Journal of Orthopaedics 70 (2025) 283–291

Contents lists available at ScienceDirect

Journal of Orthopaedics

journal homepage: www.elsevier.com/locate/jor

Reliability, minimal detectable change, and standard error of measurement of functional tests for athletes: A systematic review

Thiago Teixeira Serafim [a], Ana Paula Ramos [b], Diego Ailton Prudêncio [b,c], Filippo Migliorini [d,e,f,*], Nicola Maffulli [g,h,i], Rodrigo Okubo [b,c]

[a] Laboratory of Sport and Exercise Psychology (LAPE), Santa Catarina State University (UDESC), Florianópolis, Brazil
[b] Physical Therapy Graduate Program (PPGFT), Santa Catarina State University (UDESC), Florianopolis, Brazil
[c] Department of Physiotherapy, Santa Catarina State University (UDESC), Florianópolis, Brazil
[d] Department of Trauma and Reconstructive Surgery, University Hospital of Halle, Martin-Luther University Halle-Wittenberg, 06097, Halle (Saale), Germany
[e] Department of Life Sciences, Health, and Health Professions, Link Campus University of Rome, Via Del Casale di San Pio V, 00165, Rome, Italy
[f] Department of Orthopaedic and Trauma Surgery, Academic Hospital of Bolzano (SABES-ASDAA), 39100, Bolzano, Italy
[g] Department of Trauma and Orthopaedic Surgery, Faculty of Medicine and Psychology, University La Sapienza, 00185, Roma, Italy
[h] School of Pharmacy and Bioengineering, Keele University Faculty of Medicine, Stoke on Trent, ST4 7QB, UK
[i] Centre for Sports and Exercise Medicine, Barts and the London School of Medicine and Dentistry, Mile End Hospital, Queen Mary University of London, London, E1 4DG, UK

ARTICLE INFO

Keywords:
Assessment
Return to Sport
Return to Play
Evaluation
Athlete
Functional test
Performance test

ABSTRACT

Introduction: Functional tests must be validated for the target population. It is also important that the professionals applying them know which tests are the most reliable. Some tests have a standard error of measurement (SEM), which needs to be considered, as does the minimal detectable change (MDC) used to quantitatively perceive clinical improvement. It is important to know the psychometric properties of a functional test to consider it suitable for its use. This study aims to synthesise values of psychometric properties of functional tests in validation studies for athletic or physically active populations.

Methods: This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The search was performed in PubMed, Web of Science, SportDiscus and Cochrane in June 2025. The methodological quality of the included studies was evaluated by the Consensus-based Standards for Health Measurement Instruments (COSMIN) Risk of Bias checklist.

Results: The final review included 49 studies. The study samples ranged from 11 to 243, totalling 1713 subjects. The mean age of the subjects studied ranged from 16.47 ± 0.51 to 59.40 ± 8.70 years. The reliability values verified by ICC ranged from 0.26 to 0.99. SEM and MDC values were delivered in percentages and absolute values. All studies evaluated using the COSMIN checklist were classified as "Inadequate."

Conclusion: Functional tests used to assess athletes generally have good reliability values. However, standardisation in the application is necessary. The training of professionals who administer the tests is essential for greater reliability. Furthermore, greater stabilisation of the subject being evaluated is necessary for strength tests to reduce compensations during the test.

Level of evidence: I – Systematic review.

1. INTRODUCTION

In the context of health and the sporting environment, individual assessment is very important in diagnosing dysfunctions, providing direct treatment, supporting development, and serving as criteria for returning to sporting activity.[1,2] Within these issues, task simulation tests are used, thus acting in the most functional way possible.[3] These functional tests constitute the most complex part of an assessment

* Corresponding author. Department of Trauma and Reconstructive Surgery, University Hospital of Halle, Martin-Luther University Halle-Wittenberg, Ernst-Grube-Street 40, 06097, Halle (Saale), Germany.
E-mail addresses: thiagotserafim@outlook.com (T.T. Serafim), anaramos.fisio@gmail.com (A.P. Ramos), diegoprudencio1@hotmail.com (D.A. Prudêncio), filippo.migliorini@uk-halle.de (F. Migliorini), n.maffulli@qmul.ac.uk (N. Maffulli), rodrigo.okubo@udesc.br (R. Okubo).

https://doi.org/10.1016/j.jor.2025.08.030
Received 5 August 2025; Accepted 16 August 2025
Available online 18 August 2025

because they require several skills to complete a given task[4] and need to be used in the practical environment of physically active individuals and athletes to prevent further injuries, direct treatment, and return to sport safely.[5] The tests must be validated for the target population, minimising the chance of errors.[6] Furthermore, standardisation among the assessment team also needs to be achieved. For example, some authors perform warm-ups, but others do not. All these factors need to be standardised to increase reliability.[7] The quality of information tests provides depends, in part, on their psychometric properties.[8] Knowing the values of these functional test properties guides which test should be used in each situation and each individual or group of individuals.[9,10] Tests with high-reliability values should be preferred. Furthermore, some tests have a standard error of measurement (SEM), and this needs to be taken into account,[11] as well as minimal detectable change (MDC), which is used to quantify perceived clinical improvement.[12,13] Once we know the values of the psychometric properties of functional tests, we know which one should be used to contextualise the results of a given individual.[10] This study aims to synthesise the values of psychometric properties of functional tests in validation studies for athletic or physically active populations.

2. Methods

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).[14] The protocol for this review was registered in the International Prospective Register of Systematic Reviews (PROSPERO), CRD42020177143.

2.1. Search strategy

The search was performed in the PubMed, Web of Science, SportDiscus and Cochrane databases using the following keywords: Evaluation OR Measurement AND Psychometrics OR Reliability AND Sports OR Athlete AND Strength OR Mobility OR "Range of Motion" OR Balance OR Function NOT Child OR "Youth Sports" OR Questionnaire. The search was performed in June 2025. The process of selecting studies was conducted independently by two researchers (TTS and APR). If there was disagreement, a third reviewer (RO) with experience in systematic reviews was consulted for the final decision.

2.2. Eligibility criteria and selection of studies

Studies were deemed eligible according to the PICOS criteria (Table 1). We included in this systematic review only articles which validated the psychometric proprieties of tests using reliability, standard error of measurement (SEM) or minimal detectable change (MDC) – with outcomes of balance, strength, range of motion and agility in athletes or physically active adult individuals (>16 years old). Articles published in English were included without time limits. Articles were excluded if their population was compared to individuals younger than 16 years old, sedentary, and with neurological or cardiorespiratory disorders. Studies that evaluated the psychometric properties of questionnaires, cardiorespiratory and neurological tests, clinical tests, video analysis, kinematic analysis, smartphone use or sport-specific technical tests were also excluded.

