Clinician-friendly lower extremity physical performance measures in athletes: a systematic review of measurement properties and correlation with injury, part 1. The tests for knee function including the hop tests

Eric J Hegedus,[1] Suzanne McDonough,[2] Chris Bleakley,[3] Chad E Cook,[4] G David Baxter[5]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bjsports-2014-094094).

[1]Department of Physical Therapy, High Point University, High Point, North Carolina, USA
[2]Centre for Health and Rehabilitation Technologies, School of Health Sciences, Institute of Nursing and Health Research, University of Ulster, Newtonabbey, County Antrim, UK
[3]Ulster Sports Academy, Sport and Exercise Sciences Research Institute, University of Ulster, Carrickfergus, UK
[4]Division of Physical Therapy, Duke University, Durham, North Carolina, USA
[5]School of Physiotherapy, University of Otago, Dunedin, New Zealand

Correspondence to
Dr Eric J Hegedus, Department of Physical Therapy, High Point University, 833 Montlieu Ave, High Point, NC 27262, USA; ehegedus@highpoint.edu

Received 29 July 2014
Revised 19 November 2014
Accepted 21 November 2014
Published Online First 10 December 2014

ABSTRACT

Objective To review the measurement properties of physical performance tests (PPTs) of the knee as each pertain to athletes, and to determine the relationship between PPTs and injury in athletes age 12 years to adult.

Methods A search strategy was constructed by combining the terms 'lower extremity' and synonyms for 'performance test', and names of performance tests with variants of the term 'athlete'. In this, part 1, we report on findings in the knee. The Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines were followed and the *Co*nsensus-based *S*tandards for the selection of health *M*easurement *In*struments (COSMIN) checklist was used to critique the methodological quality of each paper. A second measure was used to analyse the quality of the measurement properties of each test.

Results In the final analysis, we found 29 articles pertinent to the knee detailing 19 PPTs, of which six were compiled in a best evidence synthesis. The six tests were: one leg hop for distance (single and triple hop), 6 m timed hop, crossover hop for distance, triple jump and single leg vertical jump. The one leg hop for distance is the most often studied PPT. There is conflicting evidence regarding the validity of the hop and moderate evidence that the hop test is responsive to changes during rehabilitation. No test has established reliability or measurement error as assessed by the minimal important change or smallest detectable change. No test predicts knee injury in athletes.

Conclusions Despite numerous published articles addressing PPTs at the knee, there is predominantly limited and conflicting evidence regarding the reliability, agreement, construct validity, criterion validity and responsiveness of commonly used PPTs. There is a great opportunity for further study of these tests and the measurement properties of each in athletes.

To cite: Hegedus EJ, McDonough S, Bleakley C, *et al*. *Br J Sports Med* 2015;**49**:642–648.

INTRODUCTION

Tests of physical performance are employed at multiple levels and throughout the sporting world.[1–3] These tests, in combination, are being used more frequently as part of pre-season screening, although test findings appear to be more specific than sensitive.[4 5] The advantage of physical performance tests (PPTs) is that the tests are easy to administer, are not time consuming and do not require a great deal of expertise. Further, PPTs do not require expensive equipment, and can be completed in multiple settings and locations.

For PPTs to be useful as outcome measures, we need to know what constitutes a meaningful change in score. Further, these tests should possess some key measurement properties such as reliability, validity and responsiveness. A meaningful change in score is often captured by the minimal clinically important difference or the minimal important change (MIC), which is the smallest change in a score detectable by the patient.[6] The MIC should be greater than the minimal detectable change in order for the PPT to identify a relevant change in the patient's status. Reliability is the degree to which a measurement is free from error.[7] The interested reader is also directed to Davidson's discussion of these topics.[8]

Validity discerns whether a test measures what it is intended to measure.[7] There are different types of validity. Criterion validity is a measure of how well the PPT under investigation correlates with a gold or criterion standard. Included in criterion validity is predictive validity, which would be, for example, how well a PPT predicts an outcome such as injury. Construct validity, the degree to which a PPT correlates with a latent construct such as strength or function, can be of either a convergent or divergent/discriminant nature.[7] In convergent validity, one would expect a PPT that measures function to correlate well with, say, another test of function such as an established self-report measure. Discriminant validity is the opposite: one would expect low correlation between two measures that assess different constructs. Whether PPTs provide useful information is of some debate[9 10] and whether each test possesses the necessary measurement properties to be considered a valuable outcome measure is also a matter of contention.[11–14]

To examine the evidence behind individual PPTs, we conducted a systematic review of measures typically used to assess lower extremity performance in athletes. Our goals in conducting this systematic review were to coalesce the literature on PPTs, subject the literature and measurement properties to a quality analysis, and provide a best evidence synthesis. We hypothesised that PPTs would have

moderate evidence regarding their measurement properties but have little or no ability to predict injury in athletes.

