Brazilian Journal of Physical Therapy 28 (2024) 100575

Brazilian Journal of Physical Therapy

https://www.journals.elsevier.com/brazilian-journal-of-physical-therapy

SYSTEMATIC REVIEW

Measurement properties of upper extremity physical performance tests in athletes: a systematic review

Germanna Medeiros Barbosa[a,*], Letícia Bojikian Calixtre[b], Hilmaynne Renaly Fonseca Fialho[a], Francisco Locks[b], Danilo Harudy Kamonseki[c,d]

[a] *Post-Graduate Program in Rehabilitation Sciences, Faculty of Health Sciences of Trairi, Universidade Federal do Rio Grande do Norte (UFRN), Santa Cruz, RN, Brazil*
[b] *Department of Physical Therapy, Universidade de Pernambuco (UPE), Petrolina, PE, Brazil*
[c] *Laboratory of Analysis and Intervention of the Shoulder Complex, Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brazil*
[d] *Department of Physical Therapy, Universidade Federal da Paraíba (UFPB), São Carlos, SP, Brazil*

Received 29 October 2022; received in revised form 15 November 2023; accepted 3 December 2023
Available online 14 December 2023

KEYWORDS
Clinimetric;
Functional tests;
Overhead;
Psychometric;
Reliability;
Validity

Abstract
Background: Upper extremity Physical Performance Tests (PPTs) have been used in sports contexts to provide functional status of the athletes. However, whether these tests present appropriate measurement properties to be considered a valuable measurement is not clear.
Objective: To systematically review the measurement properties of upper extremity PPTs in athletes.
Methods: Databases (e.g., Medline, EMBASE, CINAHL, SPORTDiscus, CENTRAL) were searched in March 2021. Two reviewers independently rated the methodological quality using the 4-point Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) checklist. Quality of evidence was graded by measurement property for each test, considering the adequacy, the sample size, and the methodological quality of the studies.
Results: Fifteen studies were included with a pooled sample of 684 athletes. The PPTs analyzed were Arm-Jump Board Test, Closed Kinetic Chain Upper Extremity Stability Test (CKCUEST), Finger Hang Test, Medicine Ball Explosive Power Test, One-Arm Hop Test, Posterior Shoulder Endurance Test, Pull-Up Shoulder Endurance Test, Repetition to Failure Assessment, Seated Medicine Ball Throw Test (SMBT), Seated Single-Arm Shot-Put Test (SSPT), Shoulder Endurance Test, Two-Arm Bent Hang Test, Unilateral Seated Shot-Put Test, and Upper Limb Rotation Test. Evidence synthesis provided moderate and high-quality evidence for sufficient inter-session and intra-session reliability of the CKCUEST, respectively. There was moderate evidence for sufficient inter-session reliability of the SSPT and for insufficient validity of the SMBT.
Conclusion: The CKCUEST and the SSPT are sufficiently reliable in athletes. More studies are needed to investigate other psychometric properties for these tests and other upper extremity PPTs.

* Corresponding author at: Faculdade de Ciências da Saúde do Trairi, Universidade Federal do Rio Grande do Norte, Rua Vila Trairi, S/N, Centro, Santa Cruz, RN CEP: 59200-000, Brazil.
E-mail: germanna.barbosa@ufrn.br (G.M. Barbosa).

https://doi.org/10.1016/j.bjpt.2023.100575

Introduction

Physical performance-based tests (PPTs) are routinely used in sports rehabilitation and prevention to assess physical function related to sports demand, such as strength, power, and agility,[1] and provide functional status of the athletes.[1–3] PPTs are a low-tech, not time-consuming, portable, and easy-to-administer way of assessment that can be performed in different environments with minimal material.[2,3]

The International Olympic Committee recommends that screening tests must be reliable, with appropriate sensitivity and specificity, affordable, easy to perform, and widely available.[4] The PPTs for the lower limb have been extensively investigated and some PPTs have appropriate measurement properties.[3,5,6] In contrast, studies on upper extremity PPTs are scarce with limited evidence related to measurement properties.[2] In 2016, a systematic review identified eleven studies that examined the measurement properties of six upper extremity PPTs,[2] and only two showed moderate positive evidence for reliability and one for validity. Nevertheless, recent studies have been published and updated information on the evidence of upper extremity PPTs is needed to guide clinicians and researchers in the assessment of athletes.

Therefore, the purpose of this systematic review is to summarize and analyze the current evidence regarding the measurement properties of upper extremity PPTs in athletes.

Methods

This systematic review followed the PRISMA checklist[7] and was prospectively registered at PROSPERO (CRD 42021241883).

Search methods for identification of studies

Electronic searches were performed in Medline (via Ovid), EMBASE, CINAHL, Cochrane Central Register for Controlled Trials (CENTRAL), SCOPUS, SPORTdiscus, and Web of Science, from inception up to July 2022. Keywords related to PPTs, athletes, upper extremity, and measurement properties were combined and adjusted for each database (Supplementary material 1). The reference lists from the included articles were screened to identify potentially relevant studies.

Study selection

Two reviewers independently analyzed titles and abstracts of the retrieved publications and, thereafter, analyzed full texts, according to eligibility criteria. The selection process was conducted by consensus, and a third reviewer was consulted in case of disagreement using the software State of the Art through Systematic Review.

Eligibility criteria

Participants

Studies were included if they assessed athletes or participants enrolled in any sports practice of any sport from both sexes, without restrictions related to age, level of sports practice (e.g., recreational, high school level, semi-professional, and professional), and presence of injury.

Type of studies

Studies with any design and language that verified the measurement properties of upper extremity PPTs in athletes were included. PPTs were defined as assessments that measure constructs related to muscle strength and power, agility, endurance, flexibility, and readiness for return-to-play that stimulate activities or gestures of sports practice,[2,5] using affordable, portable, and readily available equipment, with results reported as the number of repetitions, distance (centimeters or meters), or duration (seconds or minutes). Upper extremity was defined as the region spanning from shoulder girdle to the end of the fingers. Studies that investigated the measurement properties of technology-dependent instruments, including 2/3-dimensional motion analysis system, upper body ergometers, rowing ergometers, and dynamometer, were excluded.

Outcome measure

Primary studies were included if they reported one or more measurement properties, which were defined according to the Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) taxonomy.[8]

Data extraction

Three reviewers independently extracted the data and a fourth reviewer verified the data in case of discrepancies. A standardized form was used to extract the data, including information regarding characteristics of the study, participants, PPT, and measurement properties.

Methodology quality

Two reviewers independently evaluated the methodological quality of each measurement property of the included studies, and a third author was consulted in case of discrepancies, using the COSMIN 4-point checklist,[9,10] which scored as *very good, adequate, doubtful, or inadequate*. For scoring the quality of inter-session, intra-session, and inter-rater reliability studies, the item "assignment of the score or determination of the biological value" was not considered for analysis due to PPTs not involving biological samplings (e.g., blood and urine). Also, the item "administration of measurements" was not considered for intra-session reliability analysis, because the assessor would necessarily know the value previously obtained by that same participant when repeating that measurement. The total score was determined by taking the lowest score (worst score counts method).[10]