Table 1
PI(E)COS framework.

Criteria	Inclusion	Exclusion
Population	Physically active adults or athletes	<16 years old or sedentary adults
Intervention (Exposition)	Functional test	Questionnaire, Specific sports modality test
Comparison	Other tests	–
Outcome	Reliability	–
Study	Validation	Systematic review, Randomized clinical trial, Editorial, Case study, Correlation

2.3. Quality of studies

The methodological quality of the included studies was evaluated using the Consensus-based Standards for Health Measurement Instruments (COSMIN) Risk of Bias checklist, which consists of a tool from the COSMIN to classify the methodological quality of studies on the properties of measures of patient-reported outcome measures (PROMs).[15] The checklist consists of 10 boxes. Boxes 1 and 2 refer to content validity. Boxes 3 to 5 refer to structural validity, internal consistency (IC), and transcultural validity. Together, these boxes form the internal structure. Boxes 6–10 form the properties of remaining measurements: reliability, measurement error, criterion validity, hypothesis test for construct validity, and responsiveness.[15,16] The boxes consist of items which evaluate the measured property. Each item can be classified as (1) Very good, (2) Adequate, (3) Doubtful, (4) Inadequate, or (5) Not applicable (NA). The final score of each box is determined by the lower option of the classified items.[15] For this study, we used cross-cultural validity, reliability, and measurement (boxes 5, 6 and 7) for analysis.

3. Results

3.1. Literature search

A total of 6.197 articles were identified. After duplicate exclusion and title reading, 458 articles were selected for the abstract. After this step, the other 376 were excluded, with 82 left for a full reading. Finally, 49 studies were part of the final review (Fig. 1).

3.2. Study characteristics

The study sample ranged from 11[17] to 243,[18] totalling 1713 subjects. The mean age of the studies ranged from 16.47 ± 0.51[19] to 59.40 ± 8.70 years.[20] Two studies evaluated individuals with an injury or history of injuries.[21,22] Most studies were on physically active individuals. The results were also analysed for professional, semi-professional, and university athletes. Although all the studies involved sports, only 15 verified the weekly training frequency. Most studies evaluated the lower limbs, and some evaluated the upper limbs. Few studies evaluated the full body or different segments. Most of the studies evaluated muscular strength or endurance, and some studies also evaluated balance, flexibility, agility and reaction time. The reliability values verified by ICC ranged from 0.26[23] to 0.99.[17,19,24–26] SEM and MDC values were delivered in percentages and absolute values (Table 2).

3.3. Quality of studies

All studies evaluated by the COSMIN checklist were ultimately classified as "Inadequate". The main flaws were found in items 5.3, 7.5 and 6.5. The items with the best ratings (Very good) were 6.2, 7.2, and 7.4. All studies were classified as "Inadequate" by box 5, 17 studies were classified as "Adequate" by box 6,[19,27–42] and six studies were "Adequate" in box 7.[33,34,36,37,41,43] The final results are shown in Table 3.

4. Discussion

The main findings of the present systematic review remark on the relevance of the psychometric properties of functional tests. The reliability value in a functional test represents how well this test can be replicated. Its value is represented by the ICC and ranges from 0 to 1. The closer to 1, the higher the reliability value. Portney and Watkins suggest that values below 0.5 are classified as poor, from 0.5 to 0.75 as moderate, from 0.75 to 0.9 as good, and greater than 0.9 as excellent

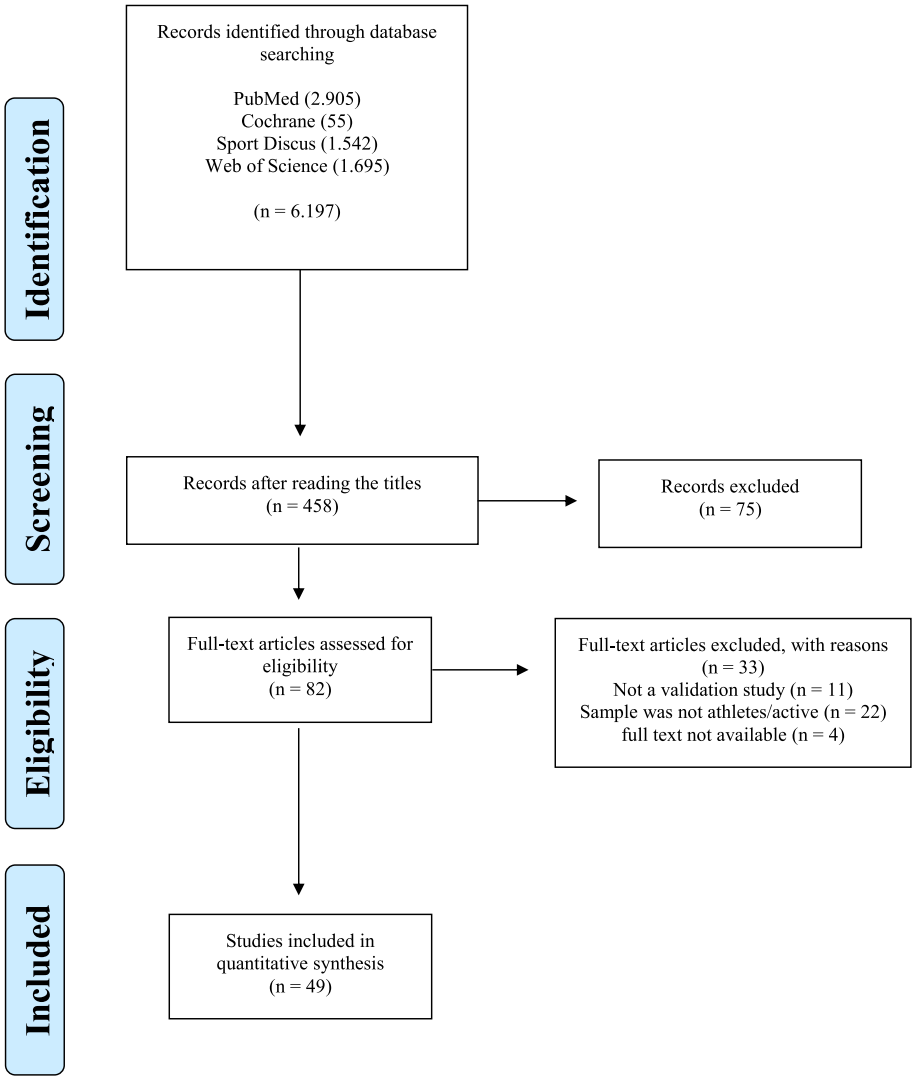

Fig. 1. Flowchart of the literature search.

