2013

Tests Examining Skill Outcomes in Sport: A Systematic Review of Measurement Properties and Feasibility

Samuel J. Robertson

Angus F. Burnett
Edith Cowan University, a.burnett@ecu.edu.au

Jodie Wilkie
Edith Cowan University, j.wilkie@ecu.edu.au

10.1007/s40279-013-0131-0
This article was originally published as: Robertson, S. J., Burnett, A. F., & Wilkie, J. (2013). Tests Examining Skill Outcomes in Sport: A Systematic Review of Measurement Properties and Feasibility. Sports Medicine, 44(4), 501-518. The final publication is available at Springer here

Tests examining skill outcomes in sport: A systematic review of measurement properties and feasibility

Samuel J Robertson[a], Angus F Burnett[b,c] and Jodie Cochrane[c]

(a) Centre for Exercise and Sports Sciences, School of Exercise and Nutrition Sciences, Deakin University, Melbourne, Victoria, Australia

(b) Department of Sports Science and Physical Education, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

(c) School of Exercise and Health Sciences, Edith Cowan University, Perth, Western Australia, Australia

Word Count: 6047

Corresponding author contact details:

Samuel J Robertson

Deakin University, Burwood Campus, 221 Burwood Hwy, Burwood, Victoria, 3125 AUSTRALIA

Email address: sam.robertson@deakin.edu.au

Phone: +61 3 9251 7277

Fax: +61 3 5227 2211

Figure Captions

Figure 1. PRISMA flow diagram

Formal running head

Tests examining skill outcomes in sport

Abstract

Background: A high level of participant skill is influential in determining the outcome of many sports. Thus, tests assessing skill outcomes in sport are commonly used by coaches and researchers to estimate an athlete's ability level, evaluate the effectiveness of interventions or for the purpose of talent identification.

Objective: The objective of this systematic review was to examine the methodological quality, measurement properties and feasibility characteristics of sporting skill outcome tests reported in the peer-reviewed literature.

Data Sources: A search of both SPORTDiscus and MEDLINE databases was undertaken.

Study Selection: Studies that examined tests of sporting skill outcomes were reviewed. Only studies that investigated measurement properties of the test (reliability or validity) were included. A total of 22 studies met the inclusion/exclusion criteria.

Study Appraisal and Synthesis Methods: A customised checklist of assessment criteria, based on previous research, was utilised for the purpose of this review.

Results: A range of sports were the subject of the 22 studies included in this review, with considerations relating to methodological quality being generally well-addressed by authors. A range of methods and statistical procedures were used by researchers to determine the measurement properties of their skill outcome tests. The majority (95%) of the reviewed studies investigated test-retest reliability, and where relevant, inter and intra-rater reliability was also determined. Content validity was examined in 68% of the studies, with most tests investigating multiple skill domains relevant to the sport. Only 18% of studies assessed all three reviewed forms of validity (content, construct and criterion) with just 14% investigating the predictive validity of the test. Test responsiveness was reported in only 9% of studies, whilst feasibility received varying levels of attention.

Limitations: In organised sport, further tests may exist which have not been investigated in this review. This could be due to such tests firstly, not being published in the peer-review literature and secondly, not having their measurement properties (i.e. reliability or validity) examined formally.

Conclusions: Of the 22 studies included in this review, items relating to test methodological quality were on the whole, well addressed. Test-retest reliability was determined in all but one of the reviewed studies, whilst most studies investigated at least two aspects of validity (i.e. content, construct or criterion-related validity). Few studies examined predictive validity or responsiveness. While feasibility was addressed in over half of the studies, practicality and test limitations were rarely addressed. Consideration of study quality, measurement properties and feasibility components assessed in this review can assist future researchers when developing or modifying tests of sporting skill outcomes.

1. Introduction

Although a clear relationship between skill and success exists in sport, there is currently a paucity of literature reviewing the characteristics of existing tests examining skill,[1] with the majority of the literature to date focusing on physical determinants of performance. Although tests of specific skill outcomes date back over fifty years,[2-7] out-dated methodology and undefined measurement properties (ie. reliability, validity and responsiveness) often limit their usefulness. Tests of skill outcomes experience widespread utility in research, in particular for the purpose of assessing the effect of coaching or scientific interventions on performance.[8-10] Recent studies have also utilised these tests to investigate the effects of nutrition,[11-14] game-specific fatigue, [15] performer focus of attention[16] and pre-skill execution routine[17] on participant performance. Further, a body of work exists in team-based field sports such as football in assessing participant skill (amongst other factors) within simulated match-play environments.[18-20]

The prevalence of skill outcome tests being used in the field is also widespread. For example, the use of data or scores obtained from appropriately designed assessments can potentially eliminate the need to collect longitudinal information on an athlete, for the purposes of rating or ranking them either individually or against their peers. Further, these tests can also been used to assist in identifying relative strengths and weaknesses of the performer,[21-23] monitor progress of an athlete within a structured training program,[22-24] provide information on predictive performance potential,[8,23] inform improved practice and training complexity/specificity[25] as well as provide a time-efficient method of defining participant ability levels.[26]

Recently, skill outcome tests have experienced considerable use for the purposes of identifying talent in sport.[8,21,22,27-29] For example, team-based competitions such as the

Australian Football League in Australia and the National Football League in the United States employ multidisciplinary testing “combines” in their player drafting processes that involve each participant receiving a score based on an outcome of a specific test. Although these events have traditionally focused on physiological assessments, in an attempt to account for additional attributes associated with producing a high level of performance in these sports, tests examining skill outcomes such as kicking, passing and throwing accuracy have also been assessed in recent times.

However, the use of skill outcome tests used either in isolation, or as part of a multi-disciplinary assessment protocol, has also been the topic of considerable discussion in recent times.[28,30-35] This debate appears to centre predominantly on a) the representative design of currently utilised testing methods and b) the ideal level of specificity and detail included in such assessments. In particular, the latter consideration has focused on whether designed tests should assess participants on a series of technical based actions or indicators, as opposed to scoring the relevant skill outcome alone (although a combination of both has been used). The decision made by test designers to utilise either approach may have contrasting advantages with relation to reliability, validity, feasibility as well as the intended purpose for undertaking the test. For example, it is evident that the processes that contribute to skilled outcomes in sporting scenarios exhibit considerable inter- and intra-individual variability,[36-39] potentially rendering assessments of such components inherently unreliable.[40] This can also be a consideration in the test design of skill outcomes, with recent work showing differences in the reliability of soccer passing versus shooting in testing scenarios.[24,29] Additionally, tests assessing outcomes of skill in isolation can also face issues in displaying adequate validity, at least in part due to the context in which they are undertaken; often not able to consider the situational, task-strategizing and decision-making components of undertaking the particular action.[41-42] Irrespective of this discussion, tests examining skill outcomes experience considerable use for a range of purposes, however, there appears to be no formal system in place with which to evaluate their measurement properties.

Regardless of whether a test has been developed for research or practical purposes, it is well established that it should display acceptable measurement properties; this has in particular been well addressed in medical and health-related fields.[43-48] However, despite widespread use, studies investigating such tests in sport may not consistently report these properties. Although tests of physiological performance have been the subject of review in recent times,[1] to our knowledge, three specific studies examining sport performance assessments specifically have been published. Of these reviews, two exclusively addressed football (soccer)[10,27] whilst also discussing in some depth, the physiological and technical contributors to performance.[1,20,27]

Therefore, in considering the suitability of sporting skill outcome tests, a number of rating items should be considered. Firstly, detailed descriptions of methodological quality and study characteristics are important so that results can be considered with relevance to the population being examined. For example, the properties displayed by a skill test when undertaken by elite participants should not be assumed as similar when being utilised with participants of lesser ability level or for example, the opposite sex. Additionally, the provision of this information allows for accurate reproduction and comparison of studies by future researchers or coaches implementing the test in the field. Such descriptions should therefore be inclusive of a number of components including specific details on the participants themselves,[44,49-50] inclusion and exclusion criteria,[44,50] consideration of sample size [44,49,51], reporting of floor and ceiling effects [44], stability of test conditions and participants between retest periods [44,49] and the test-retest interval duration.[44,51]

As multiple trials often form part of a testing protocol's scoring system[8,52] and may actually be necessary in order to gain a better representation of a participant's actual ability,[22,53] studies should also be examined for evidence of reporting reliability. Further, three main types of validity are typically stated as being important characteristics to the investigation of the quality assessment of a test, and therefore also warrant reporting. These are content validity, construct validity and criterion-related validity.[12,54-57]

Feasibility is another test property commonly examined in the health/medical literature.[46,49,58-59] In the context of this review, it can be defined as the ease in which a test can be undertaken, administered and scored or rated.[49-58] Feasibility is of particular importance to sport, where tests need to be practical for the environment they are intended to be used within, or will be likely to experience limited use by athletes, coaches and researchers. It could be reasoned that skill outcome tests have been particularly popular in their use as they are relatively easy to score and can often be undertaken without the use of high-end equipment.