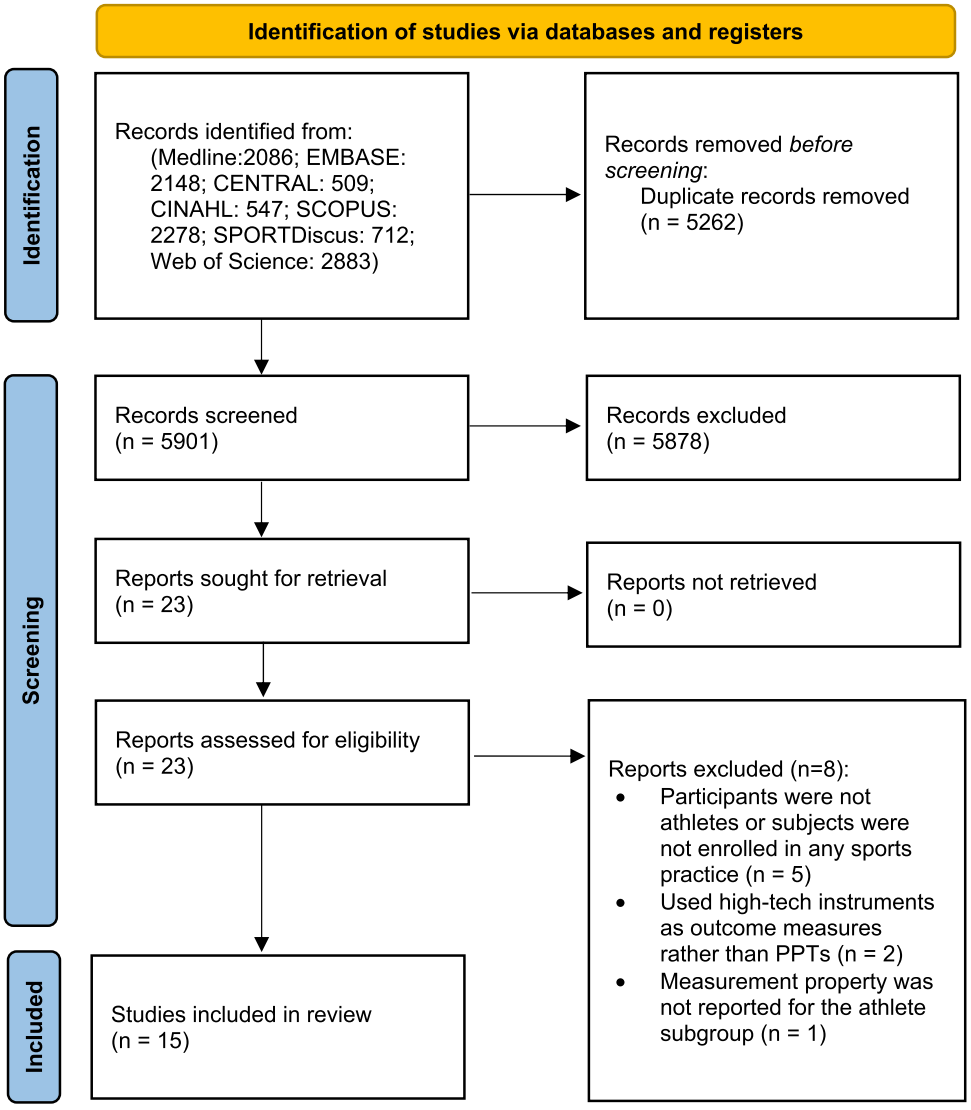

Fig. 1 PRISMA flow diagram of search strategies and results.

Quality criteria for measurement properties

The adequacy of measurement properties was assessed with the adapted version of Terwee et al.[11] Each measurement property was rated as *sufficient*, *insufficient*, or *indeterminate* (Supplementary material 2).

Grading the quality of the evidence

The quality of evidence (QoE) was graded by measurement property for each PPT, according to previous systematic reviews (Supplementary material 3).[12–14]

Results

Study selection

The literature search retrieved 11,163 studies and, of those, 5262 were duplicates, which resulted in 5901 for assessment. The assessment of title and abstract excluded 5878 because the included individuals were not athletes, or individuals were not enrolled in sports practice, and/or measurement properties of PPTs were not assessed. Twenty-three were considered in the full-text assessment and 15 studies were included (Fig. 1).[16–30]

Characteristics of the primary studies

The characteristics of included studies are described in Table 1. The number of athletes in each study ranged from 14 to 132 (pooled sample: 684; 70.6% men) and the mean age of 22.0 ± 3.1 years, which ranged from 14.7 ± 1.4 to 27.3 ± 7 years old.

Evidence synthesis of the measurement properties for each PPT reported on primary studies

Different PPTs were investigated by the primary studies (Table 1 and Fig. 2 (A-S)). The results of reliability data and measurement error are described in Table 2 and the results of the validity data described in Table 3. Methodological

Table 1 Data extraction of the included studies.

Study	Sample	Name of the PPT	Description of the PPT	Measurement property
Stockbrugger et al. (2001)[17]	*n* = 20 (10 men, 10 women) Sport: outdoor beach volleyball Level: **competitive** Age: 22.8 ± 3.7 y/o	Medicine Ball Explosive Power Test	Participants started with the feet shoulder-width apart and holding a medicine ball (3 kg) with arms straight out front at shoulder height. After the countermovement (flexing the hips and knees), participants extended the knees and trunk and threw the ball up and back over the head (optimally at about 45°). The mean distance (m) of three trials was considered the score (Fig. 2A).	Reliability
Falsone et al. (2002)[16]	*n* = 26 (all men) Sports: 13 wrestling, 13 football Level: **collegiate** Age: wrestlers 20.3 ± 1.6 y/o, football players 20.0 ± 1.7 y/o	One-Arm Hop Test	Participants were positioned with the one-arm push-up position and performed five times one-arm hops onto a 10.2 cm step as quickly as possible. The time (s) to complete five times one-arm hops was considered the score (Fig. 2B).	Reliability
Laffaye et al. (2014)[25]	*n* = 34 (all men) Sport: rock climbing (15 route specialists and 9 bouldering specialists) Level: **novice, skilled, and elite** Age: novice 21.5 ± 7 y/o, skilled 25.4 ± 7 y/o, elite 24.8 ± 6 y/o	Arm-Jump Board Test	A board with a scale in cm and two climbing holds (easy "jug" grip) 55 cm apart were placed on a wall. Participants started holding the grips and then pull-up as high as possible and touched the board with both hands. Three trials with a 3-min rest were performed, and the best trial (cm) was considered the score (Fig. 2C).	Validity
Tucci et al. (2014)[18]	*n* = 40 (20 men, 20 women) Sport: upper extremity sport-specific Level: **recreational** Age: men 23.15 ± 2.48 y/o, women 21.75 ± 1.37 y/o	CKCUEST	Two pieces of tape were parallelly placed on the floor 91.4 cm apart. Participants adopted a push-up position (women adopted a kneeling push-up position) with their hands over the tapes and alternately moved one hand to touch the dorsum of the opposite hand, as quickly as possible, during 15-s. Three trials with 45-s intervals were performed, and the average of touches was considered the score (Figures 2D and E).	Reliability
Degot et al. (2019)[19]	*n* = 27 (all men) Sport: 11 rugby, 5 judo, 3 soccer, 2 fitness, 2 basketball, 1 climbing, 1 volleyball, 1 yoga, 1 running Level: **not reported** Age: 22.5 ± 3.2 y/o	m-CKCUEST (1)	Two pieces of tape were parallelly placed on the floor at a distance of one-half of the participant's arm span. Participants adopted a push-up position with their hands over the tapes and alternately moved one hand to touch the floor outside the opposite hand, as quickly as possible, during 15-s. Three trials with 45-s intervals were performed, and the average of touches was considered the score. Muscular Endurance Index: following the three sets of 15-s, participants performed four trials of 15-s with no interval of m-CKCUEST (Fig. 2F).	Reliability
Hollstadt et al. (2020)[21]	*n* = 15 (8 men, 7 women) Sport: basketball Level: **NCAA Division I** Age: 19.5 ± 1.4 y/o	m-CKCUEST (2)	Two pieces of tape were parallelly placed on the floor 91.4 cm apart. Participants adopted a push-up position with their hands located directly under their shoulders and performed cross-body reaches to touch the contralateral piece of tape alternating each hand, as quickly as possible, during 15-s. The number of touches during one trial was considered the score (Fig. 2G).	Reliability
Kumar er al. (2020)[24]	*n* = 100 (all men) Sports: 36 Greco-Roman wrestling, 34 boxing, 30 freestyle wrestling Level: **competitive** Age: 22.9 ± 2.97 y/o	Seated Medicine Ball Throw Test	Participants were seated on the floor with their back against a wall and with minimal or no knee flexion, holding a 3 kg medicine ball with both hands, and throwing it as far as possible away from the center of their chest. Three trials with a 90-s rest were performed, and the highest trial (m) was considered the score (Fig. 2H).	Validity
Pinheiro et al. (2020)[29]	*n* = 30 individuals with shoulder pain (19 men, 11 women) Sport: 7 wt training, 4 volleyball, 4 basketball, 4 swimming, 2 functional training; 1 judo, 1 karate, 1 muay thai, 1 rugby, 1 capoeira, 1 surf, 1 badminton, 1 handball Level: **recreational or competitive** Age: 23.70 ± 4.47 y/o	Seated Single-Arm Shot-Put Test	Participants seated on the floor with their back against a wall, holding a 3 kg ball with one hand, and threw it as far as possible. Three trials with a 60-s rest were performed, and the average distance (cm) was considered the score (Fig. 2I).	Reliability
Popchak et al. (2020)[30]	*n* = 30 (19 men, 11 women) Sport: N/A Level: **recreational** Age: 24.0 ± 1.6 y/o	CKCUEST Unilateral Seated Shot-Put Test Repetition to Failure Assessment	CKCUEST: two pieces of tape 91.4 cm apart were placed on the floor. Participants assumed a push-up position and alternatingly moved one hand to touch the contralateral hand, as fast as possible, for 15-s. Three trials were performed, and the average of touches was considered the score (Figures 2D and E). Unilateral Seated Shot-Put Test: participants seated on the floor with their back against a wood box, held a 2.72 kg medicine ball with one hand and threw it as far as possible. Three trials were performed, and the average distance (cm) was the score (Fig. 2H). Repetition to Failure Assessment: participants performed shoulder ER at 0° of shoulder abduction (side-lying) (Fig. 2J), ER at 90° of shoulder abduction (prone) (Fig. 2K), and shoulder horizontal abduction at 120° of arm elevation (prone) (Fig. 2L). The resistance was 5% of the body weight for ERs and 2% for horizontal abduction. The test was ended	Reliability and validity