reliability.[44] The core stability assessment carried out by Vera-Garcia et al. had the lowest reliability value.[23] Five tests were evaluated, and none of them presented excellent results. Most are classified as moderate, and several as poor. This difference may have occurred because of the long time lag between assessments, namely one month. Lee and Granata verified trunk stability but reported higher ICC values with an interval between one-week sessions.[45] It is difficult to assess the strength of the core muscles given the number of muscles involved.[46] The resistance of these muscles is often assessed in exercises that involve many joints or a predominance of muscles that are often not part of the core, such as the hamstrings.[46,47] In other cases, balance is tested, a factor that is also influenced by other muscle groups of the lower limbs, in addition to the individual's visual field, mobility and level of attention.[48] Strength assessment is very well structured using the isokinetic dynamometer.[1] A recent systematic review found ICC reliability values above 0.70 to assess shoulder strength.[2] Overall, the reliability values in the studies found were also well evaluated,[3–5] except in two studies.[6] Impellizzeri et al. showed low reliability in assessing symmetry between limbs.[7] Muller et al. found low-reliability values for assessing lower limbs, especially in the ankle joint. Low values were also found in the multi-joint assessment of the lower limbs, which can be explained by the greater capacity for movement coordination.[6]

In many settings, isokinetic dynamometry is unfeasible given the high cost of the device, and a lower-cost alternative should be considered.[8] Manual dynamometry is a widely used example to assess isometric strength in research and clinical settings; differently, isokinetic dynamometry assesses concentric and eccentric strength at different angular velocities. The manual dynamometry demonstrated good reliability.[9–12] The great difficulty of the manual dynamometer is the standardisation of its position and even the force imposed, in some instances, by the evaluator to resist the movement.[13] To minimise external influence, the manual dynamometer should be fixed[14,15] so there is no dependence on the strength of the evaluator. Van der Made et al.[49] found good results when evaluating hamstring strength. Almeida et al.[50] evaluated external rotation strength in the lateral position and found high-reliability values. Both are carried out in a way where the manual dynamometer is fixed, and the individual evaluated remains in a stable position. The more stable the individual is during the assessment, the greater reliability.[14] Mccall et al.[42] also evaluated hamstring strength by isometric contraction in the bridge position, a cheap and easy-to-apply test with results equivalent to equipment such as a hand-held dynamometer or other more complex ones.[34] The test is very similar to the single-leg bridge test carried out with football players but with differences in the type of contraction and positioning of the subject.[51] This test does not completely isolate the hamstring muscle, but it evaluates a multi-joint movement that is highly functional for sport. Strength assessment using the RM method has good and excellent reliability values,[37] but carrying out the method is difficult when there are

Table 2
Studies generalities.

Author and year	Participants	Female	Mean age	Sport level	Train for week	Disfunction	Test	Local	Test-retest reliability	SEM	MDC
Almeida et al., 2017	49	49	21.4 ± 3.4	Amateur/ Physically active	–	Patellofemoral pain syndrome	Hip Stability Isometric Test	Lower limb	0.98	–	–
Ashall et al., 2021	19	–	26.0 ± 5.0	Semi-professional	–	No	Hand-held dynamometer	Neck	0.72–0.89	–	–
Ashworth et al., 2018	18	0	22.4 ± 4.6	Professional	–	No	Y-Test, I-Test, T-Test with Hand-Held Dynamometry	Upper limb	0.94–0.96	4.8–10.8 %	10.7–20.1 %
Ayala et al., 2012.1	50	0	21.3 ± 2.5	Amateur/ Physically active	3–4x/ week	No	Vertical (V-HJA) and Horizontal Hip joint angle (H-HJA); Passive straight-leg raise test (PSLR)	Lower limb	H-HJA: 0.93; V-HJA: 0.92; PSLR: 0.88	–	–
Ayala et al., 2012.2	243	87	21.0 ± 2.1	Professional	3–5/ week	No	Sit-and-reach test (SRT), toe touch test (TT), Passive straight leg raise test (PSLRT)	Lower limb	SRT: 0.92; TT: 0.89; PSLRT: 0.85	SRT: 8.74 %; TT: 9.86 %; PSLRT: 5.46 %	–
Bampouras et al., 2014	46	10	23.3 ± 6.8	Amateur/ Physically active	2–3x/ week	No	Concept2 Dyno	Lower and upper limb	0.89–0.98	–	–
Bazett-Jones and Squier 2020	30	16	21.5 ± 2.4	Amateur/ Physically active	3x/ week	No	Hand-held dynamometry	Lower limb	0.62–0.90	10.3–30.6 %	10.9–89.6 %
Beato et al., 2021	20	0	23.0 ± 3.0	University athletes	>2x/ week	No	Flywheel squat test	Lower limb	0.94–0.95	–	55–61W
Burnham et al., 1995	20	0	18.9 ± 1.0	Semi-professional	–	No	Hand-Held Dynamometry	Upper limb	0.55–0.81	–	–
Cahanin et al., 2021	25	11	23.1 ± 2.2	Amateur/ Physically active	–	No	Butterfly Agility Test	Lower limb	0.94	0.75s	2.08s
Chtara et al., 2020	39	18	20.8 ± 3.0	Professional	–	No	Specific Fencing Change of Direction Test	Lower limb	0.97	0.38 %	0.08 %
Clark et al., 2019	13	7	25.6 ± 5.5	Amateur/ Physically active*	>1x/ week	No	1RM Leg Press; 1RM Knee Flexion; 1RM Knee Extension	Lower limb	1RM Leg Press: 0.94–0.98; 1RM Knee Flexion: 0.75–0.95; 1RM Knee Extension: 0.78 - 0.87	1RM Leg Press: 7.2–14.3 %; 1RM Knee Flexion: 1.9–4.9 %; 1RM Knee Extension: 3.4–4.4 %	–
Corcelle et al., 2022	13	0	20.7 ± 1.6	Amateur/ Physically active	–	No	Maximal voluntary isometric contraction and Eccentric force of hamstring	Lower limbr	0.92–0.98	3.9–6.1	–
Cramer et al., 2017	20	8	29.2 ± 4.9	Amateur/ Physically active	–	No	mUQYBT	Upper limb	0.98–0.99	0.2 cm	–
Decleve et al., 2020	30	14	20.0 ± 1.7	Amateur/ Physically active	5x/ week	No	Shoulder Endurance Test	Upper limb	0.78–0.93	10.7–16.4s	29.6–45.6s
Dirnberger et al., 2013	41	0	24.4 ± 3.1	Amateur/ Physically active	2–3x/ week	No	Isokinetic knee extension and flexion	Lower limb	0.82–0.97	–	–
Evans et al., 2007	79	47	21.2 ± 2.3	Professional	–	No	Side bridge endurance test (SBET); trunk flexor endurance tests (TFET)	Trunk	0.81–0.95	–	–
Fuller et al., 2022	52	0	16.5 ± 0.9	University athletes	–	No	Neck strength test	Neck	0.82–0.95	10.7–23.7N	41.9–90.9N