METHODS

Using the PICO method, we established our research question as to whether individual PPTs of the lower extremity have any relationship to injury in athletes, age 12 years to adult (no limit). We then operationally defined PPTs as measures that assess components of sport function (strength, power, agility), determine readiness for return to sport, or predict injury of the lower extremity; and as measures that can be performed field side, courtside, or in a gym with affordable, portable and readily available equipment.

Specifically, this operational definition excluded studies that made use of three-dimensional motion capture, force plates, timing gates, treadmills, stationary bikes, metabolic carts or any other form of non-portable, unaffordable testing device. Also, this definition excluded tests of which the sole purpose was to judge movement quality or range of motion, such as the unloaded double leg squat.

We defined athletes as those individuals at level 5 or above on the Tegner scale.[15] We chose level 5 because the predominance of literature on PPTs pertains to the knee, and level 5 is the lowest level in which competitive athletes are still encompassed. In articles where the Tegner scale was not used, we accepted the terms 'recreational athlete', 'sports participation', 'intramural athlete' as indicative of level 5 activity. We also included studies where 50% or more of the participants were at Tegner level 5 or above. For articles where there was confusion between the authors about inclusion or exclusion, a consensus was reached among all authors through discussion and majority vote.

We followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA)[16 17] guidelines and the *C*onsensus-based *S*tandards for the selection of health *M*easurement *In*struments (COSMIN) checklist[18] to critique the methodological quality of each paper.

After the fact and in order to make this review more publishable, we elected to divide the reporting into two subject categories: part 1, the knee; and part 2, the rest of the lower extremity. To be included in the knee review, the studies had to identify the knee or a knee injury as the focal point of the paper. In lieu of obvious identification of the knee as the primary focus, we reasoned that correlations with knee-related outcome measures or correlational studies with constructs, such as strength as measured by knee flexor and extensor torque, should be included.

Search strategy

A search was performed in PubMed, CINAHL and SportDiscus for all dates up to 13 January 2014. The full PubMed search strategy is described in online supplementary appendix A. Systematic reviews were then located using the 'Clinical Queries' option of PubMed and the references cited in these reviews were examined for appropriate articles for inclusion. Finally, after the selection of the final studies, as outlined below, citations from these articles that appeared pertinent were read in full to determine their appropriateness for inclusion.

Study selection

The process by which studies were selected is outlined in figure 1. Two authors (EJH and CB) read the titles and abstracts of all citations from the three search engines in order to determine which articles to read in full. A third author (SM) resolved disputes between these authors. One author (EJH) then read the complete text of all remaining articles whereas all other authors read the same studies based on their area of expertise so that two researchers read all articles in full.

Data extraction and analysis of quality

Each of the studies included in the final analysis was read three times for the purposes of: (1) data extraction, (2) assessment of methodological quality and (3) assessment of the quality of the measurement properties of each PPT.

For data extraction, we chose to group the data in two ways. First, a 'Study Summary' was created (see online supplementary table S1), which summarises the study population, PPTs, aims and results of each study. Next, we examined the names of the PPTs and the methodology of each study to determine whether certain tests were used more often, and if there was a consensus in how the tests were labelled and performed (see online supplementary table S2).

Methodological quality was critiqued using the COSMIN four-point scoring system (excellent, good, fair, poor) designed for systematic reviews[19] with the worst score serving as the global score in each subsection. In addition, we followed the adaptations to COSMIN for a review on PPTs as described previously (see online supplementary appendix B).[6] Quality of measurement properties including reliability, measurement error, hypothesis testing/construct validity, criterion validity (including predictive validity) and responsiveness (both internal and external) were assessed using a rating scale of 'positive', 'indeterminate' and 'negative' for each property (see online supplementary appendix C).[20] For both these steps, one author (EJH) applied the adapted COSMIN checklist for methodological quality and quality criteria to all final articles while each of the other authors did the same based on their area of expertise so that each article had at least two authors performing quality assessment. In the event that these two authors disagreed in their assessment, feedback was obtained from the other authors and a consensus was reached. Because there was a large volume of data accrued during this process, the final included studies were separated by region into hip, thigh, knee, ankle and entire lower extremity for the first three steps: data extraction, assessment of methodological quality and assessment of the quality of the measurement properties of each PPT. All studies pertaining to the knee are presented in this paper, whereas studies pertaining to the rest of the lower extremity are presented in part 2 of this series.

