.30	0.16	-	-	-	-
			Very good	-	-	0.46	-	-	0.61
	O`Connor 2020	452	Adequate	-	-	-	-	-	0.18
	Tunay 2008	37	Very good	0.22	-	-	-	-	-

n = number of participants; mo. = months; post-op = post-operatively; post-inj. = post-injury.
Negative values (6-m Timed Hop) have been removed from table for consistent presentation of results.

Supplementary Appendix 6. COSMIN methodological quality ratings and correlation coefficients of studies included in qualitative synthesis for construct validity

Functional performance test	Study	n	COSMIN score	Quad	Ham	IKDC	KOOS sport/rec	KOOS QoL	ACL -RSI	Summary result
Medial Hop	Ebert 2021	50	Very good	0.71	-	0.53	-	-	0.39	2+ and 1-
Lateral Hop	Ebert 2021	50	Very good	0.53	-	0.33	-	-	0.22	2+ and 1-
Timed Speedy Hop	Ebert 2021	50	Very good	0.34	-	0.29	-	-	0.18	3-
10 Sec. Vertical Hop	Read 2020a	73	Very good	0.37	-	-	-	-	-	1-
Single Leg Drop Jump	O`Connor 2020	452	Adequate	-	-	-	-	-	0.16	1+ and 1
	Birchmeier 2019	52	Very good	0.66	-	-	-	-	-	
4-Jump Vertical Hop	Laudner 2015	26	Very good	0.63	0.39	-	-	-	-	1+ and 1-
Vertical Jump	Laudner 2015	26	Very good	0.64	0.58	-	-	-	-	2+
Step-Down	Roe 2021	66	Very good	-	-	0.39	0.40	0.43	-	3+
Carioca	Keays 2003	31	Very good	0.47	0.05	-	-	-	-	4+ and 1-
	Kong 2012	30	Adequate	0.40	0.48	0.45	-	-	-	
Shuttle Run	Keays 2003	31	Very good	0.50	0.15	-	-	-	-	3+ and 2-
	Kong 2012	30	Adequate	0.55	0.35	0.51	-	-	-	
Side-Step	Keays 2003	31	Very good	0.53	0.12	-	-	-	-	1+ and 1-
Co-Contraction	Kong 2012	30	Adequate	0.51	0.39	0.57	-	-	-	2+ and 1-
SEBT anterior	Clagg 2015	66	Doubtful	ns.	ns.	-	-	-	-	
	Harput 2016	90	Very good	-	-	-	-	-	0.09	2-
	Harput 2018	72	Doubtful	-	-	ns.	-	ns.		
			Very good	-	-	-	0.52	-	-	
SEBT post.med.	Clagg 2015	66	Doubtful	ns.	ns.	-	-	-	-	1-
	Harput 2016	90	Very good	-	-	-	-	-	0.27	
	Harput 2018	72	Doubtful	-	-	ns.	ns.	ns.	-	
SEBT post.lat.	Clagg 2015	66	Very good	0.28	0.29	-	-	-	-	4-
	Harput 2016	90	Very good	-	-	-	-	-	0.28	
	Harput 2018	72	Doubtful	-	-	ns.	-	ns.	-	
			Very good	-	-	-	0.51	-	-	
YBT anterior	Myers 2018	45	Adequate	0.59	0.49	-	-	-	-	2+ and 3-
	Roe 2021	66	Adequate	-	-	0.28	0.30	0.26	-	
YBT post.med.	Myers 2018	45	Adequate	0.50	0.58	-	-	-	-	1+ and 1-
YBT post.lat	Myers 2018	45	Adequate	0.29	0.45	-	-	-	-	1+ and 1-
YBT composite	Myers 2018	45	Adequate	0.27	0.36	-	-	-	-	1+ and 1-

For summary results: + = number of hypotheses confirmed, - = number of hypotheses not confirmed.
For overall rating: (+) = sufficient, (±) = inconsistent, (-) = insufficient.
Negative values (Carioca, Shuttle Run, Side-Step, Co-Contraction) have been removed from table for consistent presentation of results.
n = number of participants; IKDC = International Knee Documentation Committee Subjective Knee Form; ACL-RSI = Anterior Cruciate Ligament Return to Sport after Injury Scale; KOOS = Knee injury and Osteoarthritis Outcome Score; sport/rec = KOOS function in sport and recreation subscale; QoL = KOOS knee related quality of life subscale; ns. = not significant; SEBT = Star Excursion Balance Test; post.med = posteromedial; post.lat. = posterolateral; YBT = Y-Balance Test.