(*continued on next page*)

Table 2 (*continued*)

Author and year	Participants	Female	Mean age	Sport level	Train for week	Disfunction	Test	Local	Test-retest reliability	SEM	MDC
Garcia et al., 2023	16	0	29.5 ± 7.3	Amateur/ Physically active	–	No	Portable Traction Dynamometer	Lowe limb	0.91–0.93	13.01–17.29Nm	36.05–47.94Nm
Hadzic et al., 2012	21	12	26.2 ± 2.8	Amateur/ Physically active	–	No	Isokinetic strength test of shoulder internal and external rotators	Upper limb	0.80–0.94	6–9.9 %	–
Hartog et al., 2021	22	11	59.4 ± 8.7	Amateur/ Physically active	3x/ week	No	Q-Force lI	Lower limb	0.97	9.2–30.4 %	25.5–84.1 %
Hassen et al., 2022	36	16	20.1 ± 3.1	Professional	–	No	Specific karate agility test	Full body	0.98	1.5 %	4.18 %
Impellizzeri et al., 2008	18	0	23.0 ± 3.0	Amateur/ Physically active	–	No	Cybex NORM dynamometer for knee	Lower limb	0.29–0.87	3.2–8.7 %	8.9–24.2 %
Kambic et al., 2020	19	6	24.0 ± 3.0	Amateur/ Physically active	–	No	SMM isokinetic dynamometer for quadriceps and hamstring	Lower Limb	0.89–0.98	2.54–6.93 %	7.04–19.22 %
Lodge et al., 2020	26	0	21.2 ± 2.0	University athletes	–	No	Hamstring Solo Elite	Lowe limb	0.91	14.29–14.65 N	39.63–40.62N
Lum and Aziz 2020	30	0	26.0 ± 4.0	Professional	–	No	Isometric prone bench pull	Upper limbr	0.88–0.98	23.5–932.5N	–
McCall et al., 2015	29	0	19.6 ± 3.5	Professional	–	No	Isometric posterior lower limb strength	Lower limb	0.86–0.95	–	26.2–36.9 N
Miralles-Iborra et al., 2023	19	0	19.0 ± 1.0	Semi-professional	–	No	Isometric hamstring and quadriceps strength	Lower limb	0.80–0.90	9.1–13.5N	64–94N
Muller et al., 2007	29	0	33.8 ± 7.2	Amateur/ Physically active	3x/ week	Chronic tendon injuries	Lower limb Dynamometer system	Lower limb	0.27–0.92	5.28–16.39 %	–
O'Connor et al., 2016.1	15	0	19.4 ± 0.6	Amateur/ Physically active	–	No	Alternative trunk stability push up test	Upper limb	0.73–0.97	–	–
O'Connor et al., 2023.2	50	0	23.1 ± 4.8	Amateur/ Physically active	–	No	Hip adduction and abduction strength by sphygmomanometer and ForceFrame	Lower limb	0.70–0.92	4.3–13.7 %	12.1–37.9 %
Padulo et al., 2020	19	0	16.5 ± 0.5	Semi-professional	–	No	Portable dynamometer for quadriceps and hamstring	Lower limbr	0.87–0.99	–	–
Pojskic et al., 2019	47	14	20.2 ± 1.9	Professional	3x/ week	No	Response time test	Upper limb	0.31–0.97	24.4–65.6 ms	–
Pruyn et al., 2016	50	50	23.5 ± 2.9	Semi-professional	–	No	Vertical hop test	Lower limb	0.60–0.79	10.6–14.2 %	–
Rhodes et al., 2022	30	0	22.8 ± 5.0	University athletes	–	No	Isometric Soleus Strength Test	Lower limb	0.79–0.89	9.09–12.47 %	25.19–34.56 %
Riemann et al., 2021	35	16	24.5 ± 4.0	Amateur/ Physically active	–	No	Isokinetic Knee Dynamometer System for knee	Lower limbr	0.85–0.97	3–17.5Nm	–
Romero-Franco et al., 2016	11	4	27.9 ± 1.2	Amateur/ Physically active	3x/ week	No	Isometric strength lower limb with digital isokinetic dynamometer	Lower limb	0.76–0.99	3.9–11.9N	
Rosen et al., 2023	49	16	22.7 ± 3.4	Amateur/ Physically active	–	No	Choice-reaction hop test	Lower limbr	0.84–0.88	0.92–1.10s	2.6–3.0s

(*continued on next page*)

Table 2 (*continued*)

Author and year	Participants	Female	Mean age	Sport level	Train for week	Disfunction	Test	Local	Test-retest reliability	SEM	MDC
Ruschel et al., 2015	31	15	23.0 ± 4.0	Amateur/ Physically active	–	No	Isometric knee strength	Lower limb	0.75–0.94	11.7–18.1N	32.5–50.1N
Sánchez-Sánchez et al., 2021	19	0	21.0 ± 4.0	Semi-professional	4x/ week	No	Swing eccentric hamstring	Lower limbr	0.83–0.94	11.07–35.76 cm/s	
Sassi et al., 2009	86	34	22.5 ± 1.5	Amateur/ Physically active	–	No	Modified agility T-test (MAT)	Lower limb	0.92–0.95	–	–
Scott et al., 2022	av41	–	23.0 ± 3.5	Professional	–	No	Sub-maximal fitness test	Full body	0.80–0.94	3.3–3.5 %	
Tassignon et al., 2020	21	–	22.0 ± 1.0	Amateur/ Physically active	–	No	Reactive balance test	Lower limb	0.74–0.99	14.32–69.81 ms	39.69–193.51 ms
Thorborg et al., 2013	21	6	30. ± 8.6	Amateur/ Physically active	–	No	Isometric strength knee and hip	Lower limb	0.76–0.95	5–11 %	14–29 %
Van Bergen et al., 2023	22	0	28.5 ± 8.6	Professional	–	No	Functional Grip Strength	Upper limb	0.88–0.99	2.23–5.86 Kg	–
Velarde-Sotres et al., 2021	25	0	21.3 ± 2.4	Amateur/ Physically active	–	No	OctoBalance Test	Upper limb	0.73–0.97	–	–
Vera-Garcia et al., 2019	33	0	24.1 ± 2.9	Amateur/ Physically active	1–3x/ week	No	Three Plane Core Strength Test; Double-leg Lowering Test (DLLT); Biering-Sorensen Test (BST)	Trunk	TPCST: 0.26–0.29; DLLT: 0.55; BST: 0.81	–	–
Wollin et al., 2015	16	0	16.8 ± 0.5	Professional	–	No	Hamstring strength test	Lower limb	0.86–0.87	5–20.6 %	12.9–14 %
Yildiz et al., 2007	20	0	21.1 ± 1.8	Amateur/ Physically active	1–2x/ week	No	Ankle Isokinetic strength; one leg standing test; Single, triple, cross-over and 6 m hop test	Lower limb	Ankle Isokinetic strength: 0.86–0.89; one leg standing test: 0.92; Single course: 0.91; single distence: 0.97; triple: 0.98; cross-over: 0.89 and 6 m: 0.91	–	–

Table 3
Quality of the included studies.