Table 1 (*Continued*)

Study	Sample	Name of the PPT	Description of the PPT	Measurement property
			when the participant was unable to complete a repetition through a full ROM, maintain pace with the metronome (speed of 1-s up and 1-s down), or exhibited any compensatory movements. The number of repetitions was the score.	
Decleve et al. (2020)[26]	*n* = 91 (46 men, 45 women) Sport: overhead sports (volleyball, handball, tennis, swimming) Level: **recreational** Age: men 21.5 ± 2.27 y/o; women 21.07 ± 2.29 y/o	Upper Limb Rotation Test	Participants adopted a modified push-up position (on elbows) next to a wall and performed a trunk rotation coupled with 90° of shoulder ER and 90° of shoulder abduction touching a tape placed vertically on the wall, as quickly as possible, for 15-s. Three trials, with 45-s between trials, for each side, were performed, and the average was considered the score (Fig. 2M).	Reliability
Decleve et al. (2021a)[27]	*n* = 73 (41 men, 32 women) Sport: 39 basketball, 34 volleyball Level: **NR** Age: 14.7 ± 1.4 y/o	m-CKCUEST (3)	Two pieces of tape were placed on the floor at a distance according to the participant's inter-acromial distance. Participants adopted a push-up position with the hands over the tapes (aligned with shoulders and with inter-acromial distance) and alternately moved one hand to touch the dorsum of the opposite hand as quickly as possible during 15-s. Three trials with a 45-s interval were performed, and the average of touches was considered the score (Fig. 2N).	Reliability
Decleve et al. (2021b)[28]	*n* = 30 (16 men, 14 women) Sport: overhead sports Level: **competitive** Age: 20 ± 1.76 y/o	Shoulder Endurance Test	Participants adopted a stand-up straight position with the back against a wall and tested the arm at 90° of flexion, holding a 1-m elastic band (green Theraband® for males and red for females), and pulled the elastic band from the starting position (90° forward flexion) to an ending position (90° of shoulder ER and 90° of shoulder abduction). Participants pulled the elastic band in a cadence of 60 bpm, which increased every 20-s to 150 bpm. Cadence remained at 150 bpm until the participant presented fatigue. The duration of the test (s) was considered the score (Fig. 2O).	Reliability and validity
Degot et al. (2021)[20]	*n* = 22 (all men) Sport: 11 rugby, 5 judo, 3 soccer, 2 strength training, 2 basketball, 1 climbing, 1 volleyball, 1 yoga, 1 running Level: **University** Age: 22.5 ± 3.2 y/o	USSPT	Participants seated on the floor with half of their back and head against a wall, held a 3 kg medicine ball at shoulder-height with one hand and threw it as far as possible. Three trials with a 30-s rest were performed, and the highest trial (cm) was considered the score (Fig. 2I).	Reliability
Powell et al. (2021)[29]	*n* = 14 (8 men, 6 women) Sport: canoe Level: **elite** Age: 22.5 ± 4.48 y/o	Posterior Shoulder Endurance Test	Participants were positioned in prone, with arm resting at 90° forward flexion, glenohumeral in ER, and holding a weight (2% of body mass). A metronome was set to 60 Hz, and the participants raised their arm on the first beat, hold the arm in 90° abduction for one beat and lower on the third beat to the start position before repeating. The number of repetitions until signs or report of fatigue was considered the score (Fig. 2P).	Reliability
Draper et al. (2022)[23]	*n* = 132 (87 men, 45 women) Sport: climbing Level: **lower grade, intermediate, advanced and elite** Age: 27.3 ± 7 y/o	Finger Hang Test Two-Arm Bent Hang Test Pull-Up Shoulder Endurance Test	Finger Hang Test: participants positioned both hands onto a rung with straight arms, shoulder width apart, and preferred grip. Participants should maintain their position as long as possible. The test duration (s) until the participant is unable to hold onto the rung was considered the score (Fig. 2Q). Two-Arm Bent Hang Test: participants positioned both hands in a "pull up" position on a bar, with fingers forward, shoulder width apart and chin above the bar. Participants should maintain their position as long as possible. The duration of the test (s) until the participant is unable to maintain the chin above the height of the bar was considered the score (Fig. 2R). Pull-Up Shoulder Endurance Test: participants positioned both hands in a "pull up" position (dead hang) on a bar, with fingers forward and shoulder width apart. A metronome was set to 60 bpm/1 Hz, and the participants raised themselves up to an L-hang position (elbows flexed to 90°), up to full lock with chin above the bar, reverse down to L-hang and then finishes the repetition in a dead hang. The number of full repetitions until voluntary fatigue was considered the score (Fig. 2S).	Reliability

Abbreviations: °, degrees; bpm, beats per minute; CKCUEST, Closed Kinetic Chain Upper Extremity Stability Test; cm, centimeters; ER, external rotation; Hz, Hertz; kg, kilograms; m, meters; m-CKCUEST, modified CKCUEST; min, minutes; NCAA, National Collegiate Athletic Association; PPT, Physical Performance Test; s, seconds.