The fourth and final step, a best evidence synthesis, requires combining the information from findings regarding the methodological quality and the quality of measurement properties. The best evidence synthesis was subcategorised by PPT. In this grand summary, only studies with fair, good or excellent methodological quality were included, and the evidence for each test was rated as 'strong', 'moderate', 'limited', 'conflicting' and 'unknown'.[20 21] We used 'unknown' to indicate that either there was no evidence of the statistical property or that there was evidence, but only in studies of poor methodological quality. Further, for the synthesis, only PPTs with somewhat consistent descriptions from study to study, across at least two studies, were considered for the synthesis. The evidence from studies with sample size less than 30 participants without an a priori power analysis was classified as limited evidence.[6]

RESULTS

Included studies, tests and testing procedures

One hundred and sixty-nine articles were read in full and 60 studies were considered for analysis. Almost without exception, studies were eliminated based on the fact that there were few or

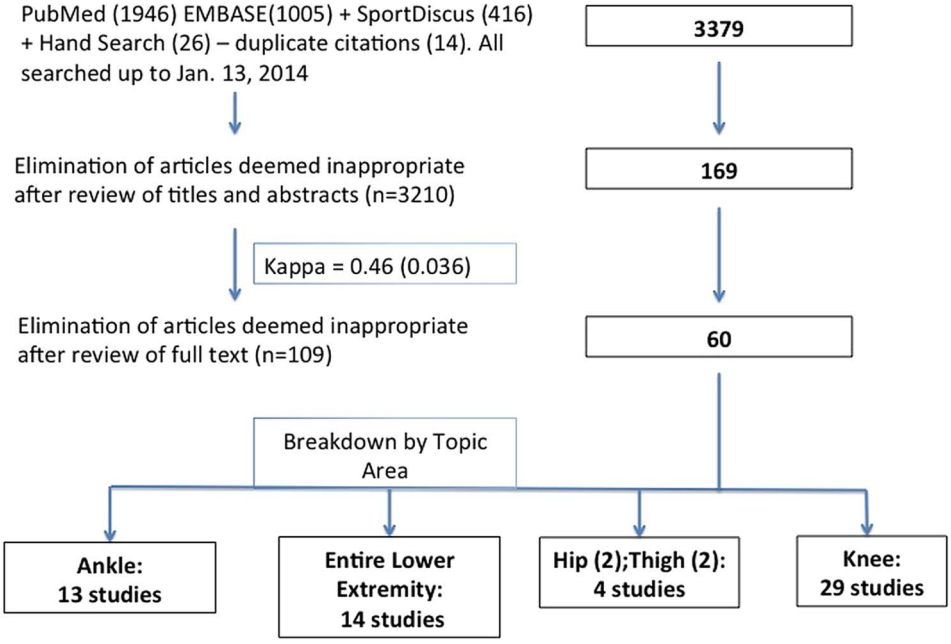

Figure 1 Process for selecting studies.

no athletes in the subject pool or because the examiners used equipment to conduct the study that would not be regularly available to most practitioners such as electronic timing gates.

Twenty-nine of the final 60 studies pertinent to the knee were included in this systematic review (figure 1). These studies reported on the properties of 19 different tests, of which 8 were examined in more than one study and, therefore, compiled in a final evidence synthesis. The most common PPTs studied were:

- one leg hop for distance: single hop (24 studies);[11–14 22–41]
- 6 m timed hop (9 studies);[11–14 23 25 27 28 40]
- crossover hop for distance (9 studies);[11–14 28 33 40 42 43]
- one leg hop for distance: triple hop (7 studies);[11–14 28 33 44]
- single leg vertical jump (7 studies);[23 25 32 34 42 43 45]
- single leg squat (5 studies);[34 42–45]
- figure of eight run (3 studies);[26 42 43]
- triple jump (3 studies).[30 32 34]

For the eight most common tests, there is great variation in what the tests are named, and in the procedures by which the tests are to be completed (see online supplementary table S2). As an example, the one leg hop for distance is the most commonly reported PPT in the literature. Where these were reported, the warm-up and number of practice hops varied widely. The number of hops comprising the test varied from 1 to 3 to 10. How the arms are to be used during the test is not standardised and the final scoring can be based on the mean of two attempts, the greater of two attempts, the greatest of three attempts, or the greatest of three successful trials. This vast variability was not limited to the one leg hop for distance; most other PPTs of the knee also demonstrated marked inconsistency.

Summary of the methodological quality of included studies

Reliability

The methodological quality of studies examining reliability of PPTs at the knee is generally poor regardless of the PPT studied (table 1; online supplementary appendix B). Only one[42] of eight total studies addressing reliability had a fair level of evidence. Bjorklund *et al*[42] reported an inter-rater reliability of $\kappa=0.75$ for the single leg vertical jump which was repeated five times and incorporated a qualitative rating of 'springiness'. There is no study with high methodological quality that examines the single leg vertical leap as it is more traditionally performed measuring a maximum jump height off of one leg.

Agreement/measurement error

No studies currently exist that have looked at the relationship of the MIC or smallest detectable change (SDC) to the limits of agreement.