Author and year	Box 5	Box 6	Box 7	Final
Almeida et al., 2017	Inadequate	Doubtful	Inadequate	Inadequate
Ashall et al., 2021	Inadequate	Doubtful	Inadequate	Inadequate
Ashworth et al., 2018	Inadequate	Doubtful	Doubtful	Inadequate
Ayala et al., 2012.1	Inadequate	Adequate	Inadequate	Inadequate
Ayala et al., 2012.2	Inadequate	Inadequate	Inadequate	Inadequate
Bampouras et al., 2014	Inadequate	Inadequate	Inadequate	Inadequate
Bazett-Jones and Squier 2020	Inadequate	Inadequate	Inadequate	Inadequate
Beato et al., 2021	Inadequate	Adequate	Inadequate	Inadequate
Burnham et al., 1995	Inadequate	Inadequate	Inadequate	Inadequate
Cahanin et al., 2021	Inadequate	Inadequate	Inadequate	Inadequate
Chtara et al., 2020	Inadequate	Adequate	Adequate	Inadequate
Clark et al., 2019	Inadequate	Adequate	Adequate	Inadequate
Corcelle et al., 2022	Inadequate	Inadequate	Inadequate	Inadequate
Cramer et al., 2017	Inadequate	Inadequate	Inadequate	Inadequate
Decleve et al., 2020	Inadequate	Inadequate	Inadequate	Inadequate
Dirnberger et al., 2013	Inadequate	Adequate	Adequate	Inadequate
Evans et al., 2007	Inadequate	Adequate	Inadequate	Inadequate
Fuller et al., 2022	Inadequate	Doubtful	Inadequate	Inadequate
Garcia et al., 2023	Inadequate	Inadequate	Inadequate	Inadequate
Hadzic et al., 2012	Inadequate	Adequate	Inadequate	Inadequate
Hartog et al., 2021	Inadequate	Doubtful	Doubtful	Inadequate
Hassen et al., 2022	Inadequate	Adequate	Adequate	Inadequate
Impellizzeri et al., 2008	Inadequate	Doubtful	Doubtful	Inadequate
Kambic et al., 2020	Inadequate	Inadequate	Inadequate	Inadequate
Lodge et al., 2020	Inadequate	Doubtful	Doubtful	Inadequate
Lum and Aziz 2020	Inadequate	Inadequate	Inadequate	Inadequate
McCall et al., 2015	Inadequate	Adequate	Inadequate	Inadequate
Miralles-Iborra et al., 2023	Inadequate	Adequate	Inadequate	Inadequate
Muller et al., 2007	Inadequate	Doubtful	Doubtful	Inadequate
O'Connor et al., 2016	Inadequate	Adequate	Inadequate	Inadequate
O'Connor et al., 2023	Inadequate	Adequate	Doubtful	Inadequate
Padulo et al., 2020	Inadequate	Adequate	Inadequate	Inadequate
Pojskic et al., 2019	Inadequate	Doubtful	Doubtful	Inadequate
Pruyn et al., 2016	Inadequate	Inadequate	Inadequate	Inadequate
Rhodes et al., 2022	Inadequate	Doubtful	Doubtful	Inadequate
Riemann et al., 2016	Inadequate	Adequate	Inadequate	Inadequate
Romero-Franco et al., 2016	Inadequate	Adequate	Inadequate	Inadequate
Rosen et al., 2023	Inadequate	Adequate	Doubtful	Inadequate
Ruschel et al., 2015	Inadequate	Doubtful	Inadequate	Inadequate
Sánchez-Sánchez et al., 2021	Inadequate	Inadequate	Inadequate	Inadequate
Sassi et al., 2009	Inadequate	Inadequate	Inadequate	Inadequate
Scott et al., 2022	Inadequate	Adequate	Adequate	Inadequate
Tassignon et al., 2020	Inadequate	Inadequate	Inadequate	Inadequate
Thorborg et al., 2013	Inadequate	Doubtful	Inadequate	Inadequate
Van Bergen et al., 2023	Inadequate	Inadequate	Inadequate	Inadequate
Velarde-Sotres et al., 2021	Inadequate	Doubtful	Inadequate	Inadequate
Vera-Garcia et al., 2019	Inadequate	Doubtful	Inadequate	Inadequate
Wollin et al., 2015	Inadequate	Adequate	Adequate	Inadequate
Yildiz et al., 2007	Inadequate	Inadequate	Inadequate	Inadequate

problems in identifying the maximum load.[52] As a result, studies estimate MR through a calculation performed after finding a submaximal load.[53,54] This can optimise time and help injured individuals not to use high loads.

The Response Time Test for Agility-Based Sports evaluated by Pojskic et al.[55] showed low results. The test has many variations, and few have exhibited poor results. What is discussed in the test is the ability to verify truly detectable changes, given that only a few seconds separate the stimulus from the response. The stimulus occurs by LED light, a factor that does not favour a high external validity.[56] The Illinois Agility Test on male football players showed more reliable values,[57] as did the MAT and agility tests.[58,59] However, it is important to take into account the specificity of the sport when deciding on which test to use.[60] Flodström et al.[61] evaluated a battery of nine functional tests within the Functional Movement Screen (FMS). This has very questionable reliability values, given the high subjectivity and classification of the tests, and this failure is very clear from the reliability values found. The assessment is carried out to check the quality of movement and neuromuscular control. Other functional tests that evaluate neuromuscular control, whether for lower or upper limbs, have higher reliability values. Tests for upper limbs (mUQYBT and CKCUEST) and lower limbs (Lateral Step Down, Leap and Catch; Single, triple, crossover, square and Timed hop test) have excellent reliability values[24,62] and range from moderate to excellent,[63,64] respectively. Pruyn et al.[65] found lower reliability values for the vertical hop test. This can be explained by the study using test-to-test ground reaction force equipment, not checking jump height. This meant the tests were carried out barefoot, which is not recommended for the hop test procedure.[66,67]

SEM represents the estimated error of some functional tests and is directly related to the reliability and MDC values. The standard error will be smaller when sample means are clustered closer to the population mean.[68] MDC is a psychometric property with a strong clinical and performance relationship. From this, we can determine the minimum value to be achieved in a re-test so that the individual or another person can notice improvement in the clinical status or training routine.[69]

Not all studies included these analyses. Additionally, the verified unit of measurement is an important component to consider. Interestingly, many studies include percentage values, which can better represent proportionally, especially when considering clinical improvement.[16] This is because some individuals will have low and others high test results, and considering an absolute value may have different representations for these same individuals.