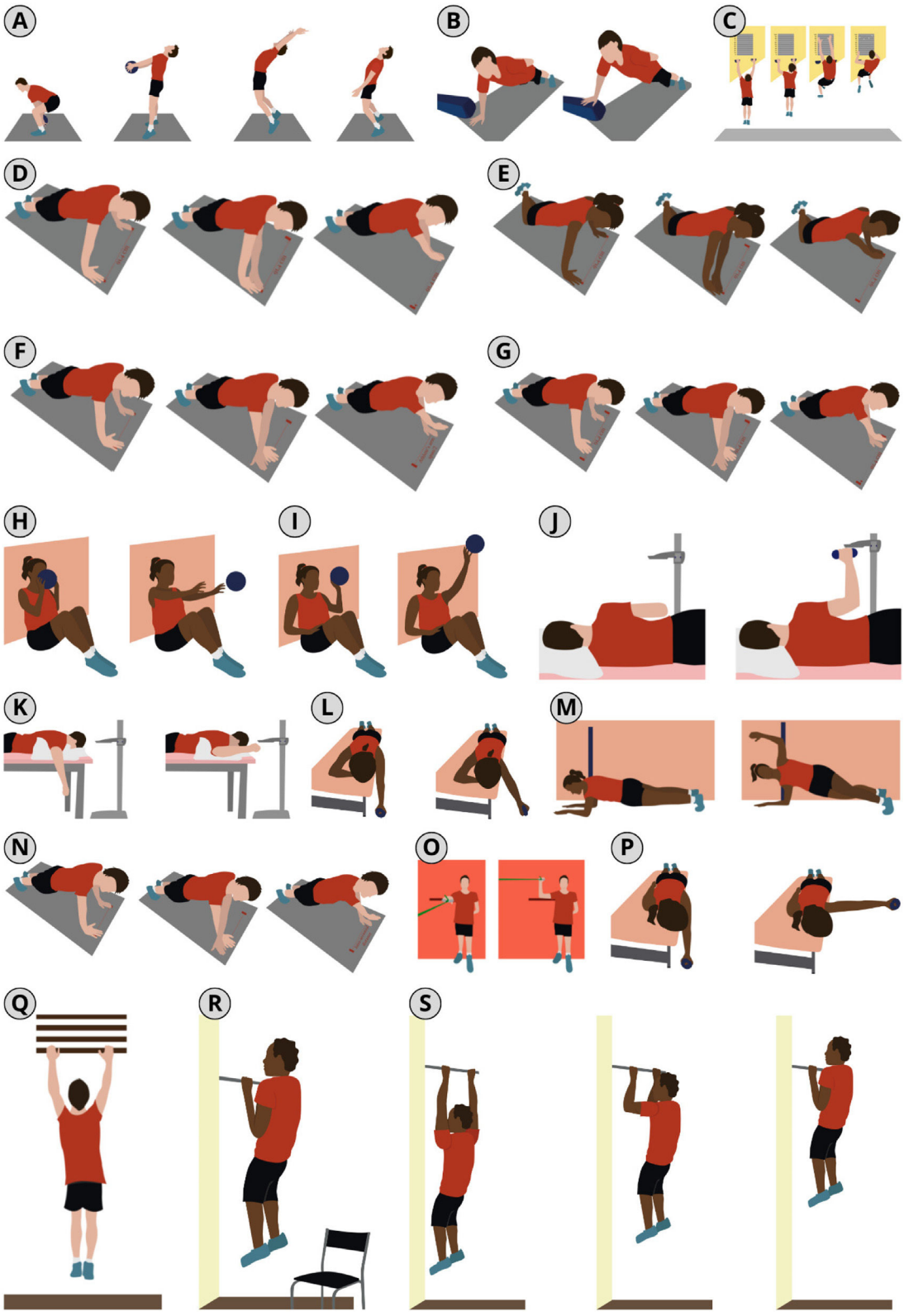

Fig. 2 Physical performance tests. Medicine Ball Explosive Power Test (A); One-Arm Hop Test (B); Arm-Jump Board Test (C); Closed Kinetic Chain Upper Extremity Stability Test (CKCUEST) (men position) (D); CKCUEST (women position) (E); Modified CKCUEST – 1 (F); Modified CKCUEST – 2 (G); Seated Medicine Ball Throw Test (H); Seated Single-Arm/Unilateral Seated Shot-Put Test (I); Repetition to Failure Assessment: external rotation (ER) at 0° of shoulder abduction (side-lying -J); ER at 90° of shoulder abduction (prone - K); and shoulder horizontal abduction at 120° of arm elevation (prone - L); Upper Limb Rotation Test (M) Modified CKCUEST – 3 (N); Shoulder Endurance Test (O); Posterior Shoulder Endurance Test (P); Finger Hang Test (Q); Two-Arm Bent Hang Test (R); Pull-Up Shoulder Endurance Test (S).

quality and quality criteria for rating the results are described in Supplementary material 4 and 5, respectively.

Medicine ball explosive power test (Fig. 2A)
One study[16] investigated the inter-session reliability in 20 competitive beach volleyball athletes. Reliability was classified as *sufficient* (ICC = 0.99) with methodological quality scored as *doubtful*.[9,10] The QoE was rated as *low* due to the low number of studies, small sample size, and *doubtful* methodology quality.

One-Arm hop test (Fig. 2B)
One study[15] with 26 uninjured collegiate athletes investigated the intra-session reliability, which was classified as *sufficient* for 13 wrestlers (ICC = 0.81) and 13 football players (ICC = 0.78) and the methodological quality was rated as *doubtful*.[9,10] The quality of the evidence was rated as *low* due to the low number of studies, small sample size, and *doubtful* methodology quality.

Arm-Jump board test (Fig. 2C)
One study[19] with 34 athletes investigated the concurrent validity using the velocity, time, index of efficiency, relative and absolute power, collected with a 3D accelerometer. Correlations varied from weak to strong and validity was classified as *indeterminate*. The methodological quality was scored as *doubtful* due to the lack of information about measurement properties of the 3D accelerometer.[9,10] Based on the methodological quality and criteria for rating the results, *no evidence* is available for the validity of this test.

CKCUEST (Figures 2D, E, F, G, and N)
Five studies[21,24–26,28] reported the measurement properties of the CKCUEST. Two studies[24,25] performed the test according to the original version described in the literature and three[21,26,28] modified the distance between the hands or the duration or interval between the series. Three studies[21,25,26] with a pooled sample size of 140 athletes investigated intra-session reliability, which was classified as *sufficient* (intra-class correlation coefficient [ICC] = 0.86–0.95). The methodological quality of those studies was scored as *very good*,[9,10] which resulted in *high* QoE. Three studies (n = 140)[21,25,26] presented intra-session standard error of measurement (SEM) and minimal detectable change (MDC), which were rated as *indeterminate* because the minimal important changes (MIC) have not been defined for the CKCUEST.

Five studies[21,24–26,28] (n = 185 athletes) verified the inter-session reliability, which was rated as *sufficient* (ICC = 0.79–0.93). The methodological quality of those studies was scored as *doubtful* or *inadequate*[9,10] and the QoE was *moderate*. Three studies[21,24,26] (pooled sample size of 130) presented inter-session SEM and MDC, which were rated as *indeterminate* because these properties have not been defined for the CKCUEST.

One study[24] investigated the concurrent validity of the CKCUEST against isokinetic shoulder external rotators (ER) and internal rotators (IR) strength, which was classified as *insufficient* validity due to a moderately positive correlation (r = 0.55 to 0.59). The methodological quality of the study was rated as *very good*,[9,10] and the QoE was *low* due to small sample size (n = 30).

Seated medicine ball throw test (Fig. 2H)
One study[18] investigated the concurrent validity of this test against the absolute peak power for the upper body during the Wingate Anaerobic Test using a modified electromagnetically braked crank-arm ergometer. Correlations were moderately positive in boxers, freestyle wrestlers, and Greco-Roman wrestlers (r = 0.40–0.54), which resulted in *insufficient* validity. The methodological quality of the study was *adequate* and the QoE was *moderate*, based on a sample of 100 athletes.

Seated single-arm/unilateral seated shot-put test (Fig. 2I)
Three studies[24,27,29] named this same test differently, which were pooled for the evidence synthesis. The methodological quality of the studies was scored as *doubtful*[9,10] for reliability due to the lack of information about the time interval,[24,29] similar assessment conditions,[29] and/or the administration of measurements.[24,29]

Inter-rater reliability was tested in 30 recreational or competitive athletes with shoulder pain from different sports.[29] Although the reliability was *sufficient* (ICC = 0.97), the *doubtful*[9,10] methodological quality and small sample size resulted in a *low* QoE. Inter-rater SEM and MDC were rated as *indeterminate* because the MIC have not been defined for this test.

The inter-session reliability, analyzed by three studies (n = 82 athletes),[24,27,29] was classified as *sufficient* (ICC = 0.92–0.94). The methodological quality of those studies was rated as *doubtful*,[9,10] which resulted in *moderate* QoE. Inter-session SEM and MDC were reported in cm by two studies[24,29] and in cm/kg$^{0.35}$ by one study,[27] which were rated as *indeterminate* because the MIC have not been defined.

Intra-session reliability was tested in 22 male athletes from different sports.[27] Reliability was considered *sufficient* in both sessions (ICC = 0.78 to 0.94), methodological quality was *adequate*[9,10] and the QoE was *low*. Intra-rater SEM and MDC were rated as *indeterminate* because the MIC have not been defined (Table 2).

One study[24] with 30 athletes investigated the concurrent validity of this test against isokinetic shoulder strength during ER and IR, which resulted in *sufficient* validity due to strong and positive correlations (r = 0.73 to 0.83). The methodological quality of the study was rated as *very good*,[9,10] and the QoE was *low* due to small sample size.

Repetition to failure test (Fig. 2J, K, and L)
One study[24] with 30 recreational athletes investigated the inter-session reliability of the posterior shoulder muscles in three different test positions: i) sideling ER at 0° abduction, ii) prone ER at 90° abduction, and iii) prone horizontal abduction at 120°. Test-retest reliability was classified as *insufficient* (ICC = 0.48–0.57) and the methodological quality was scored as *doubtful*.[9,10] The QoE was rated as *low* due to the low number of studies, small sample size, and *doubtful* methodology quality. The same study provided SEM and MDC_{95}, which were rated as *indeterminate* because the MIC have not been defined.

This study[24] has also investigated the concurrent validity of the Repetition to Failure Assessment against isokinetic shoulder ER and IR strength. The correlations were weakly positive (r = 0.20 to 0.40), which resulted in *insufficient*

Table 2 Results of the included studies that assessed reliability and measurement error.

