Sports Medicine (2022) 52:1961–1979
https://doi.org/10.1007/s40279-021-01635-2

SYSTEMATIC REVIEW

Reliability of Field-Based Fitness Tests in Adults: A Systematic Review

Magdalena Cuenca-Garcia[1,2] · Nuria Marin-Jimenez[1,2] · Alejandro Perez-Bey[1,2] · David Sánchez-Oliva[1,2,3] · Daniel Camiletti-Moiron[1,2] · Inmaculada C. Alvarez-Gallardo[1,2] · Francisco B. Ortega[4,5,6] · Jose Castro-Piñero[1,2]

Accepted: 10 December 2021 / Published online: 22 January 2022

Abstract

Background Physical fitness is a powerful predictor of morbidity and mortality, and is therefore a useful indicator for public health monitoring. To assess physical fitness, field-based tests are time-efficient, inexpensive, have minimal equipment requirements, and can be easily administered to a large number of individuals.

Objective The objective of this systematic review was to examine the reliability of existing field-based fitness tests used in adults aged 19–64 years.

Methods A systematic search of two electronic databases (MEDLINE and Web of Science) was conducted from inception to 8 June 2021 by two independent researchers. Each study was classified as high, low, or very low quality according to the description of the participants, the time interval between measurements, the description of the results, and the appropriateness of statistics. Three levels of evidence (strong, moderate, and limited) were established according to the number of studies and the consistency of their findings. The study protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO reference number, CRD42019118480).

Results Of 17,010 records identified, 129 original studies examining the reliability of field-based fitness tests in adults were considered eligible. The reliability was assessed of tests of cardiorespiratory fitness (33 studies: 30 of high quality), musculoskeletal fitness (92 studies: 78 of high quality), and motor fitness (22 studies, all of high quality). There was strong evidence indicating: (i) the high reliability of the cardiorespiratory fitness tests: 20-m shuttle run, 6-min step, and 6-min walk; (ii) the high reliability of the musculoskeletal fitness tests: handgrip strength, back-leg strength, Sorensen, trunk flexion sustained, 5-reps sit-to-stand, sit-and-reach and toe-touch, and moderate reliability bilateral side bridge and prone bridge tests; and (iii) the moderate reliability and low reliability, respectively, of the motor fitness tests *T*-test and single-leg stand. We found moderate evidence indicating the moderate or high reliability of the following tests: Chester, sit-up, partial curl-up, flexion-rotation trunk, timed stair ascent, pull-up, bent-arm hang, standing broad jump, hop sequence, trunk lift, timed-up-and-go, and hexagon agility. Evidence for the reliability of balance and gait speed tests was inconclusive. Other field-based fitness tests demonstrated limited evidence, mainly due to there being only few studies.

Conclusions This review provides an evidence-based proposal of the more reliable field-based fitness tests for adults aged 19–64 years. Our findings identified a need for more high-quality studies designed to assess the reliability of field-based tests of lower and upper body explosive and endurance muscular strength, and motor fitness (i.e., balance and gait speed tests) in adults.

✉ Nuria Marin-Jimenez
nuria.marin@uca.es

Extended author information available on the last page of the article

Key Points

Field-based fitness-test batteries have been proposed based on their reliability, criterion-validity, predictive validity, feasibility, and safety for use in preschool children, and children and adolescents (i.e., PREFIT and ALPHA-fitness test battery, respectively). However, there is no known field-based fitness-test battery with these characteristics for adults.

Reliability assessment is only the first step for the recommendation of field-based fitness tests.

Our study is a systematic review of the available evidence and provides evidence-based recommendations regarding the reliability of field-based fitness tests for use in adults aged 19–64 years.

Its contribution is to better inform researchers, clinicians, and practitioners so that they can select the best field-based physical fitness tests to be used in adults. This could be a first step towards finding a comprehensive tool to assess or predict health status through physical fitness.

1 Introduction

Physical fitness is a good indicator of the cardiometabolic health of a person of any age [1–6]. Strong and consistent evidence exists of an inverse relationship between physical fitness, especially cardiorespiratory and muscular fitness, and all-cause mortality and cardiovascular disease-related mortality [4, 7–9]. Thus, the assessment of physical fitness has become highly relevant from a clinical and public health perspective. However, this indicator is not routinely used in clinical settings to assess or predict health.

Measuring physical fitness through laboratory tests requires sophisticated and expensive equipment, is time consuming, and qualified technicians are needed to conduct these tests. Field-based tests provide a reasonable alternative, as they involve minimal equipment and are less costly. They can also be easily administered to several people at a time, thus minimizing the assessment time needed. These benefits make them a good option for routine use in different contexts by researchers, clinicians, physical education teachers, and personal health trainers.