The main factor that caused the studies to be classified as "Inadequate" was the small sample size. Although the COSMIN checklist is more specific for questionnaires and scales, this item is important so that values for the population can be considered. The study with the largest number included 243 subjects,[17] but more than one test was performed, making this sample inadequate. Sample calculation was also not performed in the studies, which is a major flaw in their design. To ensure that reliability is accurately assessed, the sample size requirement for statistical analyses must be considered.[70,71] Characterisation of the sample in greater detail also rarely occurred. Few studies have divided the results by sex, limb dominance, or sporting demand.[72] However, with very small samples, such categorisation becomes more difficult. Furthermore, few studies presented other tests and analysed agreement or correlation with the results. This is an important factor to verify whether the test is being carried out effectively for this population.[73] One of the great difficulties of science, especially linked to scales and tests, is replicating its methods to reduce errors.[74] Hence, solid methods are important in constructing validation and reliability studies. In this way, the risk of bias is reduced, and the external validity of the results for the population in question can be increased.[72,75]

This study presented only reported on three psychometric properties. However, external validity, that is, in the practical environment, is the most relevant for health professionals. The quality of the studies was not high. Unfortunately, no study achieved a result better than "Inadequate" on the Cosmin checklist. This demonstrates the need for better methods to find values of psychometric properties of functional tests aimed at the sports environment. Nevertheless, this is a broad study and a general review that can contribute to professionals needing functional tests and minimising subjectivity but still having doubts about which one to use.

The results presented in this systematic review make it easier for health professionals to evaluate their patients in the best possible way, and they can choose the best test to perform. Based on a well-performed assessment, the professional can determine the best treatment and re-evaluate with the same test. Furthermore, during the re-evaluation, the MDC helps determine the effectiveness of the treatment or program for a given variable.

According to the COSMIN checklist, this study has some limitations, such as the low quality of the validations presented. Furthermore, the heterogeneity in study design and methods made no possible to perform a formal meta-analysis. The ultimate objective of the present

investigation was to facilitate clinicians' search for a test that is reliable for their practical environment. The results obtained should point the healthcare professionals in the right direction.

5. Conclusions

Functional tests used to assess athletes generally have good reliability values. However, standardisation in the application is necessary. The training of professionals who administer the tests is essential for greater reliability. Furthermore, greater stabilisation of the subject being evaluated is necessary for strength tests to reduce compensations during the test.

Consent to publish

Not applicable.

Ethical approval

This study complies with ethical standards.

Availability of data and materials

The datasets generated during and/or analysed during the current study are available throughout the manuscript.

Author's contribution statement

TTS: Idea conception, data extraction, writing (original), study selection and data reduction, assessment of risk of bias; APR: literature search, study selection and data reduction, assessment of risk of bias; DAP: writing (original); FM: design, supervision, writing (original and review); MN: supervision, writing (review); RO: supervision, writing (original and review). All authors agreed on the final version to be published and agreed to be responsible for all aspects of the work.

Registration and protocol

International Prospective Register of Systematic Reviews (PROSPERO) number CRD42020177143.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of competing interests

The authors declare that they have no conflicts of interest.

Acknowledgements

None.

Abbreviations

CKCUEST: Closed Kinetic Chain Upper Extremity Stability Test
COSMIN Consensus-based Standards for Health Measurement Instruments
MDC Minimal Detectable Change
mUQYBT Modified Upper Quarter Y Balance Test
PICOS Population, Intervention, Comparison, Outcome, Study
PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROSPERO International Prospective Register of Systematic Reviews
SEM Standard Error of measurement