The primary aim of this systematic review was to examine the methodological quality of sporting skill outcome tests reported in the peer-review literature as well as report the types and level of measurement properties investigated in these tests. A secondary aim of this review was to examine factors related to the feasibility and limitations of the identified tests.

2. Methods

Although a number of methods for reporting items in systematic reviews exist in the literature,[43-44, 60-62] due to their lack of specificity for use in systematically assessing measurement properties of variables/tests and feasibility-related issues, a customised framework based on previous literature was required to be developed for use in this review. A similar approach has been undertaken in previous systematic reviews examining test measurement properties in other disciplines,[50,63-65] although wherever possible the COSMIN framework[44] was in particular deferred to where possible. Additional considerations relating to the design of this framework (as well as the rating items contained within) were informed by a number of additional sources including; studies assessing similar domains,[1,24] validated systematic review guidelines and checklists[54,61-62,66-67] as well as other reviews which have utilised a customised model.[50,64] This process is described in greater detail in Section 2.4.

2.1 Search Strategy

The literature search for this review was undertaken between July 2012 and March 2013 by the first author (SR) using the SPORTDiscus and MEDLINE databases. Key words utilised in the search using multiple combinations of AND/OR phrases included ‘skill’, ‘measurement’, ‘test’, ‘assessment’, reliability, ‘validity’, ‘testing’, ‘elite’, ‘sport’, ‘instrument’, ‘sporting’, ‘practical’, ‘outcome’, ‘reproducibility’, ‘task’ and ‘feasibility’. Further studies were collected following examination of citations present within the collected publications (‘snowballing’).

2.2 Inclusion Criteria

Initial pilot testing of the search strategy in February 2012 revealed multiple studies relating to the design of skill tests as far back as 1958. However, no studies prior to 1990 were found to have met the inclusion criteria described below; therefore in facilitating the search process, articles were required to be published after 1990 and were included up to and including March 2013. Additional inclusion criteria for studies examining skill outcome tests in this review were; a) each publication addressing a skills test collated from the abovementioned search strategy must have been peer-reviewed and written in English;; b) abstracts of each article were required to be present in the database search; c) articles describing the use of a multidisciplinary testing battery could be included provided the skill outcome testing component could be extracted and reviewed separately to other assessment items.

2.3 Exclusion Criteria

The following criteria resulted in exclusion of studies for this review; a) articles not reporting at least one component of either reliability or validity of the developed test; b) articles that reported physiological function or specific motor skills not directly relevant to the sport investigated or assessing a skill outcome; c) articles utilising tests that had their measurement properties investigated previously elsewhere; d) articles that stated utilising minor adaptations of tests investigated previously; and e) any articles that had been withdrawn from publication. Further, f) studies examining tests rating or scoring participants on technical processes in isolation of recordable skilled outcomes were excluded. For example, tests that rated combinations of technical criteria in order to produce a score were excluded as they were not assessing the skill outcome per se. Studies that examined both processes in addition to a skill product or outcome had the latter components extracted for review wherever possible.

2.4 Data Extraction

As the validity of using customised scored review templates for systematically reviewing measurement properties and feasibility of skill outcome tests is yet to be defined,[52] quantitative ratings for each of the reviewed items were not provided. The assessment items used in this review were based on study quality, test measurement properties (reliability, validity and responsiveness) and feasibility. Wherever possible, data pertaining to the measurement properties of each instrument were recorded.

A total of seven items were used to rate study quality and the operational definitions have been reported in Table I. These items were; the level of detail provided on study participants, whether participant inclusion/exclusion criteria were reported, the size of the participant sample, whether floor and ceiling effects were reported, whether familiarisation was undertaken with the participants prior to testing, whether the stability of both participants and testing conditions was accounted for, and lastly the reporting of the length of the test re-test interval. Although a variety of methods can be used to determine appropriate sample size,[68-70] absolute sample size values were used to allow direct comparison across studies.[44]

Information relating to test-retest reliability and inter/intra-rater reliability were also retrieved, with the type and level of reliability both assessed (operational definitions provided in Table II). Additionally, due to the large variety of statistical analyses in studies, reliability statistics for only the six most commonly reported approaches were reported. These were; coefficient of variation (CV%), intraclass correlation coefficient (ICC), correlation coefficients (r), 95% limits of agreement (inclusive of ratio limits of agreement) (LoA & RLoA respectively), typical error of measurement (TEM) and generalisability theory. Although specific ratings were provided for studies that reported ICC and r-values, no published guidelines were found relating to as what constituted an acceptable level of reliability for the remaining four statistical approaches. Consequently, ratings of numerical results were not provided in studies that reported reliability using solely these methods.

Operational definitions relating to validity are reported in Table II. Although some evidence exists supporting the use of both the kappa statistic and the content validity index (the proportion of a small group of experts that agree on a certain item being included in the assessment of a domain) to determine content validity,[55,71-72] these have not been widely reported in the sport literature. A more common method has been the use of ‘expert’ panels or coaching groups to develop test items. Whilst there are limitations to this process,[73] it nonetheless experiences substantial use in the relevant literature. Therefore, for the purposes of this review, content validity was rated according to whether a study gained concession by an expert panel for the items assessed in the test. Construct validity was considered as inclusive of both discriminative and convergent validity,[54-55,74-75] whilst criterion-related validity included a consideration of both the concurrent and predictive properties of the test.[54-55,74] In assessing these types of validity, some research has defined correlation coefficients in excess of 0.65[48] or 0.70[76] as appropriate, however support also exists for values of between 0.30 and 0.50 as being acceptable.[49,74,76-77] Although such correlation data was reported in some of the reviewed studies, due to the variety of statistical approaches utilised, studies were assessed on whether these measurement properties were investigated by the authors, as opposed to reporting results. However, the statistical approach used was reported wherever possible.

Operational definitions for responsiveness and feasibility characteristics are also reported in Table II. Test responsiveness can be assessed by calculating the ratio of the clinically relevant change to the standard deviation of the intra-participant test-retest differences,[78-79] or by referring to the test’s effect size.[58,74] Other common methods include obtaining the minimum clinically important difference (MCID) [80] or comparing of median test scores from multiple rounds of testing.[81] In this review, studies were rated on whether data relating to undertaking of any of these approaches were reported, with the length of the interval observed between these two (or more) rounds of testing also obtained. As studies should also focus on interpretability; they were also rated on whether they provided information relating to the minimum important change or difference. Finally, components relating to test feasibility and limitations were also recorded. As such, information relating to practicality, test duration, intended context, the presence of a familiarisation session/s and consideration of test limitations were all also extracted for the purposes of rating.[46,58] No appropriate published quantitative values of feasibility item types for the kind of tests investigated in this review were found, therefore studies were rated on whether each of these areas were included in the studies.

A customised Microsoft Excel™ spreadsheet was developed to record the abovementioned extracted data from each of the studies reviewed. All data from each study was extracted by two authors independently. Prior to undertaking this assessment it was stipulated that any instance where the two reviewers provided conflicting scores for any of the criteria, the paper would be re-assessed. However, this did not occur at any stage throughout the review process.

****** INSERT TABLES I & II ABOUT HERE ******

3. Results

A total of 604 articles were found as a result of the initial search strategy and snowballing processes. An outline of the search results and reasons for exclusion has been provided in Figure 1. It should be noted that 34 studies were excluded from the review as they examined tests of motor skills not directly relating to a performance outcome. Further, 10 studies were also excluded as they detailed only minor revisions of existing, original versions of tests already included in the review. As a result of applying the inclusion and exclusion criteria, a total of 22 studies remained for inclusion in the review. Of these 22 studies, five described skill outcome tests designed for use in football, three each for volleyball and golf, two for hockey, with one each for tennis, rugby league, squash, water polo, netball, rock climbing, racquetball, wheelchair basketball and quad rugby. Table III provides a description of the characteristics of the reviewed studies.

****** INSERT FIGURE 1 ABOUT HERE ******

****** INSERT TABLE III ABOUT HERE ******

3.1 Study Methodological Quality

Table IV displays the results of the study quality assessment undertaken of the skills tests. Of the studies reviewed, 59% were shown to have adequately stated participant characteristics, with 36% receiving a partial score. Only 14% of the reviewed studies stated both inclusion and exclusion criteria adequately with a further 18% of the total studies providing inclusion criteria only. A range of participant sample sizes were noted across the studies (n = 11 to 313) with 18% utilising a sample size of n > 50 and just 14% recruiting an n > 100. Floor and ceiling effects of participant scores were only reported in a small number (14%) of cases. A total of 64% studies also implemented familiarisation sessions as part of their as part of their tests. In 68% of studies the stability of both the participants and test conditions were adequately reported, with a further 14% receiving a partial rating. Test retest intervals ranged from 10 mins to 28 days, with 77% of studies reporting this detail. Same-day retesting was undertaken in 18% of these studies, whereas 68% implemented retesting sessions that were undertaken within one week of the initial assessment.