Study/PPT	Reliability (interval between test-retest)	Type of analysis	Reliability			Measurement error		
			Result	Study quality	Rating	Result	Study quality	Rating
Stockbrugger et al. (2001)[17] • Medicine Ball Explosive Power Test	Intersession (5–21 days)	ICC Standard Error of Estimate	***Inter-session*** **Medicine ball throw distance (m)** ICC = 0.996	Doubtful	+	NR	NA	NA
Falsone et al. (2002)[16] • One-Arm Hop Test	Intersession (1–2 days)	$ICC_{2,1}$ Mean Absolute Difference	***Inter-session*** **Wrestlers** ICC = 0.81 **Football players** ICC = 0.78	Very good	+	NR	NA	NA
Tucci et al. (2014)[18] • CKCUEST	Intersession (7 days) Intrasession (45-s)	$ICC_{2,3}$ (95% CI) SEM MDC_{95}	***Inter-session*** **Number of touches** Male: ICC = 0.89 (0.71, 0.96) Female: ICC = 0.85 (0.62, 0.94) **Power** Male: ICC = 0.84 (0.58, 0.94) Female: ICC = 0.82 (0.55, 0.93) **Normalized score** Male: ICC = 0.90 (0.75, 0.96) Female: ICC = 0.87 (0.67, 0.95)	Doubtful	+	NR	NA	NA
			Intra-session (trial-to-trial) **Session 1** Male: ICC: 0.93 (0.95, 0.99) Female: ICC = 0.90 (0.90, 0.99) **Session 2** Male: ICC = 0.95 (0.89, 0.98) Female: ICC= 0.95 (0.90, 0.98)	Very good	+	**Session 1** Male: SEM = 2.00 reps MDC_{95} = 2.82 reps Female: SEM = 2.76 reps MDC_{95} = 3.91 reps **Session 2** Male: SEM = 2.00 reps MDC_{95} = 2.82 reps Female: SEM = 2.76 reps MDC_{95} = 3.91 reps	Very good	?
Degot et al. (2019)[19]	Intersession (7 days)	$ICC_{3,k}$ (95% CI) SEM (95% CI)	***Inter-session***	Doubtful	+		Very good	+

Table 2 (*Continued*)

Study/PPT	Reliability (interval between test-retest)	Type of analysis	Reliability			Measurement error		
			Result	Study quality	Rating	Result	Study quality	Rating
• m-CKCUEST (1)	Intrasession (45-s)	MDC_{95} B-A plots (limits of agreement) CV%	**m-CKCUEST score** ICC = 0.89 (0.77, 0.95) **Muscular Endurance Index** ICC = 0.80 (0.61, 0.90)			**m-CKCUEST score** SEM = 0.74 (0.59, 1.02) reps MDC_{95} = 2.06 reps B-A plots: 96.3% (−2.65, 1.09) **Muscular Endurance Index** SEM = 1.22 (0.96, 1.67) reps MDC_{95} = 3.32 reps B-A plots: 100% (−0.17, 0.14)		
			Intra-session (trial-to-trial) **Session 1** ICC = 0.90 (0.82, 0.95) **Session 2** ICC = 0.88 (0.79, 0.94)	Very good	+	**Session 1** SEM = 0.74 (0.61, 0.95) reps MDC_{95} = 2.04 reps CV% = 4.38 **Session 2** SEM = 0.79 (0.65, 1.01) reps MDC_{95} = 2.18 reps CV% = 4.21	Very good	NR
Hollstadt et al. (2020)[21] • m-CKCUEST (2)	Intersession (∼ 7 days)	ICC Spearman Rho correlation	***Inter-session*** **Number of touches** Total sample: ICC = 0.90 Male: ICC = 0.88 Female: ICC = 0.79	Inadequate	+	NR	NA	NA
Pinheiro et al. (2020)[22] • Seated Single-Arm Shot-Put Test	Intra-rater (7 days) Inter-rater (NR)	$ICC_{2,3}$ (95% CI) SEM MDC_{NR}	***Intra-rater*** **SPPT** ICC = 0.94 (0.88, 0.97) **SPPT normalized** ICC = 0.93 (0.84, 0.96)	Doubtful	+	**SPPT** SEM = 16.27 cm MDC_{NR} = 45.11 cm **SPPT normalized** SEM = 3.59 MDC_{NR} = 9.97	Adequate	?
			Inter-rater **SPPT** ICC = 0.97 (0.94, 0.99) **SPPT normalized** ICC = 0.96 (0.92, 0.98)	Doubtful	+	**SPPT** SEM = 11.64 cm MDC_{NR} = 32.29 cm **SPPT normalized**	Adequate	?

Table 2 (*Continued*)

Study/PPT	Reliability (interval between test-retest)	Type of analysis	Reliability			Measurement error		
			Result	Study quality	Rating	Result	Study quality	Rating
						SEM = 2.77 MDC_{NR} = 7.70		
Popchak et al. (2020)[30] • CKCUEST • Unilateral Seated Shot-Put Test • Repetition to Failure Assessment	Intersession (4 weeks)	$ICC_{3,1}$ (95% CI) – intrasession $ICC_{3,2}$ (95% CI) – intersession SEM MDC_{95} B-A plots	***Inter-session*** **USSPT** ICC = 0.92 (0.87, 0.95) **CKCUEST** ICC = 0.80 (−0.04, 0.94) **Repetition to Failure Assessment** Sidelying ER at 0° abduction: ICC = 0.57 (0.37, 0.72) Prone ER at 90° abduction: ICC = 0.53 (0.32, 0.69) Prone horizontal abduction at 120°: ICC = 0.48 (0.26, 0.65)	Doubtful	+ (CKCUEST and USSPT) - (Repetition to Failure Assessment)	**USSPT** SEM = 28.37 cm MDC_{95} = 78.64 cm **CKCUEST** SEM = 2.31 reps MDC_{95} = 6.40 reps **Repetition to Failure Assessment** SEM = from 4.07 to 8.87 reps MDC_{95} = from 11.28 to 24.34 reps	Very good	?
Decleve et al. (2020)[26] • Upper Limb Rotation Test	Intersession (7 days) Intrasession (45-s)	$ICC_{2,k}$ (95% CI) SEM MDC_{95}	***Inter-session*** **Dominant** ICC = 0.76 (−0.06, 0.91) **Non-dominant** ICC = 0.78 (0.54, 0.92)	Doubtful	+	**Dominant** SEM = 1.18 reps MDC_{95} = 3.27 reps **Non-dominant** SEM = 1.14 reps MDC_{95} = 3.15 reps	Very good	?
			Intra-session (trial-to-trial) **Session 1** Dominant: ICC = 0.93 (0.86, 0.96) Non-dominant: ICC = 0.96 (0.94, 0.98) **Session 2** Dominant: ICC = 0.97 (0.95, 0.98) Non-dominant: ICC = 0.97 (0.96, 0.98)	Very good	+	NR	NA	NA
Decleve et al. (2021a)[27] • m-CKCUEST (3)	Intersession (7 days) Intrasession (45-s)	$ICC_{3,1}$ (95% CI) intersession / $ICC_{2,1}$ (95% CI) intra-session SEM MDC_{95}	***Inter-session*** ICC = 0.93 (0.63, 0.97)	Doubtful	+	SEM = 1.1 reps MDC_{95} = 3.04 reps	Very good	?
			Intra-session (trial-to-trial) **Session 1** ICC = 0.89 (0.81, 0.93) **Session 2** ICC = 0.86 (0.80, 0.90)	Very good	+	NR	NA	NA

Table 2 (*Continued*)

Study/PPT	Reliability (interval between test-retest)	Type of analysis	Reliability			Measurement error		
			Result	Study quality	Rating	Result	Study quality	Rating
Decleve et al. (2021b)[28] • Shoulder Endurance Test	Intersession (7 days)	$ICC_{2,1}$ (95% CI) SEM MDC_{95}	***Inter-session*** **Dominant** ICC = 0.93 (0.86, 0.96) **Non-dominant** ICC = 0.78 (0.58, 0.89)	Doubtful	+	**Dominant** SEM = 10.7 s MDC = 29.6 s **Non-dominant** SEM = 13.8 s MDC = 38.2 s	Very good	?
Degot et al. (2021)[20] • Unilateral Seated Shot-Put Test	Intersession (7 days) Intrasession (30-s)	$ICC_{3,k}$ (95% CI) SEM MDC_{95} B-A plots CV%	***Inter-session*** **Dominant** ICC = 0.92 (0.81, 0.96) **Non-dominant** ICC = 0.93 (0.82, 0.97)	Doubtful	+	**Dominant** SEM = 3 cm/kg$^{0.35}$ MDC_{95} = 10 cm/kg$^{0.35}$ CV% = 6.23 **Non-dominant** SEM = 3 cm/kg$^{0.35}$ MDC_{95} = 9 cm/kg$^{0.35}$ CV% = 6.06	Adequate	?
			Intra-session (trial-to-trial) **Session 1** Dominant: ICC = 0.90 (0.78, 0.96) Non-dominant: ICC = 0.90 (0.78, 0.96) **Session 2** Dominant: ICC = 0.94 (0.85, 0.98) Non-dominant: ICC = 0.78 (0.55, 0.91)	Adequate	+	**Session 1** Dominant: SEM = 4 cm/kg$^{0.35}$ MDC_{95} = 13 cm/kg$^{0.35}$ CV% = 8.39 Non-dominant: SEM = 4 cm/kg$^{0.35}$ MDC_{95} = 12 cm/kg$^{0.35}$ CV% = 9.33 **Session 2** Dominant: SEM = 3 cm/kg$^{0.35}$ MDC_{95} = 10 cm/kg$^{0.35}$ CV% = 6.45 Non-dominant: SEM = 6 cm/kg$^{0.35}$ MDC_{95} = 17 cm/kg$^{0.35}$ CV% = 9.87	Adequate	?
Powell et al. (2021)[29] • Posterior Shoulder Endurance Test	Intra-rater (7 days) Inter-rater (NA)	ICC (95% CI) B-A plots SEM MDC	***Inter-rater*** **Session 1** ICC = 0.74 (0.42, 0.89) **Session 2** ICC = 0.63 (0.23, 0.83)	Doubtful	+ (session 1) - (session 2)	**Session 1** SEM = 2.79 reps MDC = 7.7 reps **Session 2** SEM = 3.31 reps MDC = 9.2 reps	Very Good	?
			Inter-session **Examiner 1** ICC = 0.84 (0.67, 0.92)	Doubtful	+	**Examiner 1** SEM = 2.11reps	Very Good	?