Several field-based fitness test batteries have been proposed for use in adults [10–13] such as *Health-Related Fitness Test Battery for Adults*, *Canadian Physical Activity, Fitness and Lifestyle Appraisal* (CPAFLA) and *Adult Eurofit Fitness Test Battery*. Collectively, these include 30 field-based tests to assess the different components of physical fitness (i.e., cardiorespiratory fitness, musculoskeletal fitness, motor fitness and body composition). This wide test range makes it challenging to select appropriate tests for adults of different ages and fitness levels. Field-based fitness tests should meet several criteria related to their quality of measurements such as: (a) reliability (i.e., reproducibility of values in repeated trials on the same individual), (b) criterion-validity (i.e., output of the test correlates with the criterion measure, e.g., the gold standard), (c) predictive validity (i.e., relationship with health outcomes), (d) feasibility (i.e., degree of being conveniently done), (e) safety (i.e., number of health complications occurring during the testing procedure), and (f) responsiveness or longitudinal validity (i.e., ability of a test to detect changes over time) [14–17].

A reliable test is useful for clinical and research purposes as it offers similar results when performed on two or more occasions under the same conditions in the same individual. The reliability of a large number of field-based fitness tests has been analyzed in adult populations with different characteristics (i.e., age range, physical fitness level, or health condition). We propose, however, that it would be desirable to summarize the reliability determined so far of all existing field-based fitness tests and establish their current level of supporting evidence. The assessment of reliability is only a first step in the recommendation process for a complete tool to assess or predict health status through physical fitness. There is a clear need to better inform researchers, clinicians, and health personal trainers of the best physical fitness tests to be used in adults.

Accordingly, this study sought to systematically review studies conducted to examine the reliability of field-based fitness tests used in adults aged 19–64 years. Based on these findings, we provide an evidence-based proposal of the reliability shown by field-based fitness tests.

2 Methods

2.1 Protocol and Registration

This systematic review was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [18]. The review protocol was registered in the International Prospective Register of Systematic Reviews (PROSPERO reference number, CRD42019118480) [19].

2.2 Data Sources and Search Strategy

A systematic search was performed of the electronic databases MEDLINE via PubMed and Web of Science from database inception to 8 June 2021. The search terms used were those related to the following topics: (i) *Participants*: adult population (aged 19–64 years), (ii) *Reliability*: reliability, repeatability, reproducibility, consistency and stability, and (iii) *Physical fitness components*: physical fitness; muscular strength; range of motion, articular; postural balance; physical endurance; cardiorespiratory fitness, cardiovascular fitness, aerobic fitness, aerobic capacity, maximal oxygen consumption, VO_{2max}; motor fitness; running speed; agility. The specific names of the fitness-test batteries in adults were also considered. The three search topics were combined with the Boolean operators '*AND*' (inter topics) and '*OR*' (intra topics). When using PubMed, we included Medical Subject Heading (MeSH) terms to enhance the power of the search. The same search strategy and combination of terms was repeated in Web of Science but without using MeSH terms or their equivalent as a similar option does not exist for this database. The complete search strategy and search terms used for each database are provided in the Electronic Supplementary Material (ESM) Section 1.

2.3 Eligibility Criteria

Two researchers (MCG and DSO) independently assessed the eligibility of studies. Inclusion criteria were (i) Study design: original studies, reviews, and meta-analyses. (ii) Topic: studies examining the reliability of field-based fitness tests. (iii) Language: full reports published in English or Spanish. (iv) Age: adults aged 19–64 years. During the review, combined populations were observed (i.e., adults and older adults, adults and adolescents). In these cases, we noted whether these studies performed stratified analyzes by age groups, isolating the adult population from the rest. If so, the study was included and information concerning the adult population reported. In contrast, when the authors analyzed the whole sample together, we only included a study if participant ages were predominantly within our established age range. (v) Participants: generally healthy adults with no significant health problems, diseases (e.g., heart failure, chronic obstructive pulmonary disease, cancer, or neurologic diseases) or conditions (e.g., cognitive impairment or intellectual disability, mobility problems or any injury). Studies conducted in professional athletes or designed for exclusive use in clinical settings were excluded.

2.4 Study Selection

The study selection process was conducted in three stages. In the first stage, records were identified through the PubMed and Web of Science databases and imported into MENDELEY software (version Desktop 1.19.4). Next, duplicate files were removed, firstly automatically by the software and secondly by visual checking. In the second stage, titles and abstracts of the search results were examined for eligibility. Finally, seemingly eligible full-text reports were read for their final inclusion or exclusion in the review. The outcomes of all stages were compared between two researchers (MCG and DSO). When there was no consensus between the two researchers (< 5%), a third researcher (APB) made the final decision about inclusion. Reasons for exclusion of identified articles were recorded. Additionally, the reference lists of retrieved studies were examined to identify further articles. The authors were not blinded to the articles selected, as the reviewers who performed the quality assessment were familiar with the literature. Figure 1 shows the flow diagram of the study selection process.