3.2 Reliability

Table V displays results relating to the rating of the measurement properties and feasibility characteristics of the reviewed skills tests. Of the six statistical approaches used to assess level of reliability, 64% of studies reported ICC's, 27% used CV's, 32% utilised Pearson or Spearman product moment correlations, 18% reported 95% LoA (or RLoA), with 14% and 5% reporting TEM% and generalisability theory respectively. In just under half (41%) the studies reviewed, a good to excellent level of test-retest reliability was reported, whereas in the majority of the remaining studies (55%), a partial rating for reliability was given. Inter-rater reliability was investigated in the three studies that involved testers undertaking assessments of participants and then provided scores on their observations.[22,95-96] Inter-rater reliability was assessed using similar techniques as for test re-test reliability, with all studies in this case reporting a form of correlation coefficient (i.e. ICC or an r-value). Intra-rater reliability was examined in only 9% of studies due it most likely not being considered relevant for investigation in the majority of cases.[22,97]

3.3 Validity & Responsiveness

Content validity was assessed in 68% of the studies reviewed and was determined (at least in part) through consultation with a panel of experts or coaches in 27% of cases.[8,22,26,97-98] Only one study generated and reduced test items through mail-based Delphi rounds. [95] The remaining studies (36% of the total number reviewed) used a combination of review of literature and an assessment of actual game/competition demands.

Construct validity was determined in 64% of these studies with most utilising the existing status of the participant (professional competing, high-level amateur or amateur) as the construct for categorisations of ability. Of these studies, 71% used between-group comparisons of test scores (i.e. via t-tests or ANOVA) to determine whether differences existed between ability-level, whereas the remaining 29% used minimum clinically important differences (MCID) values or correlational or factor analysis. A total of 36% of studies also reviewed investigated criterion-related validity in their skill outcome tests. All of these determined the level of association with a concurrent measure including comparisons with expert/coach rankings provided prior to testing,[24,82,95,97,99] or comparisons of observed scores with expected participant rankings (based on external scales).[13,29,84]

Only 14% of studies examined a test's ability to predict future performance, with all of these studies utilising correlational analysis to determine the relationship of participant score with rankings and/or performance in subsequent tournaments or competitions.[26,82,97] Further, only 9% investigated the responsiveness of their testing protocol. These studies reported MCID's [22,25] and utilised data taken from a post-testing session undertaken four weeks later[48] to assess this measurement property. Additionally, 32% of studies reported the minimum important change or difference as part of their investigation.

3.4 Feasibility and Limitations

Test feasibility considerations and test limitations were addressed in 50% of the studies reviewed. A further 36% received a partial score, with the reduction in rating predominantly due to the lack of information provided regarding the limitations of the test. Of the 22 studies, 55% also reported the intended context or use for their designed skill test, or it was implied due to the purpose of the study. Of the studies providing this information, 42% stated the related protocols may be of use for the purposes of evaluating the success of interventions,[8,9,21,24,26,100-101] with 17% specifically developing their instrument to examine the effects of nutritional or ergogenic aid supplementation.[12,84] Further, 17% stated a use for their protocol in talent identification[21-22,102] with other reasons including a time efficient manner of defining and monitoring participant development,[22,99] method of benchmarking participants[93,98] and a process in which to inform an increase in practice schedule design or complexity.[25,102] Time to complete the tests was reported in 41% of studies with values ranging from 20 to 90 minutes, although it is worth noting that the longest test was part of a multidisciplinary testing battery assessing other non-skill domains.

****** INSERT TABLE IV & V ABOUT HERE ******

4. Discussion

The overarching objective of this study was to a) identify sporting skill outcome tests reported in the peer-reviewed literature and b) systematically review these studies based on their methodological quality and measurement properties reported. Considerations relating to test feasibility were also examined. Findings from the search strategy revealed there were a relatively small number of studies assessing all measurement properties (i.e. reliability, validity and responsiveness) with just over half adequately investigating some aspect of feasibility.

Despite the reporting of participant characteristics being important for the purposes of test reproducibility, they were not fully described in the majority of cases. In particular, information relating to the specific ability-level of participants as well as their anthropometric characteristics was lacking. The external reproducibility of many of the reviewed studies was also potentially compromised due to a lack of clear inclusion and exclusion criteria. Authors should be encouraged to show greater transparency by reporting these criteria in future work. Participant sample considerations in this review related to the size of the cohort(s) investigated. However, as a number of studies recruited professional or elite level participants as part of their investigation, access to a larger population of these cohorts is likely to be more difficult than in other disciplines.[24] In ensuring sample size is adequate, authors should ideally recruit participants from a range of ability levels, which in turn can also allow for a greater investigation of construct validity. Whilst not a rating item in this particular review, it should also be noted that the need for implementation of familiarisation sessions was addressed in the majority of studies where relevant. As results stemming from these preliminary sessions typically noted a retest improvement for in particular lower-level participants,[24,29,84] these authors should be commended for including such an undertaking as part of the investigation of their tests. The attention provided by many authors to ensuring both testing and participant conditions remained stable between retesting sessions should also be noted.

Whilst a range of test-retest interval durations were reported in the studies reviewed, it is difficult to provide an objective rating on what he exact duration of this test characteristic should be, as it is dependent on the nature of the test itself (i.e. the number and complexity of skilled actions being performed). Regardless, it is important for test-retest intervals to not be too short in duration as a) this may not allow for adequate examination of the assessments' temporal stability,[54] and b) often performers may still be fatigued from previous trials [68,87] (although this is likely to be more of a concern in physiologically exertive assessments). Conversely, excessively long retest intervals can result in large variation of results (thereby affecting reliability); this may be due to seemingly innocuous factors, (i.e. participant circadian variations)[103] or notable skill improvements in participants between the two trials.

An inclusion criterion for this review was that either reliability or validity of each skill test was reported in the reviewed study. Test-retest reliability was the most commonly addressed measurement property reported across the tests reviewed with all but one of the reviewed studies investigating this property. Of those that did investigate test-retest reliability, just under half displayed good to excellent repeatability. In the rare circumstances where inter-rater reliability was assessed, good to excellent levels of agreement were found. For ease of reader interpretability this review reported only the six most commonly used methods in assessing reliability and as such is not a comprehensive representation of the statistical methods available on which to assess this measurement property. Existing systematic review frameworks have recommended rating studies on whether a particular statistical technique is utilised,[44] however a discussion on this area is beyond the scope of this review and the reader is directed elsewhere for a comprehensive discourse on the pros and cons of available techniques used to assess reliability in this context.[68,87]

It is also worth noting that any investigation of test reliability should include some consideration of the amount of error present in any measurement tools used to assist in the scoring of the assessment. For example, a number of technologies such as radar measurement devices,[81,21] radar speed guns,[29,84,101], and video cameras[22,24-25,95,101] were all utilised to obtain data that was directly used in either the scoring or administering of the reviewed tests. In some circumstances information relating to digitisation techniques and analysis errors were reported; in these cases the authors should be commended for providing such detailed descriptions.[22,95,29,101] Future authors are recommended to do likewise when developing future tests where such technologies are integral to the scoring of the protocol.

Due to a lack of widely reported techniques in assessing content validity for sporting skill tests, it was not surprising that for the majority of studies reviewed, no statistical techniques were used to assess this form of validity. It is recommended that wherever possible researchers use a formal process and/or quantitative measure to assess this form of validity, such as the Delphi rounds seen in previous studies[95] or those commonly used in other disciplines (i.e. a content validity index).[54,72] The argument for this more transparent approach is supported by the consideration that although in some cases determining the content of a particular testing protocol may seem a relatively simple task, many (in fact, most) sports require multiple skills to be executed. This may mean that one individual test does not assess the entire content of skill and multiple tests may be needed to define a construct more completely.[8,21,24,27] Therefore, sports involving complex and multiple skill domains can pose a particularly difficult problem for researchers. This may be due to multiple or different skills being required within competition (i.e. passing, shooting, catching). Further, and specifically in team sports, both the type of skill requirement and their relative importance may differ between players depending on their role or position within the team. Further still, certain participants may display a high-level of aptitude in one domain yet be relatively mediocre or poor in another.