Table 2 (*Continued*)

Study/PPT	Reliability (interval between test-retest)	Type of analysis	Reliability			Measurement error		
			Result	Study quality	Rating	Result	Study quality	Rating
			Examiner 2 ICC = 0.84 (0.67, 0.92)			MDC = 5.8 reps **Examiner 2** SEM = 2.11reps MDC = 5.8 reps		
Draper et al. (2022)[23] • Finger Hang Test • Two-Arm Bent Hang Test • Pull-Up Shoulder Endurance Test	Intersession (7 days)	ICC 95% CI CA B-A plots CV%	***Inter-session*** **Finger Hang Test** Total sample: ICC = 0.88 (0.84, 0.92) CA = 0.94 Male: ICC = 0.89 (0.83, 0.93) CA = 0.94 Female: ICC = 0.87 (0.76, 0.93) CA = 0.93 **Two-Arm Bent Hang Test** Total sample: ICC = 0.89 (0.85, 0.93) CA = 0.94 Male: ICC = 0.86 (0.80, 0.91) CA = 0.93 Female: ICC = 0.91 (0.84, 0.96) CA = 0.96 **Pull-Up Shoulder Endurance Test** Total sample: ICC = 0.97 (0.92, 0.99) CA = 0.99 Male: ICC = 0.95 (0.86, 0.98) CA = 0.98 Female: ICC = 0.97 (0.92, 0.99) CA = 0.99	Doubtful (for the three tests)	+	**Finger Hang Test** Total sample: CV% = 18 Male: CV% = 16 Female: CV% = 24 **Two-Arm Bent Hang Test** Total sample: CV% = 15 Male: CV% = 13 Female: CV% = 19 **Pull-Up Shoulder Endurance Test** Total sample: CV% = 14 Male: CV% = 10 Female: CV% = 24	Adequate (for the three tests)	?

Abbreviations: −, insufficient; ?, indeterminate; +, sufficient; B-A, Bland-Altman plots; CA, Cronbach Alpha; CI, Confidence Interval; CKCUEST, Closed Kinetic Chain Upper Extremity Stability Test; cm, centimeters; CV%, Coefficient of Variation; ICC, Intraclass Correlation Coefficient; m, meters; m-CKCUEST, Modified CKCUEST; MDC, Minimal Detectable Change; NA, not applicable; NR, not reported; PPT, Physical Performance Test; reps, repetitions; SEM, Standard Error of Measurements.

Table 3 Results of the included studies that assessed validity.

Study/PPT	Type of validity	Outcomes	Study quality	Rating	Type of analysis	Results
Laffaye et al. (2014)[25] • Arm-Jump Board Test	Concurrent validity	Outcomes collected with a 3D accelerometer: • Velocity • Time • Index of Efficiency • Relative Power • Absolute Power	Doubtful	?	T-test Correlation analysis	T-test: non-significant differences between the Arm-Jump Board Test (distance reached) and the accelerometer (T [33] = 1.07) Correlation of the Arm-Jump Board Test (distance reached) versus: Velocity: $r = 0.43$ Time: $r = -0.12$ Index of efficiency: $r = 0.87$ Relative power: $r = 0.70$ Absolute power: $r = 0.68$ Low systematic bias = −0.88 cm or −1.25 % Low CI (−4.61 cm < 95 %CI < 2.70 cm)
Kumar et al. (2020)[24] • SMBT	Concurrent validity	Absolute peak power for the upper body during the WAnT using a modified electromagnetically braked crank-arm ergometer	Adequate	–	Pearson's correlation One sample *t*-test One-sample Wilcoxon Signed-Rank Test Linear regression	Correlation of the SMBT x WAnT in: All sportsmen: $r = 0.55$ ($p = 0.0002$) Boxers: $r = 0.5358$ ($p = 0.0011$) Freestyle wrestlers: $r = 0.4244$ ($p = 0.019$) Greco-Roman wrestlers: $r = 0.6448$ ($p = 0.012$) T-test: non-significant differences between the SMBT and the WAnT in boxers ($T = -1.90$), freestyle and greco-roman wrestlers ($T = 0.13$ and 0.69, respectively) and all sportsmen ($T = -0.33$). Wilcoxon Signed-Rank Test: non-significant differences between the SMBT and the WAnT in boxers ($p = 0.1348$), freestyle and greco-roman wrestlers ($p = 0.9354$ and

Table 3 (*Continued*)

Study/PPT	Type of validity	Outcomes	Study quality	Rating	Type of analysis	Results
						0.5089, respectively) and all sportsmen ($p = 0.7925$). Linear regression: All sportsmen: $p = 0.99$ Boxers: $p = 0.102$ Freestyle wrestlers: $p = 0.192$ Greco-Roman wrestlers: $p = 0.838$
Popchak et al. (2020)[30] • CKCUEST • USSPT • Repetition to Failure Assessment	Concurrent validity	Isokinetic strength assessments for shoulder movements of external (ER) and internal rotation (IR) at 60°/second and 180°/second using Biodex	Very good	-CKCUEST + USSPT - Repetition to Failure Assessment	Pearson r Correlation Coefficient (95% CI)	Correlation of the CKCUEST (number of touches) versus: Isokinetic ER 60°: $r = 0.57$ (0.37, 0.72) Isokinetic ER 180°: $r = 0.59$ (0.39, 0.73) Isokinetic IR 60°: $r = 0.55$ (0.34, 0.70) Isokinetic IR 180°: $r = 0.59$ (0.40, 0.73) Correlation of the USSPT versus: Isokinetic ER 180°: $r = 0.81$ (0.73, 0.86) Isokinetic IR 180°: $r = 0.74$ (0.64, 0.81) Correlation of the Repetition to Failure Assessment (number of repetitions) in sidelying ER at 0° abduction versus: Isokinetic ER 60°: $r = 0.25$ (0.07, 0.41) Isokinetic ER 180°: $r = 0.20$ (0.02, 0.37) Correlation of the Repetition to Failure Assessment (number of repetitions) in prone ER at 90° abduction versus: Isokinetic ER 60°: $r = 0.37$ (0.20, 0.51) Isokinetic ER 180°: $r = 0.38$ (0.21, 0.52)

Table 3 (*Continued*)

Study/PPT	Type of validity	Outcomes	Study quality	Rating	Type of analysis	Results
Decleve et al. (2021b)[28] • Shoulder Endurance Test	Construct validity	Shoulder isometric rotational strength for IR and ER with a hand-held dynamometer	Very good		Spearman Rank test (range)	Correlation of the Repetition to Failure Assessment (number of repetitions) in prone horizontal abduction at 120° versus: Isokinetic ER 60° : r = 0.41 (0.25, 0.55) Isokinetic ER 180° : r = 0.40 (0.23, 0.54) Correlations between Shoulder Endurance Test (time in seconds) x isometric IR and ER rotations (r = 0.309, 0.431)

Abbreviations: –, insufficient; ?, indeterminate; +, sufficient; 3D, three-dimensional; CI, confidence interval; CKCUEST, Closed Kinetic Chain Upper Extremity Stability Test; ER, external rotation; IR, internal rotation; PPT, Physical Performance Test; SMBT, Seated Medicine Ball Throw; USSPT, Unilateral Seated Shot Put Test; WAnT, Wingate Anaerobic Test.

validity. Although the methodological quality was *very good*,[9,10] the QoE was *low* due to small sample size (n = 30).