References

1. Lehman PJ, Carl RL. The preparticipation physical evaluation. *Pediatr Ann*. 2017;46:e85–e92.
2. Westrick RB, Miller JM, Carow SD, Gerber JP. Exploration of the y-balance test for assessment of upper quarter closed kinetic chain performance. *Int J Sports Phys Ther*. 2012;7:139–147.
3. McMaster DT, Gill N, Cronin J, McGuigan M. A brief review of strength and ballistic assessment methodologies in sport. *Sports Med*. 2014;44:603–623.
4. Reiman MP, Manske RC. *Functional Testing in Human Performance: 139 Tests for Sport, Fitness, Occupational Settings*. ninth ed. 2010.
5. Simmonds MJ. Measuring and managing pain and performance. *Man Ther*. 2006;11:175–179.
6. Nájera Catalán HE, Gordon D. The importance of reliability and construct validity in multidimensional poverty measurement: an illustration using the multidimensional poverty index for Latin America (MPI-LA). *J Dev Stud*. 2020;56:1763–1783.
7. Powden CJ, Hoch JM, Hoch MC. Reliability and minimal detectable change of the weight-bearing lunge test: a systematic review. *Man Ther*. 2015;20:524–532.
8. Salmond SS. Evaluating the reliability and validity of measurement instruments. *Orthop Nurs*. 2008;27:28–30.
9. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. *J Psychosom Res*. 2010;68:319–323.
10. Noble S, Scheinost D, Constable RT. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. *Neuroimage*. 2019;203.
11. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Condit Res*. 2005;19:231–240.
12. Furlan L, Sterr A. The applicability of standard error of measurement and minimal detectable change to motor learning research - a behavioural study. *Front Hum Neurosci*. 2018;12.
13. Fritz SL, Blanton S, Uswatte G, Taub E, Wolf SL. Minimal detectable change scores for the wolf motor function test. *Neurorehabilitation Neural Repair*. 2009;23:662–667.
14. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Rev Española Nutr Humana Dietética*. 2016;20:148–160.
15. Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, De Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res*. 2012;21:651–657.
16. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27:1171–1179.
17. Romero-Franco N, Jiménez-Reyes P, Montaño-Munuera JA. Validity and reliability of a low-cost digital dynamometer for measuring isometric strength of lower limb. *J Sports Sci*. 2017;35:2179–2184.
18. Ayala F, Sainz de Baranda P, De Ste Croix M, Santonja F. Reproducibility and criterion-related validity of the sit and reach test and toe touch test for estimating hamstring flexibility in recreationally active young adults. *Phys Ther Sport*. 2012;13:219–226. https://doi.org/10.1016/J.PTSP.2011.11.001.
19. Padulo J, Trajković N, Cular D, et al. Validity and reliability of isometric-bench for knee isometric assessment. *Int J Environ Res Publ Health*. 2020;17:1–8. https://doi.org/10.3390/IJERPH17124326.
20. Hartog J, Dijkstra S, Fleer J, van der Harst P, Mariani MA, van der Woude LHV. A portable isometric knee extensor strength testing device: test-retest reliability and minimal detectable change scores of the Q-Force II in healthy adults. *BMC Muscoskelet Disord*. 2021;22. https://doi.org/10.1186/S12891-021-04848-8.
21. Almeida GPL, Rodrigues HLDN, De Freitas BW, De Paula Lima PO. Reliability and validity of the hip stability isometric test (HipSIT): a new method to assess hip posterolateral muscle strength. *J Orthop Sports Phys Ther*. 2017;47:906–913.
22. Müller S, Baur H, König T, Hirschmüller A, Mayer F. Reproducibility of isokinetic single- and multi-joint strength measurements in healthy and injured athletes. *Isokinet Exerc Sci*. 2007;15:295–302.
23. Vera-Garcia FJ, López-Plaza D, Juan-Recio C, Barbado D. Tests to measure core stability in laboratory and field settings: reliability and correlation analyses. *J Appl Biomech*. 2019;35:223–231.
24. Cramer J, Quintero M, Rhinehart A, et al. Exploration of score agreement on a modified upper quarter Y-balance test kit as compared to the upper quarter Y-balance test. *Int J Sports Phys Ther*. 2017;12:117–124.
25. Tassignon B, Verschueren J, De Wachter J, et al. Test-retest, intra- and inter-rater reliability of the reactive balance test in healthy recreational athletes. *Phys Ther Sport*. 2020;46:47–53. https://doi.org/10.1016/J.PTSP.2020.08.010.
26. van Bergen NG, Soekarjo K, Van der Kamp J, Orth D. Reliability and validity of functional grip strength measures across holds and body positions in climbers: associations with skill and climbing performance. *Res Q Exerc Sport*. 2023;94:627–637. https://doi.org/10.1080/02701367.2022.2035662.
27. Ayala F, De Baranda PS, De Ste Croix M, Santonja F. Reproducibility and concurrent validity of hip joint angle test for estimating hamstring flexibility in recreationally active young men. *J strength Cond Res*. 2012;26:2372–2382. https://doi.org/10.1519/JSC.0B013E31823DB1E2.
28. Miralles-Iborra A, Moreno-Pérez V, Del Coso J, Courel-Ibáñez J, Elvira JLL. Reliability of a field-based test for hamstrings and quadriceps strength assessment in football players. *Appl Sci*. 2023;13:4918. https://doi.org/10.3390/APP13084918, 2023;13:4918.
29. O'Connor S, McCaffrey N, Whyte E, Moran K. The development and reliability of a simple field based screening tool to assess core stability in athletes. *Phys Ther Sport*. 2016;20:40–44. https://doi.org/10.1016/J.PTSP.2015.12.003.