Table 1 Summary of methodological quality by statistical property by test

	Statistical property				
Test	**Reliability**	**Agreement**	**Hypothesis testing**	**Criterion validity**	**Responsiveness**
One leg hop for distance: 1 hop	Poor	No studies	Fair	Good	Poor
One leg hop for distance: 3 hops	Poor	No studies	Poor	Good	No studies
6 m timed hop	Poor	No studies	Poor	Good	No studies
Crossover hop for distance	Fair	No studies	Poor	Good	Good
Triple jump	No studies	No studies	Fair	No studies	Poor
Single leg vertical jump	Fair	No studies	Mixed—good to poor	Mixed—good to poor	Mixed—good to poor

Summary quality ratings above are based on the most frequent quality rating in each category. For physical performance tests where the evidence was mixed, a range of quality ratings was given.

Hypothesis testing/construct validity

For the one leg hop for distance using a single hop, the methodological quality of the 16 studies[11 22–24 26 28 29 31 33–36 38–41] was generally fair and for the version that requires three consecutive hops (triple hop), the methodological quality was poor in two[11 28] of three[44] studies. Likewise, the 6 m timed hop and crossover hop for distance generally were studied in articles of poor methodological quality. In one study[34] that examined the convergent validity of the triple jump and isokinetic quadriceps testing, a low correlation between the two variables was found. In this study[34] of fair methodological quality, the authors concluded that functional testing and isokinetic strength testing of the quadriceps reflected two different constructs. Hypothesis testing for the single leg vertical leap was from mixed quality articles including one good,[43] one fair[23] and one poor.[42] No evidence exists with regard to the construct validity of the stair hop test.

Criterion validity

There is predominantly good-quality evidence for the criterion validity of PPTs at the knee. The exception was the single leg vertical jump where the evidence quality was mixed with one study of poor[42] and one of good[43] quality.

Responsiveness

Five studies[26 30 32 37 43] reported on the responsiveness of five PPTs at the knee; however, only one study[43] demonstrated good methodological quality. The two PPTs studied in this article[43] were the five-repetition single leg vertical leap and the crossover hop for distance.

Summary of the quality of the measurement properties

Reliability

Four studies[25 27 35 39] examined test–retest reliability of the hop test and all studies scored a positive measurement property quality rating (see online supplementary appendix C). For the other tests, reliability was examined in two studies for 6 m timed[25 27] and one study each for the single leg vertical,[42] the hop with three leaps (triple hop)[44] and the crossover hop for distance.[42]

Agreement/measurement error

There are no data available about the quality of the measurement properties of MIC or SDC with regard to PPTs in athletes.

Hypothesis testing/construct validity

The quality rating of construct validity for the hop test is generally positive when examining discriminant validity[22 29 33 38] and generally negative when describing convergent validity.[26 28 29 34–36 39 40] In examining the other PPTs, such dichotomous quality ratings, based on whether discriminant or convergent validity is examined, continue almost without exception.

Criterion validity

With regard to the hop test and the ability of the test to predict function, two studies[12 13] found a positive quality rating and two [14 27] negative quality ratings. Likewise, the 6 m timed hop showed both a positive[14] and a negative[13] quality rating with regard to predicting function.

Responsiveness

The hop test,[26 32 37] single leg vertical jump,[32 43] crossover hop[43] and triple jump[32] have a positive quality rating and appeared to change with rehabilitation after knee injury. However, according to one study,[30] the hop for distance, triple jump and stair hop were not responsive to neuromuscular training in an anterior cruciate ligament (ACL) tear prevention programme.

Best evidence synthesis by PPT

The best evidence synthesis is summarised in table 2. Worth noting again is that for this synthesis, only studies of fair or better methodological quality were considered. Also, the PPT could not vary a great deal from the usual description (eg, 10 hops instead of 1), and PPTs that did not have more than one study examining their properties were eliminated from the synthesis. Adhering to these tenets eliminated the figure of eight run and the single leg squat; this left six PPTs available for the synthesis.

Grading key

Unknown: investigated in studies of exclusively poor methodology or not investigated in any study.
Strong: multiple studies of good methodological rating or at least one study of excellent methodology.
Moderate: multiple fair methodological studies or one study of good methodology.
Limited: one study of fair methodological quality.
Conflicting: contradictory findings.

One leg hop for distance (1 hop)

Although four studies demonstrated test–retest reliability, all were of poor quality, meaning that in the final analysis, evidence of the reliability of the one leg hop for distance in athletes is unknown. Likewise, agreement as represented by the MIC or SDC is unknown. With regard to hypothesis testing/construct validity and criterion validity, the evidence is conflicting.