Appendix 7. Summary of Findings tables

Reliability and measurement error

Functional performance test	Summary or pooled result	Overall rating	Quality of evidence
Intra-rater			
Single Leg Hop (LSI)	ICC range 0.92 to 0.94; consistent results; sample size: 120	Sufficient	High
6-m Timed Hop (LSI)	ICC range 0.82 to 0.93; consistent results; sample size: 54	Sufficient	Moderate: due to imprecision (total n=50-100)
6-m Timed Hop (sec.)	ICC 0.96; sample size: 19	Sufficient	Very low: due to risk of bias (one adequate study) and imprecision (total n=<50)
Crossover Hop (LSI)	ICC range 0.84 to 0.94; consistent results; sample size: 54	Sufficient	Moderate: due to imprecision (total n=50-100)
Crossover Hop (m.)	ICC 0.98; sample size: 19	Sufficient	Very low: due to risk of bias (one adequate study) and imprecision (total n=<50)
Triple Hop (LSI)	ICC 0.88; sample size 35	Sufficient	Very low: due to risk of bias (one adequate study) and imprecision (total n=<50)
Vertical Hop (LSI)	ICC range 0.81 to 0.98; consistent results; sample size: 94	Sufficient	Moderate: due to imprecision (total n=50-100)
Vertical Hop (m.)	ICC 0.98; sample size: 19	Sufficient	Very low: due to risk of bias (one adequate study) and imprecision (total n=<50)
Stair Hop (LSI)	ICC 0.90; sample size: 19	Sufficient	Very low: due to risk of bias (one adequate study) and imprecision (total n=<50)
Stair Hop (sec.)	ICC 0.94; sample size: 19	Sufficient	Very low: due to risk of bias (one adequate study) and imprecision (total n=<50)
Inter-rater			
6-m Timed Hop (LSI, above or below 90%)	Kappa -0.11; sample size: 20	Insufficient	Very low: due to risk of bias (one adequate study) and imprecision (total n=<50)
Measurement error			
Single Leg Hop (LSI)	SEM 2.41 to 3.49; MDC 90% 8.09; MIC not defined	Indeterminate	-
6-m Timed Hop (LSI)	SEM 5.59; MDC 90% 12.96; 95% LoA -23.1, 26.8; MIC not defined	Indeterminate	-
6-m Timed Hop (sec.)	95% LoA -0.259, 0.254; MIC not defined	Indeterminate	-
Crossover Hop (LSI)	SEM 5.28; MDC 90% 12.25; MIC not defined	Indeterminate	-
Triple Hop (LSI)	SEM 4.32; MDC 90% 10.02; MIC not defined	Indeterminate	-

LSI = Leg Symmetry Index; ICC = intraclass correlation coefficient; sec. = seconds; m. = metres; SEM = standard error of measurement; MDC = minimal detectable change; LoA = Limits of Agreement; MIC = minimal important change.
For quality of evidence: starting point = high; Downgraded for imprecision by one level if total sample size = 50-100, two levels if total sample size = <50; Downgraded for risk of bias by one level if there is one study of adequate quality.

Validity and responsiveness

Functional performance test	Summary or pooled result	Overall rating	Quality of evidence
Construct validity			
Single Leg Hop	2 out of 6 hypotheses confirmed	Sufficient*	Low: due to inconsistency (very serious)
6-m Timed Hop	3 out of 6 hypotheses confirmed	Sufficient*	Moderate: due to inconsistency (serious)
Crossover Hop	2 out of 6 hypotheses confirmed	Sufficient*	Moderate: due to inconsistency (serious)
Triple Hop	2 out of 6 hypotheses confirmed	Sufficient*	Moderate: due to inconsistency (serious)
Vertical Hop	3 out of 4 hypotheses confirmed	Sufficient	Moderate: due to inconsistency (serious)
Medial Hop	2 out of 3 hypotheses confirmed	Sufficient*	Low: due to inconsistency (serious) and imprecision (total n=50-100)
Lateral Hop	2 out of 3 hypotheses confirmed	Sufficient*	Low: due to inconsistency (serious) and imprecision (total n=50-100)
Timed Speedy Hop	0 out of 3 hypotheses confirmed	Insufficient	Moderate: due to imprecision (total n=50-100)
10 Sec. Vertical Hop	0 out of 1 hypothesis confirmed	Insufficient	Moderate: due to imprecision (total n=50-100)
Single Leg Drop Jump	1 out of 2 hypotheses confirmed	Sufficient*	Low: due to inconsistency (serious) and imprecision (total n=50-100)
4-Jump Vertical Hop	1 out of 2 hypotheses confirmed	Sufficient	Very low: due to inconsistency (serious) and imprecision (total n=<50)
Vertical Jump	2 out of 2 hypotheses confirmed	Sufficient	Low: due to imprecision (total n=<50)
Step-Down	3 out of 3 hypotheses confirmed	Sufficient	Moderate: due to imprecision (total n=50-100)
Carioca	4 out of 5 hypotheses confirmed	Sufficient	Moderate: due to imprecision (total n=50-100)
Shuttle Run	3 out of 5 hypotheses confirmed	Inconsistent	-
Side-Step	1 out of 2 hypotheses confirmed	Inconsistent	-
Co-Contraction	2 out of 3 hypotheses confirmed	Inconsistent	-
SEBT anterior	0 out of 2 hypotheses confirmed	Insufficient	High
SEBT posteromedial	0 out of 1 hypothesis confirmed	Insufficient	Moderate: due to imprecision (total n=50-100)
SEBT posterolateral	0 out of 4 hypotheses confirmed	Insufficient	High
YBT anterior	2 out of 5 hypotheses confirmed	Inconsistent	-
YBT posteromedial	1 out of 2 hypotheses confirmed	Inconsistent	-
YBT posterolateral	1 out of 2 hypotheses confirmed	Inconsistent	-
YBT composite	1 out of 2 hypotheses confirmed	Inconsistent	-
Structural validity			
Test battery	No info available	Indeterminate	-
Responsiveness			
Single Leg Hop	1 out of 1 hypothesis confirmed	Sufficient	Low: due to imprecision (total n=<50)
6-m Timed Hop	0 out of 1 hypothesis confirmed	Insufficient	Low: due to imprecision (total n=<50)
Crossover Hop	1 out of 1 hypothesis confirmed	Sufficient	Low: due to imprecision (total n=<50)
Triple Hop	0 out of 1 hypothesis confirmed	Insufficient	Low: due to imprecision (total n=<50)