Upper limb rotation test (Fig. 2M)

One study[20] investigated the inter and intra-session reliability of this test in 91 uninjured recreational overhead athletes. Reliability was rated as *sufficient* (inter-session, ICC = 0.76–0.78; intra-session, ICC = 0.93–0.97), while the methodological quality was *doubtful*[9,10] for inter-session reliability and *very good* for intra-session reliability. The QoE was rated as *low* for inter and intra-session reliability due to small sample size. The methodological quality of inter-session SEM and the MDC_{95} was *very good*,[9,10] and *no evidence* was established for measurement error because the MIC have not been defined.

Shoulder endurance test (Fig. 2O)

One study[22] with 30 competitive overhead athletes investigated the inter-session reliability, that was rated as *sufficient* (ICC = 0.78–0.93) and the methodological quality was *doubtful*.[9,10] The QoE was rated as *low* due to small sample size and *doubtful* methodological quality. The methodological quality for SEM and the MDC was *very good*,[9,10] *no evidence* was established for measurement error because the MIC have not been defined.

The construct validity was analyzed against shoulder isometric IR and ER strength. The correlation was weak and positive (r = 0.309 to 0.431), which led to *sufficient* validity because the hypothesis was established and confirmed. The methodological quality of the study was *very good*,[9,10] and the QoE was *low* due to small sample size.

Posterior shoulder endurance test (Fig. 2P)

One study[23] with 12 elite canoeing athletes investigated the inter-rater and inter-session reliabilities. Inter-rater reliability was *sufficient* in session 1 (ICC = 0.74) and *insufficient* in session 2 (ICC = 0.63). The QoE was rated as *conflicting* due to the conflicting results and *doubtful* methodological quality.[23] Inter-session reliability was *sufficient* (ICC = 0.84) and resulted in *low*-quality evidence due to the low number of studies, small sample size, and *doubtful* methodological quality.[9,10] The methodological quality of inter-rater SEM and the MDC_{95} was *very good*[9,10] and no evidence was established for measurement error because the MIC have not been defined.

Finger hang test (Fig. 2Q)

One study[17] assessed the inter-session reliability of the Finger Hang Test in 132 rock climbers and presented *sufficient* reliability (ICC = 0.86–0.88). The methodological quality was *doubtful*[9,10] and QoE was *low*.

Two-Arm bent hang test (Fig. 2R)

One study[17] assessed the inter-session reliability of the Two-Arm Bent Hang Test in 132 rock climbers and presented *sufficient* reliability (ICC = 0.86–0.91). The methodological quality was *doubtful*[9,10] and QoE was *low*.

Pull-Up shoulder endurance test (Fig. 2S)

One study[17] reported the inter-session reliability of this test in rock climbers. Reliability was considered *sufficient* (ICC = 0.95–0.97). The methodological quality was *doubtful*[9,10] and QoE was *low*.

Discussion

This review synthesized the current evidence about the measurement properties of PPTs to assess the upper extremity of athletes. Although the reliability was considered sufficient (ICC $\geq$ 0.70) for almost all upper extremity PPTs, the evidence synthesis was downgraded in most of the cases due to small sample sizes and doubtful methodological quality of the primary studies.[9,10] The methodological quality of reliability studies was downgraded because of a lack of clarity about the knowledge of the assessor on the scores obtained in the previous session (on inter-session rehabilitation studies), as well as the absence of details about the setting that the instrument was administered (e.g. hospital, home, outpatient clinic, laboratory), and the given instructions for the test. Other factors, such as a wide range of time intervals between the test-retest measurements and the lack of information about the clinical stability of the athlete throughout the test-retest period negatively influenced the results. The CKCUEST was the only PPT that showed high and moderate QoE for intra and inter session reliability, respectively, and the Seated Single-Arm Shot-Put Test showed sufficient reliability. Tarara et al.[2] showed moderate evidence that both tests are reliable. However, evidence about the reliability of other upper extremity PPTs is still lacking, mainly those that consider the COSMIN risk of bias tool.[8,9]

Error of measurement values are important to assist clinical decision-making and interpreting studies' findings.[30] In this review, 10 studies[20–26,29] analyzed the SEM or MDC of PPTs, and most of them[20–26] presented very good methodological quality. However, they did not define the MIC, which is required to rate the results according to the COSMIN quality criteria.

Four studies assessed the validity of the upper extremity PPTs, one[19] with doubtful, one with adequate,[18] and two[22,24] with very good methodological qualities. However, due to the low number of studies and small sample sizes, the QoE, in general, was low[19,22,24] or moderate.[18] The Arm-Jump Board Test in rock-climbing[19] presented strong correlations with data from a 3D accelerometer (velocity, time, index of efficiency, and relative and absolute powers), but there was a lack of information about the measurement properties of the 3D accelerometer. Furthermore, the studies that investigated the concurrent validity of the CKCUEST and Unilateral Seated Shot-Put Test observed moderate to strong correlations of those tests with isokinetic strength of shoulder IR and ER.[24] The construct validity of the Shoulder Endurance Test was previously investigated[22] and showed weak correlations with isometric strength of shoulder IR and ER.

Tarara et al.,[4] conducted a systematic review with 11 included studies that investigated the measurement properties of 6 PPTs. They also used the Terwee Scale, COSMIN checklist, and modified GRADE approach. For comparison, our systematic review included 15 studies that assessed 13 PPTs, 6 of which were also included in that previous systematic review.[4] Furthermore, COSMIN checklist was updated in 2018 and 2020,[10,31] so the newest version was applied. Therefore, the results of this review provide an updated and detailed information about the measurement properties on PPTs.

Strength and limitation

This systematic review was conducted and reported following PRISMA guidelines.[7] The comprehensive search strategy, careful evaluation of the methodological quality according to the COSMIN,[8,9] and grading the level of evidence[12–14] provided updated information on measurement properties of the upper extremity PPTs. This study summarized and graded the level of evidence of the reliability, standard error, and validity of PPTs, which can assist clinicians in choosing a PPT according to the characteristics of the population, results of reliability and validity, and interpretation according to SEM and MDC, following an evidence-based approach. However, care should be taken because the QoE for most of the tests was very low, inconsistent, or with no evidence, and there is a lack of information about the responsiveness of all upper extremity PPTs.

Seventy percent of the sample were men, which may limit the generalizability of the findings, and more studies are needed to verify the measurement properties of PTTs in women. This systematic review focused on investigating measurement properties of the PPTs, so technology-dependent instruments, including hand-held dynamometer, isokinetic dynamometer, and 2/3-dimentional motion analysis, were not within the scope of this review. However, those instruments are important in the assessment of upper extremity and their measurement properties should be summarized by future reviews.

Implications for research

Further investigation on the inter-session and inter-rater reliability, measurement error, validity, and responsiveness of the upper extremity PPTs are still needed for enhancing the level of evidence. Using the COSMIN recommendations for planning, conducting, and reporting measurement properties studies will enhance the methodological quality of future studies.

The stability of clinical factors that could influence the scores obtained in the assessments, using scales or questionnaires (e.g., Visual Analogue Scale,[32] Global Rating of Change [−3 to +3][32]), and if the test-retest assessments were performed under similar conditions (e.g., familiarization with the test, same environment, and instructions) are important details that enhance the methodological quality of a reliability study. Also, it is recommended to report if the rater was blinded to the values obtained in the first assessment session; using appropriate time interval (i.e., 7 to 14 days) between the test-retest measurements to assure that patients were stable between the assessments and avoid recall bias. Additionally, it is recommended to randomize the limb or tests order, blinding athletes to the results until the second session is completed and including at least a sample size of 50 athletes to investigate measurement properties.

SEM and MDC values are important to the clinical decision-making, but there is a lack of these measurements for many PPTs, as well as information regarding the responsiveness and MICs of upper extremity PPTs, ideally using an anchor−based longitudinal approach (e.g. global rating of change) and longitudinal study designs.

Also, research is needed to investigate the validity against gold standard measurements or at least instruments with adequate measurement properties as a comparison. As an example, it seems important to assess the correlation of the Shoulder Endurance Test with shoulder horizontal abductors and extensors muscle strength or IR, ER, horizontal abductors, and extensors resistance using an isokinetic dynamometer.