As a reminder, construct validity can be subdivided into discriminant validity, low correlations with tests that are expected to test different constructs and convergent validity; the results of two tests examining the same construct will be highly correlated. The hop tests generally displayed discriminant validity but seldom displayed convergent validity. Thus, the hop test differentiates between a normal and not normal knee regardless of whether the difference in performance is between an ACL-repaired (ACLR) knee and the uninvolved knee in the same person,[22] the ACLR knee and the uninvolved knee in age-matched normals,[33] or the ACL-deficient (ACLD) knee and the uninvolved knee in age-matched normals.[38] Although the gender mix was not specified in one study,[33] the other two studies[22 38] have all male participants, giving these results limited generalisability. Further, the hop may not discriminate at all once the athlete is 2 years or longer after surgery. In two long-term follow-up studies examining participants with ACLR, the hop test was unable to discriminate between the operative and non-operative knee[41] or between competitive and non-competitive athletes with ACLR.[31]

In contrast to its discriminative ability, the hop test does not correlate well with other measures that attempt to capture function or strength. Several studies examined the correlation between patient self-report measures of function and the hop test. One study[23] of fair methodological quality reported a significant correlation between self-reported function (ability to run, sprint, jump, land, cut and twist) and the hop test, but these authors concluded that such self-ratings alone were not strong enough in isolation to be predictors of function. In all other cases, the hop test failed to correlate with or explained

Table 2 Synthesis of evidence by test

Measurement property	Unknown (???)	Strong (+++) (−−−)	Moderate (++) (−−)	Limited (+) (−)	Conflicting (±)
One leg hop for distance: 1 hop					
Reliability	???				
Agreement	???				
Hypothesis testing					±
Criterion validity					±
Responsiveness			++		
One leg hop for distance: 3 hops					
Reliability	???				
Agreement	???				
Hypothesis testing	???				
Criterion validity					±
Responsiveness	???				
6 m timed hop					
Reliability	???				
Agreement	???				
Hypothesis testing				+	
Criterion validity					±
Responsiveness	???				
Crossover hop for distance					
Reliability				−	
Agreement	???				
Hypothesis testing				+	
Criterion validity					±
Responsiveness				+	
Triple jump					
Reliability	???				
Agreement	???				
Hypothesis testing				−	
Criterion validity	???				
Responsiveness	???				
Single leg vertical jump					
Reliability	???				
Agreement	???				
Hypothesis testing				+	
Criterion validity	???				
Responsiveness	???				

only a small amount of the variance in self-rated functional outcomes.[35 36 39] In other words, results of the hop test generally fail to predict functional outcomes. There is also no evidence that results of the hop test predict injury. In addition to the failure of the hop test to correlate with self-report measures, it seems to assess a different construct than strength as measured by isokinetic torque production. Although one study[23] found a correlation between isokinetic quadriceps weakness at 60°/s and lower hop scores, two other studies found no correlation between the hop test and either quadriceps torque at 60°,[34] 90°,[35] or 180°/s[34], or hamstring torque at 60° or 180°/s.[34]

Finally, with regard to responsiveness, there is moderate evidence from one good[37] and one fair[26] quality study that the hop test is responsive. The hop test displays internal responsiveness since outcomes improve as the athlete progresses through rehabilitation.

One leg hop for distance (3 hops)/triple hop

Evidence regarding the one leg hop for distance with three hops, most commonly known as the triple hop, is largely inconclusive. The only evidence currently available regarding the measurement properties of the triple hop is that the test has conflicting criterion validity. Three studies, all in patients with ACL deficiency, found that the triple hop does not predict which athletes will be able to cope with ACLD[28] nor does it predict function at 1 year[13] as captured by the International Knee Documentation Committee (IKDC) form,[46] self-rated global function, or the Knee Outcome Survey-Activities of Daily Living (KOS-ADL) Scale.[14 47] Another study that used the IKDC as a functional outcome measure, found mixed results: the triple hop performed at baseline had no ability to predict function 1 year after ACLR while a triple hop performed at 6 months postoperation did predict 1 year self-reported function.[12] No studies are available that investigated whether triple hop results predict injury.

The 6 m timed hop

Similar to the triple hop, the reliability, agreement, responsiveness and ability of the 6 m timed hop to predict injury are unknown and the evidence about criterion validity is conflicting. This PPT does not appear to predict a change in usual or worst pain,[27] who will cope with an ACL tear,[28] or what sort of

functional outcome will be attained,[12 13] nor is it sensitive enough to detect asymmetry in patients who are ACLD.[11] However, the 6 m timed hop performed at 6 months after surgery does predict self-rated functional outcome at 1 year.[12] In one study of fair methodological quality,[23] the 6 m timed hop correlated well with self-reported limitations in running, twisting, cutting, sprinting and jumping/landing. Therefore, the evidence regarding construct validity is positive but limited.

Crossover hop for distance

Evidence about agreement and the crossover hop is unknown and reliability has limited negative evidence.[42] However, there is limited but positive evidence with regard to construct validity and responsiveness. Bjorklund *et al*[43] found the crossover hop to possess discriminant validity in that the test can detect differences in the surgically repaired knee and the unaffected knee at 4 as well as 8 months after ACL repair. These same authors found a moderate effect size with regard to detecting change post-ACLR with rehabilitation at the 4-month and 8-month marks. Finally, there is conflicting evidence about the criterion validity of the crossover hop. This PPT does not appear to be a predictor of self-rated function[12–14] nor is it sensitive enough to detect abnormal limb symmetry in an ACLD population.[11] However, test results make up one variable that helps predict who will cope with an ACL deficiency,[28] and when the test is performed at 6 months after ACLR, it correlates with self-reported function at 1 year.[12] There were no studies that examined the ability of the crossover hop for distance to predict knee injury in athletes.