2.5 Data Collection Process

Data extraction was independently double-checked by two researchers (APB and DSO). The extracted data from the included studies were (i) study (i.e., first author's name and year of publication); (ii) sample size used for the analysis and characteristics of the study sample (i.e., sex, age, and health condition); (iii) physical fitness test; (iv) study design (i.e., time interval between measurements); (v) statistical methods; and (vi) main outcome and conclusions.

2.6 Risk of Bias Assessment

A meta-analysis was not conducted due to the heterogeneity of statistical methods employed in the original studies, the large number of tests, and the limited number of studies per test. The risk of bias was assessed by two independent researchers (APB and DSO). Overall agreement between the two reviewers was 90% (kappa = 0.76). Disagreements were resolved by consensus in a meeting with a third researcher (NMJ). Risks of bias of the original studies were determined according to a quality assessment list for reliability studies [20] based on four criteria: (i) adequate description of the participants; (ii) adequate description of the time interval between measurements; (iii) adequate description of the results; and (iv) appropriateness of statistics. Each criterion was scored from 0 to 2 (where 2 is the best score). The scores of each criterion were summarized (0–8) for all studies. Studies were then categorized as being of very low quality (score < 2), low quality (score 2–5), or high quality (score ≥ 6) (see ESM Table 1).

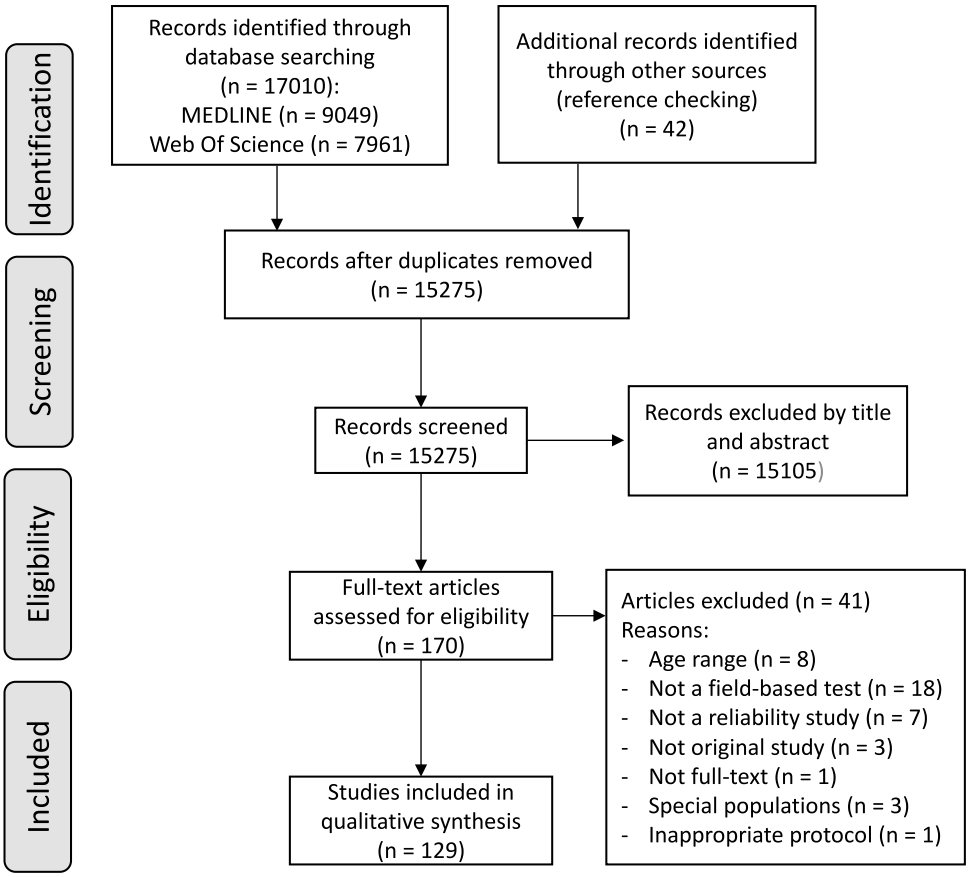

Fig. 1 PRISMA flowchart showing the selection of original studies

2.7 Level of Evidence

To make our final conclusions on the reliability of the fitness tests identified, three levels of evidence were established [21]: (i) strong evidence: consistent findings in three or more high-quality studies; (ii) moderate evidence: consistent findings in two high-quality studies; and (iii) limited evidence: consistent findings in multiple low-quality studies, inconsistent results in multiple high-quality studies, or results based on a single study. When we found strong or moderate evidence that a field-based fitness test was or was not reliable, we discussed its level of reliability.