When considering these factors, it is not surprising that there has been some recent debate regarding the appropriateness of assessing different components of skill in isolation of each other, particularly in the football codes.[30-31] Whilst the approach of concurrently assessing multiple components has precedent in the two of the five football-specific studies reviewed here,[12,52] a decision on which skills to include in a test design is likely to depend on the intended use of the protocol. For example, some sports may be better disposed to isolated extraction and testing of items better than others (such as golf, which requires clearly differentiated skills performed in relatively 'closed' environments). As shown in Table III, skill outcomes/domains such as 'accuracy', 'placement', 'passing', 'shooting' and 'time to complete' tasks were commonly assessed within the studies included in this review. Some authors also implemented minimum skill execution speed [29,84] or temporal[12] constraints to the design of the protocol with others including the use of dual-task methodology to more accurately assess participant skill.[25] An obvious benefit of the addition of these types of environmental constraints to test protocols can be the improvement of the external validity and/or representative design of the test. With particular reference to skill tests, this term is perhaps best described as “how the (test) design…may allow for the maintenance of coupled perception and action processes that reflect the functional behaviour of athletes in specific performance contexts.”[35]

Despite the undoubted importance of these methodological considerations, ensuring there is a balance between improving the representative design of a test and maintaining or improving its measurement properties (in particular, protecting against a loss of reliability) can be a quandary for researchers when designing protocols. The development of a test displaying good measurement properties should ideally allow for more specific, concurrent evaluation of the technical processes and actions contributing to the skill outcome. Such an approach can also then allow better investigation of the ‘how’ and ‘why’ of the performance achieved (if relevant to the specific study). However, the initial goal of the researcher should be to develop appropriate measurement properties as a priority. For example, evolution and amendments of tests over time occurs in other disciplines, and it is evident that tests have undergone considerable change from initial versions through processes such as increasing time efficiency and/or representativeness.[104] Future research and discussion may seek to include better representative task design however, a lack of a clear definition in this context makes this difficult at present.

With reference to construct validity, although discriminative test characteristics were typically investigated by studies in this review, limited evidence of the investigation of convergent validity was noted. This is can be a particularly perplexing form of validity for investigators in sports performance to assess, as often one of the defining motivations for development of a new test may be because of a gap in the literature and therefore, there may be no similar test to compare the new method to.[54] This may at least in part explain why there were only a small number of cases noted in this review. However, as the number of skill tests continues to increase, such investigations may become both more useful and relevant to researchers. For example, examination of convergent validity may inform the development of a more comprehensive testing assessment than in existing versions and/or help to reduce the length of such protocols (i.e. thereby also increasing test feasibility).[54,75] Particularly, if a test requires expensive equipment or is of a particularly long duration, it is unlikely to experience continued use by those working in the field. Whilst the ability of a test to relate to a concurrent measure of the same construct is important for its criterion-related validity, a test displaying a proven ability to predict actual performance (predictive validity) could be considered an even more important characteristic of a test. However, as shown in this review, very few studies of sporting skill outcomes have examined this property.

Similarly, the evaluation of a test's responsiveness was rarely investigated in the studies included in this review. This is despite the fact that responsiveness is routinely investigated in other fields of research such as epidemiology,[78] or when examining quality of life [79] or rehabilitation outcomes.[81,105-106] Similarly to test-retest reliability, investigation of a test's responsiveness requires access to the same group of participants for repeat assessments and therefore, can be difficult when examining samples such as elite athletes who may have competition and/or training schedules that conflict with the ideals of test designers. In particular, when using these populations, investigators need to consider the ethical implications of excessive testing whilst ensuring the benefits from the testing outweigh any potential athlete burden. Ongoing, mutually beneficial collaborations with sporting bodies can potentially present researchers with suitable opportunities to investigate this particular measurement property of their tests.

Whilst the need for a test to display acceptable measurement properties is clearly important, its usefulness as a tool for researchers and coaches is reduced if it not feasible or practical. Whilst less than half of the studies in this review stated the potential use of their tests as well as their limitations, a number of practical considerations went largely undiscussed. For example, other considerations such as the availability and cost of equipment,[59,85] the ease of incorporating the test with participants of different ability-levels,[59] level of participant enjoyment, number of participants to be tested,[59] and the availability of skilled examiners[85] were not routinely reported. Some investigations into test feasibility in other fields have utilised standardised expert or coach interviewing to rate some of the test components post-testing. This included the perceived value of the assessment (by the rater, participant and coach), ease of scoring,[59,85] time taken to explain and set up the test,[54,59,105] as well as the availability of equipment provided.[58-59,105] Therefore, it is evident that feasibility requires further consideration in studies of the nature reviewed here.

Whilst the duration of a test may be dependent on both the sport and the skill itself, it is logical to suggest that implementation of the test should be shorter than the actual competition itself. Tests of excessive duration may have the potential to induce fatigue [68] and/or cause the performer (or their coach/coaches) to lose interest or motivation in undertaking the assessment. This may be of particular concern when undertaking tests with younger participants, where increased pressure may also cause poor and unrepresentative performance of participants.

Duration of a test will however also be highly dependent upon the number of trials undertaken, which in turn, is influenced by the number of trials required to gain a true representation of a participant's ability. In many sports, a single trial may suffice and may actually be representative of the task being assessed however there may be a need for multiple trials in some skill tests. This may particularly be the case in sports of a continuous nature. This consideration, most likely combined with an intention to produce adequate reliability (termed the Spearman-Brown prophecy) was noted in almost all of the tests reviewed. However, although quite likely to be well justified in these cases, in most studies the number of repeated trials utilised appeared to be decided arbitrarily. Test designers should look to base the optimal number of trials on objective evidence. For example, in other disciplines particular testing items may have their weightings adjusted according to their importance to the testing construct.[104,107] Further, item reduction techniques such as Rasch analysis and item concept-retention can also be used to reduce the number of items within an instruments while also maintaining high levels of test-retest reliability.[104,107]

4.1 Limitations

A limitation of this review was the inability to undertake any form of meta-analysis. This was due to the considerable variety of statistical procedures used to determine test measurement properties. Additionally, it should be noted that findings from this review may not be generalizable due to the relatively small number of sports examined in the studies contained therein. As different sports will always contain different skill components and expressions of performance, the sports investigated here provide only an overview of the sports contained within. Further, it is likely that tests currently exist in use within practical environments that have not been reviewed here due to not being reported in the literature.

5. Conclusions

This review assessed the methodological quality, measurement properties and feasibility of 22 studies reporting tests of sporting skill. Methodological quality of the studies was mixed, with minimal attention provided on inclusion and exclusion criteria and optimising sample size. Implementation of familiarisation sessions and a consideration of participant and testing condition stability were present in the majority of studies. A range of methods and statistical procedures have been used by researchers to determine the measurement properties of their skill outcome tests, thereby making direct comparison of studies difficult. Test-retest reliability was determined in all but one of the reviewed studies, whilst most investigated at least two aspects of validity (i.e. content, construct or criterion-related validity). However, a distinct lack of specific investigation into both the predictive validity and responsiveness of skill outcome tests was noted. While some aspect of feasibility was addressed in just under half of the studies, considerations relating to test practicality were not formally investigated in any of the studies. As the items for this review were extracted from a number of existing models reported in other disciplines, future work may look to develop a specific framework for use in the sports sciences. Until then, a consideration of the study quality characteristics, measurement properties and feasibility items outlined in this review can assist future researchers in the development and or modification skill tests in sport.

Acknowledgements

The authors report no conflict of interest with the information presented in this review, although two of the studies included in this review were authored/co-authored by the first and second author. Further, no funding was received by any of the authors for preparing and writing this review. All three authors met the requirements for authorship in this journal and provided significant contributions to this paper.

References

1. Currell K, Jeukendrup, AE. Validity, reliability and sensitivity of measures of sporting performance. Sports Med 2008; 38 (4): 297-316.

2. Cale AA. The investigation and analysis of softball skill tests for college women [dissertation]. Maryland: University of Maryland, 1962.

3. Cobb JW. The determination of the merits of selected items for the construction of a baseball skill test for boys of Little League age [dissertation]. Indiana: Indiana University, 1958.

4. Collins DR, Hodges PB. A comprehensive guide to sports skills tests and measurement. Maryland, United States: Rowman & Littlefield Education, 2001.

5. Fein JT. Construction of Skill Tests for Beginning Collegiate Women Fencers [dissertation]. Iowa: University of Iowa, 1964.

6. Safrit MJ. Construction of skill tests for beginning fencers [dissertation]. Madison: University of Wisconsin, 1962.

7. Sopa AM. The construction of an indoor batting skills test for junior high school girls [dissertation]. Madison: University of Wisconsin, 1967.

8. Robertson SR, Burnett AF, Newton RU. Development and validation of the approach-iron skill test for use in golf. Eur J Sport Sci. Epub 2013; Jan 10

9. Sunderland C, Cooke K, Milne H, et al. The reliability and validity of a field hockey skill test. Int J Sports Med 2006; 27 (5): 395-400.