Triple jump

Evidence regarding the reliability, agreement, criterion validity and responsiveness of the triple jump is unknown; however, one study of fair methodology reported on construct validity and found a negative correlation with isokinetic testing of the quadriceps and hamstrings.[34] There were no studies that examined the ability of the triple jump to predict knee injury in athletes.

Single leg vertical jump

As in the triple jump, evidence regarding the reliability, agreement, criterion validity and responsiveness of the single leg vertical jump is unknown. There is limited positive evidence of the construct validity of the test. One study[23] demonstrated a correlation of the single leg vertical jump with self-assessed difficulty in pivoting and cutting, isokinetic quadriceps weakness and patellofemoral compression pain. Importantly, one study[43] of good methodology was eliminated from the synthesis because the methodology (5 consecutive hops with a qualitative evaluation of 'springiness') was significantly different from the usual (maximum jump height on a single effort). There is no evidence that results on the single leg vertical jump predict injury.

DISCUSSION

Eight PPTs were studied by more than one group of authors and six were further examined in the best evidence synthesis. The methodological quality of the tests ranged from poor to good and when combined with the quality of the measurement properties, the level of evidence was generally limited or conflicting.

The exception to this trend was the responsiveness of the one leg hop for distance where evidence of responsiveness was moderately positive. The hop test displays internal responsiveness and can be used to track rehabilitation progress.

Other rather significant findings emerged as a result of the best evidence synthesis. First, the naming of PPTs and the methods by which each is conducted vary greatly. There is a clear and urgent need to standardise terminology and methodology of these performance tests for the sports and orthopaedic community. The advantages of PPTs are their simplicity to conduct and interpret; as a consequence, these are routinely used by coaches, researchers, physical therapists and physicians. The lack of standardised terminology and methodology impairs communication and limits the generalisability of findings.

Second, the clinical applicability of the PPTs can certainly be questioned since we know very little about the measurement properties. No PPT for athletes with knee pathology displays reliability, agreement, construct validity, criterion validity and responsiveness. In fact, only the one leg hop for distance, 6 m timed hop, and crossover hop possess more than one of these measurement properties and we are unsure of the MIC or the SDC of any of these tests, thus limiting the value of these tests as outcome measures in the clinic. Further, the only information about the reliability of these tests is that the triple jump may lack reliability.

Third, results regarding construct validity seem to be mostly dichotomous; these PPTs display divergent or discriminative validity but seem to lack convergent validity. In other words, if the clinical goal is to detect differences between an uninvolved knee (healthy) and an involved (surgery or ACLD) knee, many of these single legged tests are helpful. However, if the goal is to correlate these PPTs with strength (isokinetic quadriceps or hamstrings torque) or to the patient's own estimation of function (self-report outcome measure), then, generally speaking, these tests would fail. Poor association may not be a negative characteristic but rather a reflection that self-report of function, strength measured isokinetically and function as captured by a PPT are simply different constructs.[48 49]

Finally, criterion validity has mixed evidence based on the ability of the studied PPTs to predict functional outcome. The hop and 6 m timed hop appear to be the best PPTs at predicting function as measured by self-report outcome measure.[13 14] The answer to the question of whether any of the PPTs predict injury in athletes remains unknown.

LIMITATIONS

As with any systematic review, there are limitations that need to be acknowledged. First, although the COSMIN checklist has been used in several reviews of PPTs, the checklist was originally developed for reviews of questionnaire-based self-report measures and, therefore, the measurement properties of the COSMIN itself can be questioned.[6 21 50] Also, there is no standardised search strategy for PPTs and we limited our results to studies published in English, therefore, the possibility exists that some information about these tests was missed or overlooked. Finally, most of the injured populations in the included studies had an ACL tear, which limits the generalisability of our findings.

CONCLUSIONS

Physical performance tests are used widely by a broad array of professionals seeking to gather information about rehabilitation progression, symmetry between legs and risk for injury. Despite the ubiquity of PPTs in the literature, the paucity of evidence on measurement properties, the wide array of test methodologies and the lower methodological quality of the studies in the field indicate that there is ample opportunity for research in this area. Until more is known about these PPTs, caution is urged in making any firm clinical conclusions based on their results and in deciding whether an observed change in these outcome measures is meaningful.