2.8 Statistical Analysis

Reliability refers to the consistency of measurements in repeated trials on the same individual, or the absence of measurement errors. Better reliability implies the better precision of single measurements [22]. The main components of variability between repeated test results are systematic error (due to learning, training, or fatigue effects or loss of motivation) and random error (due to inherent biological or mechanical variation, or inconsistencies in the measurement protocol) [23, 24]. Several statistical methods have been used to assess inter-rater (several observers examining the same participant) and intra-rater (same observer, also referred to as *test–retest*) reliability and agreement. Coefficients like Pearson's correlation (r) and intraclass correlation coefficient (ICC) are commonly used to describe *relative reliability* (i.e., the consistency of measurements on individuals in a group relative to others). Pearson's correlation coefficient depends greatly on the range of values of the sample. A high correlation between repeated scores reflects well the stability of position or rank order within a particular sample; however, it does not detect systematic errors. Thus, we need to be cautious when comparing test and retest correlations between reliability studies or extrapolating the results to other sample groups with different characteristics [22]. A high ($r \geq 0.8$) and statistically significant correlation coefficient is deemed to indicate a high degree of correlation (i.e., high reliability) [24, 25]. The ICC reflects both systematic and random errors, is considered an appropriate measure of agreement, and is often reported in place of Pearson's correlation; however, the ICC is affected by between-subjects variability. In this review, an ICC < 0.8 was considered to indicate low reliability, one between 0.8 and 0.9 moderate reliability, and one > 0.9 high reliability [26]. However, neither Pearson's correlation nor ICC offer information on the heteroscedasticity of the sample. The methods used to describe *absolute reliability* (consistency of

repeated measurements for individuals) include the standard error of measurements (SEM), coefficient of variation (CV), Bland and Altman's mean difference (also called systematic bias) and ± 95% limits of agreement (mean of the differences ± 1.96 × standard deviation), paired t-test and analysis of variance [25]. The SEM quantifies the precision of individual scores in a test. It is not affected by between-subjects variability (i.e., is considered a fixed characteristic of any measure, regardless of the sample of subjects under investigation), and is useful in the absence of heteroscedasticity [27]. It is expressed in standardized values (as a percentage of the mean value of the measurements (SEM = mean of the difference scores between two trials × 100/mean of the first trial), a value ≤ 15% is here considered acceptable) or unstandardized values (the lower the value, the greater the reliability) [24, 27]. The CV method assumes that greatest test–retest variation occurs in individuals scoring the highest values in the test (useful in the presence of heteroscedasticity) and provides useful information about the random error of measure (expressed as a percentage, ≤ 10% is here considered acceptable) [24, 27]. Paired statistical or analysis of variance for repeated measures is useful for detecting mean differences (i.e., systematic errors) in reliability studies. The Bland and Altman's method and limits of agreement have been widely used to evaluate reliability for within-subject variation, taking into account both systematic and random errors and helping identify the presence of heteroscedasticity [22]. The reliability of nominal and ordinal unpaired data is often analyzed with the Cohen or the weighted kappa (κ) coefficient (> 0.9 is here considered acceptable/high agreement) or the percent agreement statistic (> 80% is here considered acceptable/high agreement) [28]. Any reliability study should not be based on a single statistic method. Acceptable reliability should be based on statistical methods to describe absolute and relative reliability to help overcome possible limitations (i.e., relative and absolute reliability methods to compare test reliability across the different studies). In this review, ICC was the statistical method most frequently used to assess relative reliability. The SEM is not affected by between-subjects variability as the ICC is. To compare test reliability across the different studies, whenever possible we calculated SEM expressed as a percentage of the mean. High ICC values and low SEM values suggest high levels of reliability and reproducibility, regardless of the characteristics of individuals.

3 Results

In our electronic search, 17,010 records were retrieved and 42 additional records identified through other sources (i.e., reference checking). After removing duplicates, 15,275 articles were identified, of which 15,105 were excluded after reading the title and abstract, and 41 articles after reading the full text. Reasons for exclusion were: (i) age range ($n = 8$); (ii) not a field-based test ($n = 18$); (iii) not a reliability study ($n = 7$); (iv) not an original study ($n = 3$); (v) not a full test ($n = 1$); (vi) special populations ($n = 3$); and (vii) inappropriate protocol ($n = 1$). Finally, 129 studies were considered eligible for our systematic review (see PRISMA flowchart in Fig. 1). Of these 129 studies, 114 (88%) were classified as 'high quality' and 16 (12%) as 'low quality'; no article was classified as 'very low quality' (see ESM Table 2). The sample sizes of the individual studies ranged from 13 to 260. The time interval between tests ranged from 1 to 54 days, but in most it was 1 or 2 weeks, except for those examining intra-session test–retest data.