10. Russell M, Kingsley M. Influence of exercise on skill proficiency in soccer. Sports Med 2011; 41 (7): 523-39.

11. Duncan MJ, Taylor S, Lyons M. The effect of caffeine ingestion on field hockey skill performance following physical fatigue. Res Sports Med 2012; 20 (1): 25-36.

12. Currell K, Conway S, Jeukendrup AE. Carbohydrate ingestion improves performance of a new reliable test of soccer performance. Int J Sport Nutr Exerc Metab 2009; 19 (1): 34-46.

13. Bottoms LM, Hunter AM, Galloway SD. Effects of carbohydrate ingestion on skill maintenance in squash players. Eur J Sport Sci 2006; 6 (3): 187-95.

14. Russell M, Benton D, Kingsley M. Influence of carbohydrate supplementation on skill performance during a soccer match simulation. J Sci Med Sport 2012; 15 (4): 348-54.

15. Russell M, Benton D, Kingsley M. The effects of fatigue on soccer skills performed during a soccer match simulation. Int J Sports Physiol Perform 2011; 6 (2): 221-33.

16. McKay B, Wulf G. A distal external focus enhances novice dart throwing performance. Int J Sport Exerc Psych 2012; 10 (2): 149-56.

17. McCann P, Lavallee D, Lavallee, R. The effect of pre-shot routines on golf wedge shot performance. Eur J Sport Sci 2001; 1 (5): 1-10.

18. Russell M, Rees G, Benton D, et al. An exercise protocol that replicates soccer matchplay. Int J Sports Med 2011; 32 (7): 511-18.

19. Nicholas CW, Nuttall FE, Williams C. The Loughborough Intermittent Shuttle Test: a field test that simulates the activity pattern of soccer. J Sports Sci 2000; 18 (2): 97-104

20. Williams JD, Abt G, Kilding AE. Ball sport endurance and sprint test ($BEAST_{90}$): validity and reliability of a 90-minute soccer performance test. J Strength Cond Res 2010; 24 (12): 3209-18.

21. Robertson SJ, Burnett AF, Newton RU, et al. Development of the Nine-Ball Skills Test to discriminate elite and high-level amateur golfers. J Sports Sci 2012; 30 (5): 431-7.

22. Gabbett TJ, Georgieff B. The development of a standardized skill assessment for junior volleyball players. Int J Sports Physiol Perform 2006; 1 (2): 95-107.

23. Pyke F. Introduction. In: Gore C, editor. Physiological tests for elite athletes. Champaign, Illinois: Human Kinetics, 2000: xii-xiv.

24. Russell M, Benton D, Kingsley M. Reliability and construct validity of soccer skills tests that measure passing, shooting, and dribbling. J Sports Sci 2010; 28 (13): 1399-408.

25. Gabbett T, Wake M, Abernethy B. Use of dual task methodology for skill assessment and development: examples from rugby league. J Sports Sci 2011; 29 (1): 7-18.

26. Brent S, Draper N, Hodgson C, et al. Development of a performance assessment tool for rock climbers. Eur J Sport Sci 2009; 9 (3): 159-67.

27. Ali A. Measuring soccer skill performance: a review. Scand J Med Sci Sports 2011; 21 (2): 170-83.

28. Lidor R, CôTé JE, Hackfort D. ISSP position stand: to test or not to test? The use of physical skill tests in talent detection and in early phases of sport development. Int J Sport Exerc Psychol 2009; 7 (2): 131-46.

29. Ali A, Williams C, Hulse M, et al. Reliability and validity of two tests of soccer skill. J Sports Sci 2007; 25 (13): 1461-70.

30. Kingsley M, Russell M, Benton D. Authors' response to letter to the editor: “The need for ‘representative task design’ in evaluating efficacy of skills tests in sport: a comment on Russell, Benton and Kingsley (2010)” J Sports Sci 2012; 30 (16): 1731-3.

31. Vilar L, Araújo D, Davids K, et al. The need for 'representative task design' in evaluating efficacy of skills tests in sport: a comment on Russell, Benton and Kingsley (2010) J Sports Sci 2012; 30 (16): 1727-30.

32. Philips E, Davids K, Renshaw I, Portus M. Expert performance in sport and the dynamics of talent development. Sports Med 2010; 40 (4): 271-83.

33. Vaeyens R, Lenoir M, Williams MA, et al. Talent identification and development programmes in sport: current models and future directions. Sports Med 2008; 38 (9): 703-14.

34. McNamara A, Collins D. Comment on "Talent identification and promotion programmes of Olympic athletes". J Sports Sci 2011; 29 (12): 1353-6.

35. Pinder RA, Renshaw I, Davids K. The role of representative design in talent development: a comment on "Talent identification and promotion programmes of Olympic athletes." J Sports Sci 2013; 31 (8): 803-6.

36. Glazier P. Movement variability in the golf swing: theoretical, methodological and practical issues. Res Q Exerc Sport 2011; 82 (2): 157-61.

37. Fleisig G, Chu Y, Weber A, et al. Variability in baseball pitching biomechanics among various levels of competition. Sports Biomech 2009; 8 (1): 10-21.

38. Langdown BL, Bridge M, Li F. Movement variability in the golf swing. Sports Biomech 2012; 11 (2): 273-87.

39. Betzler NF, Monk SA, Wallace ES, et al. Variability in clubhead presentation characteristics and ball impact location for golfers' drives. J Sports Sci 2012; 30 (5): 439-48.

40. Bartlett R. Movement variability and its implication for sports scientists and practitioners: an overview. Int J Sports Sci Coach 2008; 3 (1): 113-24.

41. Ovens A, Smith W. Skill: making sense of a complex concept. J Phys Ed New Zeal 2006; 39 (1): 72-84.

42. Smith W. Skill acquisition in physical education: a speculative perspective. Quest 2011; 63 (3): 265-74.

43. Moher D, Liberati A, Tetzlaff J, et al. Reprint-preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Phys Ther 2009; 89 (9): 873-80.

44. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010; 19 (4): 539-49

45. Auld ML, Boyd RN, Moseley GL, et al. Tactile assessment in children with cerebral palsy: a clinimetric review. Phys Occup Ther Pediatr 2011; 31 (4): 413-39.

46. Fliess-Douer O, Vanlandewijck YC, Manor GL, et al. A systematic review of wheelchair skills tests for manual wheelchair users with a spinal cord injury: towards a standardized outcome measure. Clin Rehab 2010; 24 (10): 867-86.

47. Helmerhorst HJ, Brage S, Warren J, et al. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. Int J Behav Nutr Phys Act 2012; 9 (1): 103-57.

48. Williams MA, McCarthy CJ, Chorti A, et al. A systematic review of reliability and validity studies of methods for measuring active and passive cervical range of motion. J Manipulative Physiol Ther 2010; 33 (2): 138-55.

49. Packham T, MacDermid JC, Henry J, et al. A systematic review of psychometric evaluations of outcome assessments for complex regional pain syndrome. Disabil Rehabil 2012; 34 (13): 1059-69.

50. Dobson F, Morris ME, Baker R, et al. Gait classification in children with cerebral palsy: a systematic review. Gait Posture 2007; 25 (1): 140-52.

51. Howe TE, Dawson LJ, Syme G, et al. Evaluation of outcome measures for use in clinical practice for adults with musculoskeletal conditions of the knee: a systematic review. Man Ther 2012; 17 (4): 100-18.

52. Mirkov D, Nedeljkovic A, Kukolj M, et al. Evaluation of the reliability of soccer-specific field tests. Journal Strength Cond Res 2008; 22 (4): 1046-50.

53. Knudson DV, Morrison CS. Qualitative analysis of human movement. 2nd ed. Illinois: Human Kinetics: 2002.

54. Streiner DL, Norman GR. Health measurement scales: a practical guide to their development and use. 3rd ed. Oxford, United Kingdom: Oxford University Press, 2005.

55. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 3rd ed. New Jersey: Pearson/Prentice Hall: 2008.

56. Westen D, Rosenthal R. Quantifying construct validity: two simple measures. J Pers Soc Psychol 2003; 84 (3): 608-18.

57. Squires A. A valid step in the process: a commentary on Beckstead (2009). Int J Nurs Stud 2009; 46 (9): 1284-5.

58. Braeken AP, Kempen GI, Eekers D, et al. The usefulness and feasibility of a screening instrument to identify psychosocial problems in patients receiving curative radiotherapy: a process evaluation. BMC Cancer 2011; 11 (1): 479-90.

59. Gabel CP, Melloh, M, Burkett B, et al. Lower limb functional index: development and clinimetric propertie. Phys Ther 2012; 92 (1): 98-110.

60. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007; 7 (1): 10

61. Gardner MJ, Machin D, Campbell MJ, et al. Statistical checklists. In: Altman DG, Machin D, Bryant TN et al., editors. Statistics with confidence: confidence intervals and statistical guidelines. 2nd ed. Bristol: BMJ Books, 2000: 191-201.

62. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002; 11 (3): 193-205.

63. Hill BE, Williams G, Bialocerkowski AE. Clinimetric evaluation of questionnaires used to assess activity after traumatic brachial plexus injury in adults: a systematic review. Arch Phys Med Rehabil 2011; 92 (12): 2082-9.

64. Solway S, Brooks D, Lacasse Y, et al. A qualitative systematic overview of the measurement properties of functional walk tests used in the cardiorespiratory domain. Chest 2001; 119 (1): 256-70.

65. Lohr KN, Aaronson NK, Alonso J, et al. Evaluating quality-of-life and health status instruments: development of scientific review criteria. Clin Ther 1996; 18 (5): 979-92.

66. Campion MA. Article review checklist: a criterion checklist for reviewing research articles in applied psychology. Pers Psychol 2006; 46 (3): 705-18.

67. Valderas JM, Ferrer M, Mendivil J, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health 2008; 11 (4): 700-8.

68. Hopkins WG. Measures of reliability in sports medicine and science. Sports Med 2000; 30 (1): 1-15.

69. Beck TW. The importance of a priori sample size estimation in strength and conditioning research. J Strength Cond Res 2013; 27 (8): 2323-37.

70. Nevill AM, Holder RL, Cooper SM. Statistics, truth and error reduction in sport and exercise sciences. Eur J Sport Sci 2007; 7 (1): 9-14.

71. Cohen J, A. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20 (1): 37-46.

72. Haynes SN, Richard DC, Kubany, ES. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol Assess 1995; 7 (3): 238-47.

73. Beckstead JW. Content validity is naught. Int J Nurs Stud 2009; 46 (9): 1274-83.

74. Stinson JN, Kavanagh T, Yamada J, et al. Systematic review of the psychometric properties, interpretability of self-report pain measures for use in clinical trials in children and adolescents. Pain 2006; 125 (1 & 2): 143-57.

75. Ravens-Sieberer U, Auquier P, Erhart M, et al. The KIDSCREEN-27 quality of life measure for children and adolescents: psychometric results from a cross-cultural survey in 13 European countries. Qual Life Res 2007; 16 (8): 1347-56.

76. Munro BH. Statistical methods for health care research. 5th ed. Philadelphia (USA): Lippincott, Williams & Wilkins: 1999.

77. Van Saane N, Sluiter JK, Verbeek JH, et al. Reliability and validity of instruments assessing job satisfaction – a systematic review. Occup Med 2003; 53 (3): 191-200

78. Beckerman H, Roebroeck ME, Lankhorst GJ, et al. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001; 10 (7): 571-8.

79. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40 (2): 171-8

80. Hopkins WG. Probabilities of clinical or practical significance. Sportscience [serial online]. 2002:6. Available from URL: http://sportsci.org/jour/0201/wghprob.htm.[accessed 2012 Dec 11]

81. Crossley KM, Bennell KL, Cowan SM, et al. Analysis of outcome measures for persons with patellofemoral pain: which are reliable and valid? Arch Phys Med Rehabil 2004; 85 (5): 815-22.

82. Downs SB, Wood TW. Validating a Special Olympics volleyball skills assessment test. Adapt Phys Act Q 1996; 13 (2): 166-79.

83. Bartlett J, Smith L, Davis K, et al. Development of a valid volleyball skills test battery. J Phys Educ Rec Dance 1991; 62 (2): 19-21.

84. Ali A, Foskett A, Gant N. Validation of a soccer skill test for use with females. Int J Sports Med 2008; 29 (11): 917-21.

85. Stevens B, Gibbins S. Clinical utility and clinical significance in the assessment and management of pain in vulnerable infants. Clin Perinatol 2002; 29 (3): 459-68.

86. Baumgartner TA, Jackson AS. Measurement for evaluation in physical education and exercise science. 5th ed. Michigan, United States: Brown & Benchmark, 1998.

87. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998; 26 (4): 217-38.

88. Law MC, MacDermid J. Evidence-based rehabilitation: A guide to practice, Slack Incorporated, 2007 Outcome measures rating form. CanChild Centre for Childhood Disability Research 2nd ed. Thorofare (NJ): 2004: 367.

89. de Boer MR, Moll AC, de Vet HC, et al. Psychometric properties of vision-related quality of life questionnaires: a systematic review. Ophthalmic Physiol Opt 2004; 24 (4): 257-73.

90. Landis J, Koch G. The measurement of observer agreement of categorical data. Biometrics 1977; 33 (1): 159-74.

91. Pietrobon R, Coeytaux RR, Carey TS, et al. Standard scales for measurement of functional outcome for cervical pain or dysfunction: a systematic review. Spine 2002; 27 (5): 515–22.

92. Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important difference: a review of concepts and methods. Spine 2007; 7 (5): 541-6.

93. Bartlett J, Smith L, Davis K, et al. Development of a valid volleyball skills test battery. J Phys Educ Rec Dance 1991; 62 (2): 19-21.

94. Streiner DL. A checklist for evaluating the usefulness of rating scales. Can J Psychiatry 1993; 38 (2): 140-8.

95. Yilla AB, Sherrill C. Validating the Beck battery of quad rugby skill tests. Adapt Phys Act Q 1996; 15 (2): 155-67.

96. Porter JM, Landin D, Hebert EP, et al. The effects of three levels of contextual interference on performance outcomes and movement patterns in golf skills. Int J Sports Sci Coach 2007; 2 (3): 243-55.

97. Lam ET, Zhang JJ. The development and validation of a racquetball skills test battery for young adult beginners. Meas Phys Educ Exerc Sci 2002; 6 (2): 95-126.

98. Bock-Jonathon BB, Venter RE, Bressan ES. A comparison between skill and decision-making ability of netball players at club level: pilot study. S Afr J Res Sport Phys Educ Rec 2007; 29 (1): 29-38.

99. De Groot S, Balvers IJ, Kouwenhoven SM, et al. Validity and reliability of tests determining performance-related components of wheelchair basketball. J Sports Sci 2012; 30 (9): 879-87.

100. Royal KA, Farrow D, Mujika I, et al. The effects of fatigue on decision making and shooting skill performance in water polo players. J Sports Sci 2006; 24 (8): 807-15.

101. Vergauwen L, Spaepen AJ, Lefevre J, et al. Evaluation of stroke performance in tennis. Med Sci Sports Exerc 1998; 30 (8): 1281-8.

102. Lemmink KA, Elferink-Gemser MT, Visscher C. Evaluation of the reliability of two field hockey specific sprint and dribble tests in young field hockey players. Br J Sports Med 2004; 38 (2): 138-42.

103. Atkinson G. Reilly T. Circadian variation in sports performance. Sports Med 1996; 21 (4): 292-312

104. Beaton DE, Wright JG, Katz JN. Development of the QuickDASH: comparison of three item-reduction approaches. J Bone Joint Surg Am 2005; 87 (5): 1038-46.

105. Kilkens OJ, Post MW, Dallmeijer AJ, et al. Wheelchair skill tests: a systematic review. Clin Rehab 2003; 17 (4): 418-30.

106. Faber MJ, Bosscher RJ, van Wieringen PC. Clinimetric properties of the performance-orientated mobility assessment. Phys Ther 2006; 86 (7): 944-54.

107. Robinson JM, Cook JL, Purdam C, et al. The VISA-A questionnaire: a valid and reliable index of the clinical severity of Achilles tendinopathy. Br J Sports Med 2001; 35 (5): 335-41.

Table I. Details of review items relating to study methodological quality.

Assessment item	Operational definition	Assessment criteria
Sample size	Was the sample size included in the analysis adequate? [44]	n ≥ 100: ++++ n = 50-99: +++ n= 30-49: ++ n < 30: +
Details of study participants	Sex, age, participant numbers, ability-level, and (where relevant) anthropometrical data provided. [1,48,50-63]	Yes – all participant details reported Partial – one or two levels of detail not present NR
Inclusion/exclusion criteria	Detail relating to the inclusion and exclusion criteria as utilised in study methodology. [1,50,54,63,76]	Yes – both exclusion/inclusion criteria reported Partial – exclusion or inclusion criteria reported NR
Familiarisation session	The undertaking of a test familiarisation session with all participants prior to main testing. [31-84-85]	Yes – information relating to familiarisation session reported NR
Test retest interval	Duration relating to the interval between repeated bouts of testing. [44,51]	Yes – time of retest interval reported NR NA
Floor & ceiling effects	Number and/or percentage of participants who had the lowest and highest possible total score. [44]	Yes – both upper and lower values or percentages reported Partial – either upper or lower values or percentages reported NR
Stability of participants and test conditions	Were the participants and testing conditions (i.e. equipment and environment) stable between testing sessions? [44,49]	Yes – specific stability of conditions reported Partial – stability implied by study design NR NA

'+': less than 30 participants recruited for the study; '++': between 30 and 49 participants recruited for the study; '+++': between 50 and 99 participants recruited for the study; '++++': more than 100 participants recruited for the study; **NA** = not applicable to the particular investigation; **NR** = not reported

Table II. Details of review items relating to measurement properties and feasibility.