*Overall rating based on correlations with quadriceps and hamstrings strength (downgraded for inconsistency).
For quality of evidence: starting point = high; Downgraded for inconsistency by one level (serious) or two levels (very serious); Downgraded for imprecision by one level if total sample size = 50-100 or two levels if total sample size = <50.

Supplementary Appendix 8. Description of studies published prior to 2000 and their results

Author (Year)	Sample size	Females, no. (%)	Age, mean ± SD	Population	Timing of functional performance test, mean ± SD	Functional performance test	Measurement property assessed and result
							Construct validity*
Borsa (1998)	29	14 (48)	29 ± 2	ACLD	42 ± 12 mo. post-inj.	Single Leg Hop	*r*=0.06 (quad)
Lephart (1992)	41	9 (22)	23	ACLR	27 mo. post-op.	Carioca	*r*=0.30 (quad) 0.22 (ham)
						Shuttle run	*r*=0.42 (quad) 0.23 (ham)
						Cocontraction	*r*=0.20 (ham) 0.16 (quad)
Petschnig (1997)	62	0 (0)	B:28 ± 2 C:29 ± 1	ACLR	13 ± 1 mo. post-op. 55 ± 4 mo. post-op	Single Leg Hop Triple Hop	B: *r*=0.50, C: *r*=0.38 (quad) B: *r*=0.39, C: *r*=0.36 (quad)
						Vertical Hop	B: *r*=0.48, C: *r*=0.50 (quad)
Petschnig (1998)	55	0 (0)	B:28 ± 9 C:30 ± 6	ACLR	13 ± 3 wks. post-op. 54 ± 13 wks. post-op.	Single Leg Hop Triple Hop	B: *r*=0.45, C: *r*=0.51 (quad) B: *r*=0.48, C: *r*=0.55 (quad)
						Vertical Hop	B: *r*=0.01, C: *r*=0.51 (quad)
Wilk (1994)	50	16 (32)	25	ACLR	26 wks. post-op.	Single Leg Hop	*r*=0.62 (quad)
						6m Timed Hop	*r*=0.60 (quad)
						Crossover Hop	*r*=0.69 (quad)
							Intra-reliability
Brosky (1999)	15	0 (0)	26 ± 7	ACLR	29 ± 9 mo. post-op.	Single Leg Hop 6m Timed Hop Vertical Hop	ICCs 0.88-0.97
Paterno (1996)	13	3 (23)	22 ± 3	ACLR	33 ± 12 wks. post-op.	Single Leg Hop	ICC 0.89
							Reliability (not specified)
Kramer (1992)	38	16 (42)	25 ± 7	ACLR	22 ± 7 mo. post-op.	Single Leg Hop	ICC 0.93
							Structural validity**
Risberg (1994)	35	18 (51)	26 ± 8.2	ACLR	18 ± 3.9 mo. post-op	Vertical Jump	Factor 1: 0.79 Factor 2: 0.02
						Figure-of-eight	Factor 1: 0.83 Factor 2: 0.37
						Stairs-Running	Factor 1: 0.86 Factor 2: 0.22
						Triple Hop	Factor 1: 0.33 Factor 2: 0.86
						Stairs Hopple	Factor 1: 0.03 Factor 2: 0.88

*Parenthesis indicates comparator instrument;
**The tests loaded on two factors representing a common function: Two-legged tests (Factor 1) and One-legged tests (Factor 2).
Negative values (Carioca, Shuttle Run, Co-Contraction, 6m Timed Hop) have been removed for consistent presentation of results.
ACLR = anterior cruciate ligament reconstruction; mo. = months; post-op. = post-operatively; ACLD = anterior cruciate ligament deficient; post-inj. = post-injury; wks. = weeks.