Implications for practice

PPTs are frequently used in clinical practice to assess athletes' performance, rehabilitation progress, predict the risk of new injuries, and guide prevention and rehabilitation programs.[2,5] The results of the present review indicate that the CKCUEST and Seated Single-Arm/Unilateral Seated Shot-Put Test are reproducible to be used in clinical practice. The Seated Medicine Ball Throw is a valid test to be used to evaluate upper body power. The other tests mentioned in this review should be used with caution because the measurement properties were not sufficient to support clinical practice.

Conclusion

This systematic review identified that the CKCUEST presented sufficient inter-session and intra-session reliability, based on moderate and high-quality of evidence, respectively. The Seated Single-Arm Shot-Put Test also presented sufficient inter-session reliability, based on moderate quality of evidence. The CKCUEST, Unilateral Seated Shot-Put Test, and Repetition to Failure Assessment Test demonstrated a low level of evidence of sufficient validity and the Seated Medicine Ball Throw presented moderate quality of evidence of insufficient validity.

Funding

This study did not receive financial support.

Declaration of Competing Interest

The authors have no relevant conflicts of interests to disclose.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.bjpt.2023.100575.

References

1. Hegedus EJ, Cook CE. Return to play and physical performance tests: evidence-based, rough guess or charade? *Br J Sports Med*. 2015;49(20):1288–1289. https://doi.org/10.1136/bjsports-2015-094796.
2. Tarara DT, Fogaca LK, Taylor JB, Hegedus EJ. Clinician-friendly physical performance tests in athletes part 3: a systematic review of measurement properties and correlations to injury for tests in the upper extremity. *Br J Sports Med*. 2016;50(9):545–551. https://doi.org/10.1136/bjsports-2015-095198.
3. Borms D, Cools A. Upper-extremity functional performance tests: reference values for overhead athletes. *Int J Sports Med*. 2018;39(6):433–441. https://doi.org/10.1055/a-0573-1388.
4. Ljungqvist A, Jenoure P, Engebretsen L, et al. The international olympic committee (IOC) consensus statement on periodic health evaluation of elite athletes March 2009. *Br J Sport Med*. 2009;43(9):631–643. https://doi.org/10.1136/bjsm.2009.064394.
5. Hegedus EJ, McDonough SM, Bleakley C, Baxter D, Cook CE. Clinician-friendly lower extremity physical performance tests in athletes: a systematic review of measurement properties and correlation with injury. Part 2-the tests for the hip, thigh, foot and ankle including the star excursion balance test. *Br J Sports Med*. 2015;49(10):649–656. https://doi.org/10.1136/bjsports-2014-094341.
6. Harrison JJ, Yorgey MK, Csiernik AJ, Vogler JH, Games KE. Clinician-friendly physical performance tests for the knee. *J Athl Train*. 2017;52(11):1068–1069. https://doi.org/10.4085/1062-6050-52.11.19.
7. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:N71. https://doi.org/10.1136/bmj.n71.
8. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006.
9. 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(5):1171–1179. https://doi.org/10.1007/s11136-017-1765-4.
10. Mokkink LB, Boers M, van der Vleuten CPM, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol*. 2020;20(1):1–13. https://doi.org/10.1186/s12874-020-01179-5.
11. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012.
12. Schellingerhout JM, Heymans MW, Verhagen AP, De Vet HC, Koes BW, Terwee CB. Measurement properties of translated versions of neck-specific questionnaires: a systematic review. *BMC Med Res Methodol*. 2011;11(8). https://doi.org/10.1186/1471-2288-11-87.
13. NE D'hondt, Pool JJM, Kiers H, Terwee CB, Veeger DHEJ. Validity of clinical measurement instruments assessing scapular function: insufficient evidence to recommend any instrument for assessing scapular posture, movement, and dysfunction-A systematic review. *J Orthop Sports Phys Ther*. 2020;50(11):632–641. https://doi.org/10.2519/jospt.2020.9265.
14. Nazari G, Lu S, MacDermid JC, et al. Psychometric properties of performance-based functional tests in patients with shoulder pathologies: a systematic review and meta-analysis. *Arch Phys Med Rehabil*. 2020;101(6):1053–1063. https://doi.org/10.1016/J.APMR.2019.11.011.
15. Falsone SA, Gross MT, Guskiewicz KM, Schneider RA. One-arm hop test: reliability and effects of arm dominance. *J Orthop Sports Phys Ther*. 2002;32(3):98–103. https://doi.org/10.2519/jospt.2002.32.3.98.
16. Stockbrugger BA, Haennel RG. Validity and reliability of a medicine ball explosive power test. *J strength Cond Res*. 2001;15(4):431–438.

17. Draper N, Giles D, Taylor N, et al. Performance assessment for rock climbers: the international rock climbing research association sport-specific test battery. *Int J Sports Physiol Perform*. 2021;16(9):1242–1252. https://doi.org/10.1123/ijspp.2020-0672.
18. Kumar A, Singh RK, Apte VV, Kolekar A. Comparison between seated medicine ball throw test and Wingate test for assessing upper body peak power in elite power sports players. *Indian J Physiol Pharmacol*. 2021;64:286. https://doi.org/10.25259/IJPP_36_2021.
19. Laffaye G, Collin JM, Levernier G, Padulo J. Upper-limb power test in rock-climbing. *Int J Sports Med*. 2014;35(8):670–675. https://doi.org/10.1055/s-0033-1358473.
20. Decleve P, Attar T, Benameur T, Gaspar V, Van Cant J, Cools AM. The "upper limb rotation test": reliability and validity study of a new upper extremity physical performance test. *Phys Ther Sport*. 2020;42:118–123. https://doi.org/10.1016/j.ptsp.2020.01.009.
21. Declève P, Van Cant J, Cools AM. Reliability of the Modified CKCUEST and correlation with shoulder strength in adolescent basketball and volleyball players. *Brazilian J Phys Ther*. 2021;25(5):536–543. https://doi.org/10.1016/j.bjpt.2021.02.002.
22. Declève P, Van Cant J, Attar T, et al. The shoulder endurance test (SET): a reliability and validity and comparison study on healthy overhead athletes and sedentary adults. *Phys Ther Sport*. 2021;47:201–207. https://doi.org/10.1016/j.ptsp.2020.12.005.
23. Powell A, Levy E, Heneghan NR, Horsley I. Intra-rater reliability, inter-rater reliability and minimal detectable change of the posterior shoulder endurance test in elite athletes. *Phys Ther Sport*. 2021;49:62–67. https://doi.org/10.1016/j.ptsp.2021.02.004.
24. Popchak A, Poploski K, Patterson-Lynch B, Nigolian J, Lin A. Reliability and validity of a return to sports testing battery for the shoulder. *Phys Ther Sport*. 2021;48:1–11. https://doi.org/10.1016/j.ptsp.2020.12.003.
25. 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 Musculoskelet Disord*. 2014;15(1):1–9. https://doi.org/10.1186/1471-2474-15-1.
26. Degot M, Blache Y, Vigne G, et al. Intrarater reliability and agreement of a modified closed kinetic chain upper extremity stability test. *Phys Ther Sport*. 2019;38:44–48. https://doi.org/10.1016/j.ptsp.2019.04.017.
27. Degot M, Blache Y, Vigne G, Franger G, Neyton L, Rogowski I. Intra- and intersession reliability and agreement of the unilateral seated shot-put test outcome measures in healthy male athletes. *BMC Sports Sci Med Rehabil*. 2021;13(1):72. https://doi.org/10.1186/s13102-021-00301-4.
28. Hollstadt K, Boland M, Mulligan I. Test-retest reliability of the closed kinetic chain upper extremity stability test (CKCUEST) in a modified test position in division I collegiate basketball players. *Int J Sports Phys Ther*. 2020;15(2):203–209.
29. Pinheiro JS, Monteiro OLS, Pinheiro CAB, et al. Seated single-arm shot-put test to measure the functional performance of the upper limbs in exercise practitioners with chronic shoulder pain: a reliability study. *J Chiropr Med*. 2020;19(3):153–158. https://doi.org/10.1016/j.jcm.2020.01.001.
30. Portney LG, Watkins MP. *Foundations of Clinical Research : Applications to Practice*. Pearson/Prentice Hall; 2009.
31. Mokkink L.B., Prinsen C.A.C., Patrick D.L., et al. COSMIN methodology for systematic reviews of patient - reported outcome measures (PROMs). 2018;(February):1–78.
32. Macdermid JC, Walton DM, Avery S, et al. Measurement properties of the neck disability index: a systematic review. *J Orthop Sports Phys Ther*. 2009;39(5):400–416. https://doi.org/10.2519/jospt.2009.2930.