30. O' Connor C, McIntyre M, Delahunt E, Thorborg K. Reliability and validity of common hip adduction strength measures: the ForceFrame strength testing system versus the sphygmomanometer. *Phys Ther Sport*. 2023;59:162–167.
31. Riemann BL, Watson MD, Davies GJ. Reliability and validity of a novel isokinetic knee dynamometer system. *Acta Bioeng Biomech*. 2021;23:107–116. https://doi.org/10.37190/ABB-01936-2021-03.
32. Romero-Franco N, Jiménez-Reyes P, Montaño-Munuera JA. Validity and reliability of a low-cost digital dynamometer for measuring isometric strength of lower limb. *J Sports Sci*. 2017;35:2179–2184. https://doi.org/10.1080/02640414.2016.1260152.
33. Scott TJ, McLaren SJ, Lovell R, Scott MTU, Barrett S. The reliability, validity and sensitivity of an individualised sub-maximal fitness test in elite rugby league athletes. *J Sports Sci*. 2022;40:840–852. https://doi.org/10.1080/02640414.2021.2021047.
34. Wollin M, Purdam C, Drew MK. Reliability of externally fixed dynamometry hamstring strength testing in elite youth football players. *J Sci Med Sport*. 2016;19:93–96.
35. Beato M, Fleming A, Coates A, Dello Iacono A. *Validity and Reliability of a Flywheel Squat Test in Sport*. 2020. https://doi.org/10.1080/02640414.2020.1827530.
36. Chtara H, Negra Y, Chaabene H, Chtara M, Cronin J, Chaouachi A. Validity and reliability of a new test of change of direction in fencing athletes. *Int J Environ Res Publ Health*. 2020;17:1–13. https://doi.org/10.3390/IJERPH17124545.
37. Clark NC, Reilly LJ, Davies SC. Intra-rater reliability, measurement precision, and inter-test correlations of 1RM single-leg leg-press, knee-flexion, and knee-extension in uninjured adult agility-sport athletes: considerations for right and left unilateral measurements in knee injury c. *Phys Ther Sport*. 2019;40:128–136.
38. Dirnberger J, Huber C, Hoop D, Kösters A, Müller E. Reproducibility of concentric and eccentric isokinetic multi-joint leg extension measurements using the IsoMed 2000-system. *Isokinet Exerc Sci*. 2013;21:195–202.
39. Evans K, Refshauge KM, Adams R. Trunk muscle endurance tests: reliability, and gender differences in athletes. *J Sci Med Sport*. 2007;10:447–455.
40. Hadzic V, Ursej E, Kalc M, Dervisevic E. *Reproducibility of Shoulder Short Range of Motion in Isokinetic and Isometric Strength Testing*. 2012. https://doi.org/10.1016/j.jesf.2012.10.005.
41. Ben Hassen S, Negra Y, Uthoff A, Chtara M, Jarraya M. Reliability, validity, and sensitivity of a specific agility test and its relationship with physical fitness in karate athletes. *Front Physiol*. 2022;13, 841498. https://doi.org/10.3389/FPHYS.2022.841498.
42. McCall A, Nedelec M, Carling C, Le Gall F, Berthoin S, Dupont G. Reliability and sensitivity of a simple isometric posterior lower limb muscle test in professional football players. *J Sports Sci*. 2015;33:1298–1304.
43. Dirnberger J, Huber C, Hoop D, Kösters A, Müller E. Reproducibility of concentric and eccentric isokinetic multi-joint leg extension measurements using the IsoMed 2000-system. *Isokinet Exerc Sci*. 2013;21:195–202.
44. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. third ed. 2015.
45. Lee HW, Granata KP. Process stationarity and reliability of trunk postural stability. *Clin Biomech (Bristol, Avon)*. 2008;23:735–742. https://doi.org/10.1016/J.CLINBIOMECH.2008.01.008.
46. Oliva-Lozano JM, Muyor JM. Core muscle activity during physical fitness exercises: a systematic review. *Int J Environ Res Publ Health*. 2020;17:1–42. https://doi.org/10.3390/IJERPH17124306.
47. Butowicz CM, Ebaugh DD, Noehren B, Silfies SP. Validation of two clinical measures of core stability. *Int J Sports Phys Ther*. 2016;11:15. /pmc/articles/PMC4739044/ . Accessed November 8, 2023.
48. Pu F, Sun S, Wang L, et al. Investigation of key factors affecting the balance function of older adults. *Aging Clin Exp Res*. 2015;27:139–147. https://doi.org/10.1007/S40520-014-0253-8/TABLES/5.
49. van der Made AD, Paget LDA, Altink JN, et al. Assessment of isometric knee flexor strength using hand-held dynamometry in high-level rugby players is intertester reliable. *Clin J Sport Med*. 2019. Publish Ah:0–5.
50. Almeida GPL, Rodrigues HLDN, De Freitas BW, De Paula Lima PO. Reliability and validity of the hip stability isometric test (HipSIT): a new method to assess hip posterolateral muscle strength. *J Orthop Sports Phys Ther*. 2017;47:906–913. https://doi.org/10.2519/JOSPT.2017.7274.
51. Freckleton G, Cook J, Pizzari T. The predictive validity of a single leg bridge test for hamstring injuries in Australian Rules Football Players. *Br J Sports Med*. 2014;48:713–717.
52. Niewiadomski W, Gąsiorowska A, Cybulski G, Laskowska D, Langfort J. Determination and prediction of one repetition maximum (1RM): safety considerations. *J Hum Kinet*. 2008;19:109–120.
53. Mayhew JL, Johnson BD, Lamonte MJ, Lauber D, Kemmler W. Accuracy of prediction equations for determining one repetition maximum bench press in women before and after resistance training. *J Strength Condit Res*. 2008;22:1570–1577.
54. Paul K, Mayhew J, Peterson F. A modified YMCA bench press test as a predictor of 1 repetit. *J Strength Condit Res*. 2002:440–445. Journal of Strength and Conditioning Research.
55. Pojskic H, Pagaduan J, Uzicanin E, et al. Reliability, validity and usefulness of a new response time test for agility-based sports: a simple vs. complex motor task. *J Sports Sci Med*. 2019;18:623–635.
56. Khorsan R, Crawford C. How to assess the external validity and model validity of therapeutic trials: a conceptual approach to systematic review methodology. *Evid base Compl Alternative Med*. 2014;2014.
57. Hachana Y, Chaabène H, Nabli MA, et al. Test-retest reliability, criterion-related validity, and minimal detectable change of the Illinois agility test in male team sport athletes. *J Strength Condit Res*. 2013;27:2752–2759.
58. Sassi RH, Dardouri W, Yahmed MH, Gmada N, Mahfoudhi ME, Gharbi Z. Relative and absolute reliability of a modified agility t-test and its relationship with vertical jump and straight sprint. *J Strength Condit Res*. 2009;23:1644–1651. https://doi.org/10.1519/JSC.0B013E3181B425D2.
59. Zemková E. Agility index as a measurement tool based on stimuli number and traveling distances. *J Strength Condit Res*. 2017;31:2141–2146.
60. Granacher U, Borde R. Effects of sport-specific training during the early stages of long-term athlete development on physical fitness, body composition, cognitive, and academic performances. *Front Physiol*. 2017;8 OCT:810.
61. Flodström F, Heijne A, Batt ME, Frohm A. The nine test screening battery - normative values on a group of recreational athletes. *Int J Sports Phys Ther*. 2016;11:936–944.
62. Tucci HT, Martins J, Sposito GDC, Camarini PMF, De Oliveira AS. Closed Kinetic Chain Upper Extremity Stability test (CKCUES test): a reliability study in persons with and without shoulder impingement syndrome. *BMC Muscoskelet Disord*. 2014;15.
63. Yildiz Y, Sekir U, Hazneci B, Ors F, Saka Tolga, Aydin T. Reliability of a functional test battery evaluating functionality, proprioception and strength of the ankle joint. *Turk J Med Sci*. 2009;1:115–123.
64. Haitz K, Shultz R, Hodgins M, Matheson GO. Test-Retest and interrater reliability of the functional lower extremity evaluation. *J Orthop Sports Phys Ther*. 2014;44:947–954.
65. Pruyn EC, Watsford ML, Murphy AJ. Validity and reliability of three methods of stiffness assessment. *J Sport Heal Sci*. 2016;5:476–483.
66. Reid A, Birmingham TB, Stratford PW, Alcock GK, Giffin JR. Hop testing provides a reliable and valid outcome measure during rehabilitation after anterior cruciate ligament reconstruction. *Phys Ther*. 2007;87:337–349.
67. Gustavsson A, Neeter C, Thomeé P, et al. A test battery for evaluating hop performance in patients with an ACL injury and patients who have undergone ACL reconstruction. *Knee Surg Sports Traumatol Arthrosc*. 2006;14:778–788.
68. Streiner DL. Maintaining standards: differences between the standard deviation and standard error, and when to use each. *Can J Psychiatr*. 1996;41:498–502.
69. Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*. 2007;7:541–546.
70. Arifin WN. A web-based sample size calculator for reliability studies. *Educ Méd J*. 2018;10:67–76.
71. Arifin WN. Introduction to sample size calculation. *Educ Méd J*. 2013;5.
72. Mokkink LB, Terwee CB, Gibbons E, et al. *Inter-rater Agreement and Reliability of the COSMIN (COnsensus-Based Standards for the Selection of Health Status Measurement Instruments) Checklist*. 2010.
73. Berry KJ, Mielke PW. A generalization of cohen's kappa agreement measure to interval measurement and multiple raters. *Educ Psychol Meas*. 1988;48:921–933.
74. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science*. 2015;349, aac4716.
75. Matheson GJ. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. *PeerJ*. 2019;2019.