Assessment item	Operational definition	Assessment criteria
Reliability/ Measurement error		
Test-retest reliability	The consistency of performer/s scoring over repeated rounds of testing.[74] ICC or correlation coefficient values ≥ 0.8 rated as good to excellent,[54-55,77,86-90] ≥ 0.4 to < 0.8 rated as poor to average.[47,54] CV%, Generalisability theory, TEM% and 95% LoA (& RLoA) also reported.	Yes – provided and shows 'good' to 'excellent' reliability Partial – provided but a) relative reliability not investigated or b) 'poor' to 'average' reliability shown NR
Intra/inter-rater reliability	Inter-rater: the level of agreement between scoring/assessing when undertaken by two or more raters.[86] Intra-rater: defined as the agreement among two or more trials administered or scored by the same rater.[86]	Yes – either or both investigated Partial – reported but a) no reliability coefficient provided or b) 'poor' to 'average' reliability shown (as per test-retest definition) NR NA
Validity		
Content	How well a specific test measure what it intends to measure.[1,51,54-55,74] Do the items included in the test cover the entirety of those relevant to assessing a particular skill outcome measure? [44,63,89]	Yes – face, logical and/or content validity results reported NR
Construct	The ability of the testing instrument to measure a theoretical construct of performance.[55-56] How well do scores achieved on a particular test relate to a) other methods of assessment or b) ranking of the same theoretical construct?[24,55-56] Discriminative: the ability of the test to discriminate between performers of different ability (as rated by another measure)[24,54,76] Convergent: the ability of the test to relate with alternate measures of either the same construct or other associated variables.[54,76]	Yes – discriminative and/or convergent validity results reported NR
Criterion-related	The ability of a test to show good agreement with an external measure or gold standard protocol.[49,54-55,90-91] Concurrent: relationship of test score to participant score/rankings in an alternate form of measurement.[49,54] Predictive: relationship of test score with future results in a relevant	Yes –predictive or concurrent validity results reported NR

	sporting competition or performance.[49,54]	
Responsiveness (sensitivity)	The ability of a test to detect worthwhile and ‘real’ skill improvements in its intended population,[59,77-78] between initial bout of testing and subsequent rounds[48,59,68]	Yes – results relating to test responsiveness reported and test-retest interval stated. NR
Minimum important change or difference provided	Information relating to the minimum important change or minimum important difference provided in Results or Discussion section.[44,92]	Yes – minimum important change provided NR
Feasibility & limitations		
Practicality & limitations	The ease in which a test can be undertaken, administered and scored.[46,49,58,84-85] Limitations relating to findings and interpretability of the test acknowledged and stated in study.[58]	Yes – feasibility/practicality and limitations discussed Partial – one of feasibility/practicality and limitations discussed NR
Test context	Information relating to the anticipated use and context of the test provided?[46]	Yes – information relating to test context reported NR
Test duration	Expected or actual duration of the testing protocol reported.[93-94]	Yes – duration of test/trial reported NR

CV = Coefficient of variation; **ICC** = intraclass correlation coefficient; **LOA** = limits of agreement; **NA** = not applicable to the particular investigation; **NR** = not reported; **RLOA** = ratio limits of agreement; **TEM** = typical error of measurement

Table III. Study characteristics of the 22 articles included in the review.

Sport	Author(s)	Test name	Domain(s) tested	Outcome measure	Participant characteristics
Football	Ali et al. (2007) [29]	Loughborough soccer passing test	Passing (multiple trials)	Time (s)	Elite male (n=24) Non elite male (n=24)
Football		Loughborough soccer shooting test	Shooting left foot Shooting right foot (multiple trials)	Score (pts) Time (s) Ball velocity	Elite male (n=24) Non elite male (n=24)
Football	Mirkov et al. (2008)[52]	Unnamed	Standing kick Zig-zag with ball	Distance (m) Time (s)	Professional senior male (n=20)
Football	Ali et al. (2009) [84]	Loughborough soccer passing test	Passing (multiple trials)	Time (s)	Elite female (n=19) Non-elite female (n=16)
Football	Currell et al. (2009) [12]	Unnamed	Dribbling Kicking accuracy Heading	Time (s) Score (pts)	Recreational male (n=11)
Football	Russell et al. (2010)[24]	Unnamed	Passing Shooting Dribbling	Precision (cm) Success (%) Ball speed (m/s)	Professional male (n=10) Recreational male (n=10)
Golf	Porter et al. (2007)[96]	Unnamed	Putting Pitching	Score (pts)	Adult male undergraduate (n=23)
Golf	Robertson et al. (2012)[21]	Nine-ball skills test	Iron club straight shot Iron club fade shot Iron club draw shot	Score (pts)	Elite male (n=14) High-level amateur male (n=16)
Golf	Robertson et al. (2013)[8]	Approach-iron skill test	Iron club accuracy	Score (pts)	Elite male (n=26) High-level amateur male (n=23)
Hockey	Lemmink et al. (2004)[102]	Shuttle sprint & dribble test	Dribble time Peak dribble & sprint	Time (s)	Young male (n=22) Young female (n=12)
		Slalom sprint & dribble test	Dribble time		
Hockey	Sunderland et al. (2006)[9]	Field hockey skill test	Dribbling Passing Shooting	Time (s)	University male (n=20) University female (n=19)

Netball	Bock-Jonathon et al. (2007)[98}	Unnamed	Passing accuracy Repeated passing Pivot and pass	Score (n) Time (s)	University female players (n=30)
Quad rugby	Yilla & Sherrill (1998)[95]	Beck battery of rugby skills tests	Manoeuvrability Pass for accuracy Picking Catching Pass for distance	Score (pts) Time (s) Count (#)	Male (n=65)
Racquetball	Lam & Zhang (2002)[97]	Racquetball skills test battery	Service placement Power drive Power shot placement Ceiling shot Wall rally	Score (pts)	College students mixed (n=131)
Rock-climbing	Brent et al. (2009)[26]	Rock-over climbing test	Height reached	Level attained	Elite, advanced, intermediate and novice climbers (n=46)
Rugby league	Gabbett et al. (2011)[25]	Draw and pass Proficiency task	Draw and pass	Score (pts)	High-skilled male (n=20) Lesser-skilled male (n=17)
Squash	Bottoms et al. (2006)[13]	Boast & drive skill test	Forehand drive Backhand drive	Score (pts)	National male players (n=16)
Tennis	Vergauwen et al. (1998)[101]	Leuven tennis performance test	First service Second service Neutral situations Defensive situations Volleys	Errors (%) Ball velocity (km/hr) Distance to sideline (cm)	Professional male (n=7) Semi-professional male (n=10) Amateur male (n=10)
Volleyball	Bartlett et al. (1991)[93]	NCSU volleyball skills test battery	Serve Forearm pass Set	Score (pts)	College students male/female (n=313)
Volleyball (special Olympics)	Downs & Wood (1996)[82]	Volleyball skills assessment test	Serve Forearm pass Setting skill Spiking	Score (pts)	State-based male (n = 101) State-based female (n = 29)
Volleyball	Gabbett & Georgieff (2006) [22]	Unnamed	Spiking Setting Serving Passing	Score (pts)	National, state and novice mixed (n=30)

Water polo	Royal et al. (2006)[100]	Unnamed	Shooting accuracy	Score (%)	Junior elite male (n=14)
Wheelchair basketball	De Groot et al. (2012)[99]	Unnamed	Pass for accuracy Free throw accuracy Maximal pass Lay ups Pick up the ball Spot shot	Time (s) Score (pts)	Mixed ability male (n=19)

Tests examining skill outcomes in sport

Table IV. Study methodological quality items of the reviewed skill tests.