Cardiorespiratory fitness was analyzed in 33 studies, of which 30 were classified as high quality. The most common field-based fitness tests for assessing cardiorespiratory fitness were the 20-m shuttle run test (ten studies, nine of them classified as high quality), the 6-min walk test (four studies, all of high quality) and the 6-min step test (four studies, all of high quality). The reliability of field-based fitness tests assessing musculoskeletal fitness was investigated in 92 studies, 78 of which were classified as high quality. The most common field-based fitness tests for musculoskeletal fitness were the handgrip strength test (23 studies, 22 of high quality) to assess isometric strength; the Sorensen test (11 studies, all of high quality), the partial curl-up test (six studies, five of high quality), the trunk flexion sustained test (11 studies, all of high quality), the push-ups test (five studies, four of high quality), and the sit-to-stand test (eight studies, seven of high quality) to assess endurance strength; and the sit-and-reach test (19 studies, 13 of high quality) and toe-touch test (ten studies, eight of high quality) to measure flexibility. The reliability of field-based fitness tests assessing motor fitness was investigated in 22 studies, all classified as high quality. The most common field-based tests for motor fitness were the timed up-and-go test (three studies) and the *T*-test (three studies) (see ESM Table 3). ICC was the statistical method most frequently used to assess relative reliability. To assess absolute reliability, SEM, CV, analysis of variance, and the Bland–Altman method were employed (see ESM Table 3).

4 Discussion

4.1 Cardiorespiratory Fitness

Cardiorespiratory fitness reflects the overall capacity of the cardiovascular and respiratory systems to supply oxygen during maintained physical activity [29]. It has been traditionally assessed through maximal (e.g., 20-m shuttle run test) and submaximal (e.g., 2-km walk, 6-min walk, or 6-min step

tests) field-based fitness tests. The reliability of the 20-m shuttle run test was investigated in nine high-quality studies [30–38]. In seven of these studies, the 20-m shuttle run test demonstrated high test–retest reliability based on coefficients of correlation (ICC = 0.93–0.96 [31, 33, 34]; r = 0.85–96 [30, 33–35, 37, 38]) in adults aged 18–45 years. Caution is required when comparing reported coefficients between different samples (mainly of different ages and fitness levels). In four of these studies, SEM was calculated showing the low dispersion of measurement errors between test–retest (all SEMs < 15% [31, 33–35]). In one study conducted in active males of mean age 21.8 years, the 20-m shuttle run test returned no significant mean test–retest differences ($p = 0.90$), and a mean difference between test–retest (i.e., bias) and limits of agreement of 0.4 ± 2.7 mlO_2/kg/min [32]. These findings are in agreement with those of other studies performed in young adults [33, 36]. One of these studies detected a mean difference and limits of agreements between test–retest of 0.6 ± 6.8 mlO_2/kg/min [36]. In the other study, the 20-m shuttle run test was performed three times in active young adults, and the mean difference and limits of agreement were -1.1 ± 4.7 mlO_2/kg/min between sessions 1 and 2, and 0.0 ± 5.5 mlO_2/kg/min between sessions 2 and 3 [33]. These studies concluded that the 20-m shuttle run test offers high reliability with reasonably narrow limits of agreement and consistent results between test–retest in young adults (i.e., no significant mean test–retest differences, $ps > 0.05$). However, another study performed in men of mean age 34.8 years showed (despite an ICC ≥ 0.95) a lack of significant differences between test–retest ($p \leq 0.05$), and a mean difference between test–retest and limits of agreements of 0.8 ± 3.1 mlO_2/kg/min [34]. Other studies analyzed the reliability of different versions of the 20-m shuttle run test (i.e., 20-m square shuttle run test [36], 15-m square shuttle run test [39], 15-m shuttle walk/run test [40] and 10-m shuttle walk test [41, 42]). The 20-m and 15-m square shuttle run tests showed high reliability in adults aged 18–29 years, and could be a good option to reduce the test's turning angle from 180° to 90° ($r \geq 0.78$, $p > 0.05$ between test–retest, CV ≤ 5.2%). The 15-m shuttle walk/run and the10-m shuttle walk tests demonstrated high reliability in adults aged 40–59 years (ICCs ≥ 0.93, $p > 0.05$ between test–retest). These tests could be an alternative to the 20-m shuttle run test in adults whose fitness level is low or who find it difficult to run (e.g., obese individuals or those with low back pain).