Summary box

- There are six physical performance tests (PPTs) pertinent to the knee that have been substantially studied so that we have some idea of their metrics (reliability, agreement, validity, responsiveness) in an athletic population: the one leg hop for distance, the triple hop for distance, the 6 m timed hop, the crossover hop for distance, the triple jump, and the single leg vertical leap.
- The one leg hop for distance is the most studied PPT at the knee and yet we know only that this test is discriminative in males with ACL tears and that it is responsive to rehabilitation after ACL tear.
- For all other PPTs at the knee, there is limited, conflicting or unknown evidence regarding their measurement properties.
- The ability of PPTs to predict knee injury is unknown.
- Caution is urged in making any firm clinical conclusions based on the results of PPTs when testing the knee and in deciding whether an observed change in these outcome measures is meaningful in athletes.

Acknowledgements The authors would like to acknowledge Ms Connie Schardt, MLS, AHIP, FMLA, for her assistance with search strategies.

Contributors EJH, SM and CB planned the study, reviewed the citations, examined the articles for quality and edited the final manuscript. GDB and CC examined articles for quality and edited the manuscript. EJH wrote the manuscript.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The authors are happy to share on receipt of a written request by the corresponding author.

REFERENCES

1 Noyes FR, Barber-Westin SD, Fleckenstein C, *et al*. The drop-jump screening test: difference in lower limb control by gender and effect of neuromuscular training in female athletes. *Am J Sports Med* 2005;33:197–207.
2 Frohm A, Heijne A, Kowalski J, *et al*. A nine-test screening battery for athletes: a reliability study. *Scand J Med Sci Sports* 2012;22:306–15.
3 Smith HC, Johnson RJ, Shultz SJ, *et al*. A prospective evaluation of the Landing Error Scoring System (LESS) as a screening tool for anterior cruciate ligament injury risk. *Am J Sports Med* 2012;40:521–6.
4 Chorba RS, Chorba DJ, Bouillon LE, *et al*. Use of a functional movement screening tool to determine injury risk in female collegiate athletes. *N Am J Sports Phys Ther* 2010;5:47–54.
5 Kiesel K, Plisky PJ, Voight ML. Can serious injury in professional football be predicted by a preseason functional movement screen? *N Am J Sports Phys Ther* 2007;2:147–58.
6 Bartels B, de Groot JF, Terwee CB. The six-minute walk test in chronic pediatric conditions: a systematic review of measurement properties. *Phys Ther* 2013;93:529–41.
7 Portney LG, Watkins MP. *Foundations of clinical research: applications to practice*. 3rd edn. Upper Saddle River, NJ: Pearson Prentice Hall, 2013.
8 Davidson M, Keating J. Patient-reported outcome measures (PROMs): how should I interpret reports of measurement properties? A practical guide for clinicians and researchers who are not biostatisticians. *Br J Sports Med* 2014;48:792–6.
9 Freckleton G, Pizzari T. Risk factors for hamstring muscle strain injury in sport: a systematic review and meta-analysis. *Br J Sports Med* 2013;47:351–8.
10 Gabbe BJ, Finch CF, Wajswelner H, *et al*. Predictors of lower extremity injuries at the community level of Australian football. *Clin J Sport Med* 2004;14:56–63.
11 Noyes FR, Barber SD, Mangine RE. Abnormal lower limb symmetry determined by function hop tests after anterior cruciate ligament rupture. *Am J Sports Med* 1991;19:513–18.
12 Logerstedt D, Grindem H, Lynch A, *et al*. Single-legged hop tests as predictors of self-reported knee function after anterior cruciate ligament reconstruction: the Delaware-Oslo ACL cohort study. *Am J Sports Med* 2012;40:2348–56.
13 Grindem H, Logerstedt D, Eitzen I, *et al*. Single-legged hop tests as predictors of self-reported knee function in nonoperatively treated individuals with anterior cruciate ligament injury. *Am J Sports Med* 2011;39:2347–54.
14 Hurd WJ, Axe MJ, Snyder-Mackler L. A 10-year prospective trial of a patient management algorithm and screening examination for highly active individuals with anterior cruciate ligament injury: part 2, determinants of dynamic knee stability. *Am J Sports Med* 2008;36:48–56.
15 Tegner Y, Lysholm J. Rating systems in the evaluation of knee ligament injuries. *Clin Orthop Relat Res* 1985;198:43–9.
16 Liberati A, Altman DG, Tetzlaff J, *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* 2009;339:b2700.
17 Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* 2009;339:b2535.
18 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.
19 Terwee CB, Mokkink LB, Knol DL, *et al*. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012;21:651–7.
20 Terwee CB, Bot SD, de Boer MR, *et al*. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.
21 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.
22 Augustsson J, Thomee R, Karlsson J. Ability of a new hop test to determine functional deficits after anterior cruciate ligament reconstruction. *Knee Surg Sports Traumatol Arthrosc* 2004;12:350–6.
23 Barber SD, Noyes FR, Mangine RE, *et al*. Quantitative assessment of functional limitations in normal and anterior cruciate ligament-deficient knees. *Clin Orthop Relat Res* 1990;255:204–14.
24 Battaglia MJ II, Cordasco FA, Hannafin JA, *et al*. Results of revision anterior cruciate ligament surgery. *Am J Sports Med* 2007;35:2057–66.
25 Brosky JA Jr, 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.
26 Carter ND, Jenkinson TR, Wilson D, *et al*. Joint position sense and rehabilitation in the anterior cruciate ligament deficient knee. *Br J Sports Med* 1997;31:209–12.
27 Crossley KM, Thancanamootoo K, Metcalf BR, *et al*. Clinical features of patellar tendinopathy and their implications for rehabilitation. *J Orthop Res* 2007;25:1164–75.
28 Eastlack ME, Axe MJ, Snyder-Mackler L. Laxity, instability, and functional outcome after ACL injury: copers versus noncopers. *Med Sci Sports Exerc* 1999;31:210–15.
29 Gauffin H, Pettersson G, Tegner Y, *et al*. Function testing in patients with old rupture of the anterior cruciate ligament. *Int J Sports Med* 1990; 11:73–7.
30 Holm I, Fosdahl MA, Friis A, *et al*. Effect of neuromuscular training on proprioception, balance, muscle strength, and lower limb function in female team handball players. *Clin J Sport Med* 2004;14:88–94.
31 Jerre R, Ejerhed L, Wallmon A, *et al*. Functional outcome of anterior cruciate ligament reconstruction in recreational and competitive athletes. *Scand J Med Sci Sports* 2001;11:342–6.
32 Koutras G, Pappas E, Terzidis IP. Crossover training effects of three different rehabilitation programs after arthroscopic meniscectomy. *Int J Sports Med* 2009;30:144–9.
33 Myer GD, Schmitt LC, Brent JL, *et al*. Utilization of modified NFL combine testing to identify functional deficits in athletes following ACL reconstruction. *J Orthop Sports Phys Ther* 2011;41:377–87.
34 Ostenberg A, Roos E, Ekdahl C, *et al*. Isokinetic knee extensor strength and functional performance in healthy female soccer players. *Scand J Med Sci Sports* 1998;8:257–64.
35 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.
36 Ross MD. The relationship between functional levels and fear-avoidance beliefs following anterior cruciate ligament reconstruction. *J Orthop Traumatol* 2010;11:237–43.
37 Svensson M, Sernert N, Ejerhed L, *et al*. A prospective comparison of bone-patellar tendon-bone and hamstring grafts for anterior cruciate ligament reconstruction in female patients. *Knee Surg Sports Traumatol Arthrosc* 2006;14:278–86.
38 Tegner Y, Lysholm J, Lysholm M, *et al*. A performance test to monitor rehabilitation and evaluate anterior cruciate ligament injuries. *Am J Sports Med* 1986;14:156–9.
39 Vandermeulen D, Birmingham T, Forwell L. The test-retest reliability of a novel functional test; the lateral hop for distance. *Physiotherapy Canada* 2001;52:50–5.