Sport	Author(s)	Details of study participants	Inclusion/ exclusion criteria	Sample size	Floor & ceiling effects	Familiarisation session	Stability of participants & test conditions	Test-retest interval
Football	Ali et al. (2007) [29]	Yes	Partial	++	NR	Yes	Yes	1 day
		Yes	Partial	++	NR	Yes	Yes	1 day
Football	Mirkov et al. (2008)[52]	Yes	NR	+	NR	Yes	NR	NR
Football	Ali et al. (2009) [84]	Yes	Partial	+	NR	Yes	Yes	7 days
Football	Currell et al. (2009) [12]	Yes	NR	+	NR	Yes	Yes	7 days
Football	Russell et al. (2010)[24]	Yes	Yes	+	NR	Yes	Yes	2 days
Golf	Porter et al. (2007)[96]	Partial	NR	+	NR	Yes	Partial	7 days
Golf	Robertson et al. (2012)[21]	Partial	Yes	+	Yes	NR	Yes	10 mins
Golf	Robertson et al. (2013)[8]	Partial	Yes	++	Yes	NR	Yes	10 mins
Hockey	Lemmink et al. (2004)[102]	Partial	NR	++	NR	NR	Yes	14-28 days
Hockey	Sunderland et al. (2006)[9]	No	NR	++	NR	Yes	Yes	3-14 days
Netball	Bock-Jonathon et al. (2007)[98}	Partial	NR	++	NR	NR	NR	NR
Quad rugby	Yilla & Sherrill (1998)[95]	Yes	Partial	+++	Yes	NR	Partial	NR
Racquetball	Lam & Zhang (2002)[97]	Partial	NR	++++	NR	Yes	Partial	2-7 days
Rock climbing	Brent et al. (2009)[26]	Yes	NR	++	NR	Yes	Yes	7-14 days
Rugby league	Gabbett et al. (2011)[25]	Yes	Partial	++	NR	Yes	Yes	NR
Squash	Bottoms et al. (2006)[13]	Yes	NR	+	NR	Yes	Yes	NR
Tennis	Vergauwen et al. (1998)[101]	Yes	NR	+	NR	NR	Yes	7 days
Volleyball	Bartlett et al. (1991)[93]	Partial	NR	++++	NR	NR	NR	2 days
Volleyball (special Olympics)	Downs & Wood (1996)[82]	Yes	NR	++++	NR	Yes	Yes	4 days
Volleyball	Gabbett & Georgieff (2006) [22]	Yes	NR	++	NR	Yes	Partial	2 days
Water polo	Royal et al. (2006)[100]	Yes	NR	+	NR	Yes	Yes	5 mins
Wheelchair basketball	De Groot et al. (2012)[99]	Yes	NR	+	NR	NR	Yes	<7 days

'+': less than 30 participants recruited for the study; '++': between 30 and 49 participants recruited for the study; '+++': between 50 and 99 participants recruited for the study; '++++': more than 100 participants recruited for the study; **NA** = Not applicable to this particular investigation; **NR** = Not reported

Table V. Measurement properties and feasibility of the reviewed skill tests.

Sport	Author(s)	Reliability (r, ICC, CV, TEM & 95% LoA)	Validity type(s) (statistical approaches or results in brackets)	Responsiveness (time interval in brackets)	Minimum important change or difference	Feasibility, practicality & limitations	Test context	Test duration
Football	Ali et al. (2007) [29]	Partial (test-retest) (r = 0.43 to 0.64; ICC = 0.42 to 0.64; CV% = 11.2 to 16.0; LoA	Construct (Student's t-test) Criterion (median-split analysis)	NR	Yes	NR	NR	~20 mins
		Partial (test-retest) (r = 0.24 to 0.32, ICC = 0.23 to 0.31, CV% = 49.4 to 65.3); LoA	Construct (Student's t-test) Criterion (median-split analysis)	NR	Yes	NR	NR	~20 mins
Football	Mirkov et al. (2008)[52]	Partial (test-retest) (ICC = 0.76 to 0.81, TEM% = 0.21 to 2.81, CV% = 3.3 to 9.2)	Content	NR	Yes	Yes	Yes	NR
Football	Ali et al. (2009) [84]	Partial (test-retest) (r = 0.55 to 0.73, CV% = 16.7 to 17.1)	Construct (Student's t –test) Criterion (median-split analysis)	NR	NR	Partial	Yes	~20 mins
Football	Currell et al. (2009) [12]	Yes (test-retest) (CV% = 0.7 to 6.8)	Content	NR	NR	Partial	Yes	~90 mins
Football	Russell et al. (2010)[24]	Partial (test-retest) (r = 0.38 to 0.78, ICC = 0.37 to 0.77, CV% = 2.2 to 23.5; LoA & RLOA)	Content Construct (independent sample t-test) Criterion-related (mean-split analysis)	NR	Yes	Yes	Yes	47 mins
Golf	Porter et al. (2007)[96]	Partial (test-retest) (ICC = 0.72 to 0.76) (Inter-rater) (ICC = 0.98)	Construct (t-test)	NR	NR	Partial	NR	NR
Golf	Robertson et al. (2012)[21]	Partial (test-retest) (ICC = 0.67, CV% = 27.5)	Content Construct (ANOVA)	NR	NR	Yes	Yes	20-30 mins
Golf	Robertson et al. (2013)[8]	Partial[a] (test-retest) (95% LoA = 0.2 to 2.1 pts)	Content Construct (ANOVA)	NR	NR	Yes	Yes	50-65 mins
Hockey	Lemmink et	Partial (test-retest)	NR	NR	NR	Partial	Yes	NR

	al. (2004)[102]	(ICC = 0.71 to 0.91)						
Hockey	Sunderland et al. (2006)[9]	Yes (test-retest) (r = 0.96, ICC = 0.96)	Construct (correlation) (r = 0.61 to 0.83)	NR	Yes	Yes	Yes	NR
Netball	Bock-Jonathon et al. (2007)[98}	NA	Content Construct (Mann-Whitney)	NR	NR	Yes	Yes	NA
Quad rugby	Yilla & Sherrill (1998)[95]	Yes (test-retest) (r = 0.94 to 0.99) (inter-rater) (r = 0.98)	Content Construct (factor analysis) Criterion-related (concurrent) (r = 0.53 to 0.98)	NR	NR	NR	NR	NR
Racquetball	Lam & Zhang (2002)[97]	Yes (test-retest) (generalisability theory) Yes (intra-rater) (ICC = 0.87)	Content Criterion (concurrent & predictive) (r = -0.48)	NR	NR	Yes	NR	20-25 mins
Rock climbing	Brent et al. (2009)[26]	Yes (test-retest) (ICC = 0.90)	Content Construct (ANOVA) Criterion (concurrent) (r = 0.61) (predictive)	NR	NR	Partial	Yes	NR
Rugby league	Gabbett et al. (2011)[25]	Yes (test-retest) (ICC = 0.86, TEM% = 5.3)	Content Construct (ANOVA)	Yes (4 weeks)	NR	Yes	Yes	NR
Squash	Bottoms et al. (2006)[13]	Partial (test-retest) (r = 0.68)	Criterion (concurrent) (r = -0.62)	NR	NR	NR	Yes	NR
Tennis	Vergauwen et al. (1998)[101]	Partial (test-retest) (ICC = 0.15 to 0.91)	Content Construct (ANOVA)	NR	NR	Partial	Yes	NR
Volleyball	Bartlett et al. (1991)[93]	Partial (test-retest) (ICC = 0.65 to 0.88)	Content	NR	NR	Partial	NR	<40 mins
Volleyball (special Olympics)	Downs & Wood (1996)[82]	Yes (test-retest) (ICC = 0.83 to 0.88)	Content Construct Criterion (concurrent) Predictive (r = 0.88 to 0.96)	NR	NR	Yes	NR	NR
Volleyball	Gabbett &	Yes (test-retest)	Content	Yes	Yes	Yes	Yes	NR

Tests examining skill outcomes in sport

	Georgieff (2006) [22]	(ICC = 0.85 to 0.94, TEM% = 0.2 to 0.9) Intra-rater (ICC = 0.85 to 0.98, TEM% = 5.1 to 6.9) Inter-rater (ICC = 0.90 to 0.94, TEM% = 7.0 to 10)	Construct (MCID)	(8 weeks)					
Water polo	Royal et al. (2006)[100]	Yes (test-retest) (ANOVA)	Content	NR	NR	Partial	Yes	NR	
Wheelchair basketball	De Groot et al. (2012)[99]	Partial (test-retest) (ICC = 0.41 to 0.99, 95% LoA = -0.3 – 0.2 to -14.9 – 11.2)	Construct (discriminative) (ANOVA) (convergent)	NR	Yes	Yes	Yes	75 mins	

[a] Received a partial rating, as no relative measure of reliability reported for comparison across studies.

ANOVA = analysis of variance; **CV** = coefficient of variation; **ICC** = Intraclass correlation coefficient; **LoA** = limits of agreement; **MCID** = minimum clinically important difference; **NA** = Not applicable to this particular investigation; **NR** = Not reported; **r** = correlation; **RLoA =** ratio limits of agreement; **SDD** = smallest detectable difference; **TEM** = typical error of measurement