The reliability of the *6-min walk test* was examined in four high-quality studies [43–46]. The test–retest reliability of the 6-min walk test was considered high based on coefficients of correlation (ICCs = 0.93–0.96 [45, 46], $r = 0.90$ [43]) in three of these studies, and moderate in one (ICC = 0.82 [44]) in adults aged 18–64 years. The 6-min walk test showed the low dispersion of measurement errors between test–retest (all SEMs ≤ 3.2% [45, 46]). In one of these studies performed in obese adults aged 21–62 years, in addition to an ICC of 0.96 and considering mean significant difference between test–retest ($p < 0.001$), the 6-min walk test showed a mean test–retest difference of 18 m, limits of agreement between − 46 m and 80 m, and a CV of 4.7% [46]. Thus, the 6-min walk test demonstrated good reproducibility in obese individuals based on a low CV and high ICC. The reliability of the *5-min walk test,* a modified version of the 6-min walk test, was analyzed in 44 participants with low back pain and 48 controls aged 21–63 years. Test–retest reliability was moderate in the low back pain group (ICC = 0.87) and low in the control group (ICC = 0.60). However, SEMs were low in both groups (≤ 12%) [47]. Other field-based fitness tests for assessing cardiorespiratory fitness based on walking/running in a given time or space used in reliability studies were: 2-km walk [48], 400-m walk [49], 1.5-mile test (walking or running) [50], Université-Montréal [51] and Cooper's 12-min run [37].

The reliability of the *6-min step test* was examined in four high-quality studies [52–55]. In all of them, the 6-min step test showed high test–retest reliability based on ICCs ranging from 0.90 to 0.97 [52–55]) in healthy adults [52–54] and adults at risk of cardiovascular disease [55] aged 20–64 years. The 6-min step test revealed the low dispersion of measurement errors between test–retest (all SEMs ≤ 2.8% [52, 54, 55]). The reliability of the *Chester step test* was investigated in two studies conducted in adults aged 19–52 years [56, 57]. These studies analyzed reliability through the Bland–Altman method and showed mean differences between test–retest and limits of agreement of 0.8 ± 3.7 mlO_2/kg/min [56] and -0.7 ± 4.5 mlO_2/kg/min (mean significant difference between test–retest $p > 0.05$) [57]. In both studies, the Chester step test revealed the low dispersion of measurement errors between test–retest (all SEMs < 3%), indicating that this test had a high test–retest reliability. Other field-based tests for assessing cardiorespiratory fitness based on steps used in reliability studies were the 2-level step [58] and the OSU step [59].

Levels of Evidence: There was strong evidence indicating that: (i) the 20-m shuttle run test demonstrates high test–retest reliability in young adults; and (ii) the 6-min walk and 6-min step submaximal tests show high test–retest reliability in adults aged 18–64 years. These submaximal tests might be an alternative in adults with a low level of physical fitness or difficulty in running (e.g., obese individuals). There was moderate evidence indicating that the Chester step test provides results with high test–retest reliability in adults aged 19–52 years. Due to the low number of studies (a single study), there was limited evidence indicating that: (i) the 400-m walk and 2-km walk tests produce results with high test–retest reliability in adults aged 30–64 years

Table 1 Levels of evidence of cardiorespiratory fitness tests

Field-based fitness test	Strong	Moderate	Limited
Cardiorespiratory fitness tests			
Shuttle run tests			
20-m shuttle run	●		
20-m square shuttle run			○
15-m square shuttle run			○
15-m shuttle walk/run			○
10-m shuttle walk			○
Distance and time-based run/walk tests			
6-min walk	●		
5-min walk			○
1.5-mile run/walk			●
2-km walk			●
400-m walk			●
University Montreal			●
Cooper's 12-min run			●
Step tests			
6-min step	●		
Chester step		●	
2-level step			●
OSU step			○

●=indicates high reliability
◐=indicates moderate reliability
○=indicates low/null reliability
▒=indicates inconclusive reliability

and could be an option to assess cardiorespiratory fitness in adults with a low level of fitness or difficulty in running; (ii) the 1.5-mile test (walking or running) demonstrates high test–retest reliability in young adults aged 19–26 years; (iii) the Cooper's 12-min run test produces results with high reliability in young adults, and the Université de Montréal Track test does so in middle-age adults; (iv) the 2-level step test provides highly reliable results in adults aged 55–64 years; and (v) the OSU step test shows low reliability in young adults (see Table 1).