40 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.
41 Zouita Ben Moussa A, Zouita S, Dziri C, *et al*. Single-leg assessment of postural stability and knee functional outcome two years after anterior cruciate ligament reconstruction. *Ann Phys Rehabil Med* 2009;52:475–84.
42 Bjorklund K, Skold C, Andersson L, *et al*. Reliability of a criterion-based test of athletes with knee injuries; where the physiotherapist and the patient independently and simultaneously assess the patient's performance. *Knee Surg Sports Traumatol Arthrosc* 2006;14:165–75.
43 Bjorklund K, Andersson L, Dalen N. Validity and responsiveness of the test of athletes with knee injuries: the new criterion based functional performance test instrument. *Knee Surg Sports Traumatol Arthrosc* 2009;17:435–45.
44 Witvrouw E, Lysens R, Bellemans J, *et al*. Which factors predict outcome in the treatment program of anterior knee pain? *Scand J Med Sci Sports* 2002;12:40–6.
45 Purdam CR, Cook JL, Hopper DM, *et al*. Discriminative ability of functional loading tests for adolescent jumper's knee. *Phys Ther Sport* 2003;4:3–9.
46 Irrgang JJ, Anderson AF, Boland AL, *et al*. Responsiveness of the International Knee Documentation Committee Subjective Knee Form. *Am J Sports Med* 2006;34:1567–73.
47 Irrgang JJ, Snyder-Mackler L, Wainner RS, *et al*. Development of a patient-reported measure of function of the knee. *J Bone Joint Surg Am* 1998;80:1132–45.
48 Prince SA, Adamo KB, Hamel ME, *et al*. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act* 2008;5:56.
49 Kennedy D, Stratford PW, Pagura SM, *et al*. Comparison of gender and group differences in self-report and physical performance measures in total hip and knee arthroplasty candidates. *J Arthroplasty* 2002;17:70–7.
50 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.