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 Baxter5

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 Consensus-based Standards for the selection of health Measurement Instruments (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.

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 debate9 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. We chose level 5 because the predominance of judge movement quality or range of motion, such as the

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 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). 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). 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’. We used ‘unknown’ to indicate that 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.

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
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 study42 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.
Hypothesis testing/construct validity

For the one leg hop for distance using a single hop, the methodological quality of the 16 studies was generally fair and for the version that requires three consecutive hops (triple hop), the methodological quality was poor in two of three studies. Likewise, the 6 m timed hop and crossover hop for distance generally were studied in articles of poor methodological quality. In one study 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, the quadriceps re

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 and one of good quality.

Responsiveness

Five studies reported on the responsiveness of five PPTs at the knee; however, only one study demonstrated good methodological quality. The two PPTs studied in this article 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 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 hop and one study each for the single leg vertical hop and the hop with three leaps (triple hop) and the crossover hop for distance.

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 and generally negative when describing convergent validity. 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 found a positive quality rating and two negative quality ratings. Likewise, the 6 m timed hop showed both a positive and a negative quality rating with regard to predicting function.

Responsiveness

The hop test, single leg vertical jump, crossover hop and triple jump have a positive quality rating and appeared to change with rehabilitation after knee injury. However, according to one study, 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 evidence, 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, or the ACL knee and the uninvolved knee in age-matched normals, or the ACL-deficient (ACLD) knee and the uninvolved knee in age-matched normals. Although the gender mix was not specified in one study, the other two studies 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 or between competitive and non-competitive athletes with ACLR.

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, 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...
only a small amount of the variance in self-rated functional outcomes. 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 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°, 90°, or 180°/s, or hamstring torque at 60° or 180°/s. No studies are available that investigated whether triple hop results predict injury.

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 nor does it predict function at 1 year as captured by the International Knee Documentation Committee (IKDC) form, self-rated global function, or the Knee Outcome Survey-Activities of Daily Living (KOS-ADL) Scale. 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. 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, who will cope with an ACL tear, or what sort of

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Synthesis of evidence by test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement property</td>
<td>Unknown (???)</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>One leg hop for distance: 1 hop</td>
<td>Reliability</td>
</tr>
<tr>
<td>Agreement</td>
<td>???</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>???</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>???</td>
</tr>
<tr>
<td>Responsiveness</td>
<td></td>
</tr>
<tr>
<td>One leg hop for distance: 3 hops</td>
<td>Reliability</td>
</tr>
<tr>
<td>Agreement</td>
<td>???</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>???</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>???</td>
</tr>
<tr>
<td>Responsiveness</td>
<td></td>
</tr>
<tr>
<td>6 m timed hop</td>
<td>Reliability</td>
</tr>
<tr>
<td>Agreement</td>
<td>???</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>???</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>???</td>
</tr>
<tr>
<td>Responsiveness</td>
<td></td>
</tr>
<tr>
<td>Crossover hop for distance</td>
<td>Reliability</td>
</tr>
<tr>
<td>Agreement</td>
<td>???</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>???</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>???</td>
</tr>
<tr>
<td>Responsiveness</td>
<td></td>
</tr>
<tr>
<td>Triple jump</td>
<td>Reliability</td>
</tr>
<tr>
<td>Agreement</td>
<td>???</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>???</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>???</td>
</tr>
<tr>
<td>Responsiveness</td>
<td></td>
</tr>
<tr>
<td>Single leg vertical jump</td>
<td>Reliability</td>
</tr>
<tr>
<td>Agreement</td>
<td>???</td>
</tr>
<tr>
<td>Hypothesis testing</td>
<td>???</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>???</td>
</tr>
<tr>
<td>Responsiveness</td>
<td></td>
</tr>
</tbody>
</table>
functional outcome will be attained,\textsuperscript{12, 11} nor is it sensitive enough to detect asymmetry in patients who are ACLD.\textsuperscript{11} However, the 6 m timed hop performed at 6 months after surgery does predict self-rated functional outcome at 1 year.\textsuperscript{12} In one study of fair methodological quality,\textsuperscript{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.\textsuperscript{42} However, there is limited but positive evidence with regard to construct validity and responsiveness. Bjorklund et al.\textsuperscript{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\textsuperscript{12–14} nor is it sensitive enough to detect abnormal limb symmetry in an ACLD population.\textsuperscript{11} However, test results make up one variable that helps predict who will cope with an ACL deficiency,\textsuperscript{28} and when the test is performed at 6 months after ACLR, it correlates with self-reported function at 1 year.\textsuperscript{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.\textsuperscript{14} 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\textsuperscript{21} 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\textsuperscript{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 uninjured 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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.
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.