4.2 Musculoskeletal Fitness

Musculoskeletal fitness is defined as the ability of a specific muscle or muscle group to generate force (muscular strength) to resist repeated contraction over time, to maintain a maximal voluntary contraction for a prolonged period of time (muscular endurance), or to carry out a maximal/dynamic contraction in a short period of time (explosive strength) [29]. Flexibility is defined as the ability of a specific muscle or muscle group to move freely through a full range of motion [29].

4.2.1 Maximal Isometric Strength

The reliability of the *handgrip strength* test was analyzed in 22 high-quality studies in adults aged 18–64 years [31, 44, 60–79]. Most of them examined handgrip strength-test reliability using the JAMAR dynamometer as the measurement tool (14 studies, aged range 18–64 years). In all of these studies, the handgrip strength test demonstrated high test–retest reliability according to coefficients of correlation (ICCs = 0.90–0.99 [60–62, 66, 67, 70, 71, 73–79], $r = 0.80$ [66]). The test–retest reliability of the handgrip strength test using other models of dynamometers (i.e., analog and digital TKK [68], DynEx [76], BTE-Primus [75, 80], Grippit [78], Rolyan [79], Lode [63], Smedle [44], MicroFET 4 [69], Lafayette [64], and Grip-ball [65]) was investigated showing a high reliability based on ICCs > 0.90. While only a few studies have addressed the reliability of the TKK dynamometer, this method has several benefits over JAMAR, for example: (i) the grip span of the TKK dynamometer can be continuously adjusted for differences in hand size using age- and sex-specific equations, whereas JAMAR has five positions [81–84]; and (ii) the TKK dynamometer does not need regular calibration, while JAMAR requires calibration every year.

The reliability of the *back-leg strength test* (using as instruments the Takei and the Baseline back-leg-chest dynamometers) was examined in three high-quality studies [85–87]. In all of them, the back-leg strength test showed high test–retest reliability based on coefficients of correlation (all $rs > 0.8$ [86] and ICCs > 0.9 [85, 87]) and acceptable CV (< 10%) [87] in adults aged 19–46 years. In addition,

in one of these studies, the back-leg strength test returned no significant test–retest differences ($p > 0.05$) [86]. The SEMs observed in these studies were low (all < 15%), thus reflecting the low dispersion of measurement errors between test–retest.

Levels of Evidence: There was strong evidence indicating that: (i) irrespective of the type of dynamometer used, the handgrip strength test shows high test–retest reliability in adults aged 18–64 years; and (ii) the back-leg strength test provides results with high test–retest reliability in adults aged 19–46 years (see Table 2).

4.2.2 Endurance Strength (Dynamic or Isometric)

4.2.2.1 Trunk Endurance Strength Numerous tests have been proposed to measure trunk muscle performance. Many

Table 2 Levels of evidence of musculoskeletal and motor fitness tests

Field-based fitness test	Strong	Moderate	Limited
Musculoskeletal fitness tests			
Maximal isometric strength			
Handgrip strength (Jamar)	●		
Handgrip strength (TKK)			●
Handgrip strength (DynEx)			●
Handgrip strength (BTE Primus)		●	
Handgrip strength (Grippit)			●
Handgrip strength (Royal)			●
Handgrip strength (Smedley)			●
Handgrip strength (Lode)			●
Handgrip strength (MicroFET 4)			●
Handgrip strength (Grip-Ball)			●
Handgrip strength (Lafayette)			●
Back-leg strength (Baseline, Takey)	●		
Endurance strength (dynamic or isometric)			
Trunk endurance strength			
Original/Modified Sorensen	●		
Bilateral side bridge	◐		
Prone bridge	◐		
Sit-up		◐	
Original/Modified partial curl-up		◐	
Original/Modified trunk flexion sustained	●		
Flexion-rotation trunk		◐	
Lower body endurance strength			
5-reps sit-to-stand	●		
10-reps sit-to-stand			●
30-s sit-to-stand			●
60-s sit-to-stand			◐
Original/Modified timed stair ascent		●	
Modified step-up			●
Single-leg squats (isometric/dynamic)			◐
Upper body endurance strength			
Original/Modified push-up			■
Original/Modified pull-up		●	
Original/Modified bent-arm hang		◐	
Lift and reach task			■
Explosive strength			
Lower body explosive strength			
Standing broad jump		◐	
Countermovement vertical jump			◐
Jump and reach			●
Hop sequence		◐	
Figure-eight hop			●
Up-and-down hop			●
Side-to-side hop			◐
Upper-Trunk body explosive strength			

Table 2 (continued)

Modified medicine ball put				●
Front/side abdominal power				●
Flexibility				
Hamstring-low back flexibility				
Original/Modifications sit-and-reach	●			
Toe-touch	●			
Trunk lift		●		
Side-bending				■
Shoulder-neck mobility				○
Hand behind back				●
Motor fitness tests				
Static Balance				
Original/Modifications single-leg stand (open/closed eyes)	○			
Parallel, semi-tandem and tandem stand (open/closed eyes)				■
Dynamic Balance				
Star excursion balance				■
Tandem walking				◐
Functional balance				◐
Gait speed - Agility				
Timed up-and-go		●		
Original/Modified *T*-Test	◐			
Original/Modified hexagon agility		◐		
Original/Modifications gait speed				■
4-square step				◐
10x5 shuttle run				◐
Zig-zag run				●
Edgren side step				○
Illinois agility				○
Pro-agility				●
Agility track				●
Plate tapping				○

●=indicates high reliability
◐=indicates moderate reliability
○=indicates low/null reliability
■=indicates inconclusive reliability

of these tests are isometric trunk holding tests (i.e., the Sorensen, bridge and trunk flexion sustained tests) used to measure the endurance capacity and fatigability of the trunk muscles. Moreover, dynamic trunk tests such as full-range sit-up (mainly foot held), curl-up or partial curl-up (mainly foot free) or flexion-rotation trunk tests have been proposed.

The reliability of the *Sorensen test* to assess the isometric endurance trunk extensor muscles was analyzed in 11 high-quality studies in adults aged 18–64 years [47, 88–97]; four of them examined a modified version of the Sorensen test with the participant lying on the floor and held by a partner [90–92, 95]. In healthy adults, the Sorensen test based on coefficients of correlation demonstrated a high reliability in three studies (ICCs = 0.91–0.93 [47, 92], $r = 0.97$ [90]); moderate reliability in three (ICCs = 0.80–0.85 [88, 95, 96]) and low reliability in one ($r = 0.74$ [94]). In adults with low back pain, the Sorensen test provided results with a high reliability based on coefficients of correlation in four studies (ICCs = 0.91–0.96 [47, 88, 91], $r = 0.97$ [90]), moderate reliability in two (ICCs = 0.88 [93, 96]) and low reliability in one (ICC = 0.59 [97]. In another reliability study utilizing the Bland–Altman method in adults with low back pain aged 30–58 years, the Sorensen test showed no systematic significant differences between sessions ($p > 0.05$), a mean difference between test–retest of 0.15 s and limits of agreement between − 1.61 and 1.93 s [89]. Overall, the SEMs observed in these studies were low (all < 15% [47, 91, 94–97]), thus

reflecting the low dispersion of measurement errors between test–retest.

The reliability of the *prone and bilateral side bridge tests* was analyzed in three and four high-quality studies, respectively [87, 93, 95, 98–101]. The test–retest reliability of the prone bridge test (supported by the elbow and foot) was considered moderate to high, with ICCs ranging from 0.86 to 0.95 in three studies conducted in adults aged 19–45 years [87, 100, 101]. The bilateral side bridge test demonstrated a high test–retest reliability (ICCs > 0.90) in one study in adults with low back pain [93] and moderate to high test–retest reliability (ICCs = 0.80–0.91) in three studies in healthy adults aged 18–57 years [95, 98, 99]. The prone and bilateral side bridge tests showed the low dispersion of measurement errors between test–retest based on low SEMs (all < 15% [87, 93, 95, 98, 99, 101]). The bridge test requires minimal inexpensive equipment and is simple to implement in different settings and in a wide range of populations according to its easy held isometric position.

The reliability of the sit-up test was analyzed in two high-quality studies [31, 93], the partial curl-up test in five high-quality studies [93, 94, 102–104], the trunk flexion sustained test in 11 high-quality studies [89–94, 98, 100, 103, 105, 106] and the flexion-rotation trunk test in two high-quality studies [107, 108]. The *sit-up test* showed moderate test–retest reliability based on ICCs of 0.83 and the low dispersion of measurement errors between test–retest (SEMs < 5%) in adults aged 18–55 years [31, 93]. The test–retest reliability of the *partial curl-up test* was considered moderate to high based on coefficients of correlation (ICCs = 0.89–0.98 [93, 102, 103], $r = 0.93$ [104]) in adults aged 18–57 years. The dispersion of measurement errors between test–retest was low (SEMs < 15% [93, 102, 103]). Another study performed in adults aged 35–44 years investigated the test–retest and inter-rater reliability of a modified version of the partial curl-up test (i.e., feet without support), finding low levels of agreement (Kappa's ≤ 0.78) [94].

The reliability of the *trunk flexion sustained test* was analyzed in healthy adults [92, 94, 98, 100, 103, 105, 106] and adults with low back pain [89–91, 93] aged 18–59 years. In healthy adults aged 18–57 years, the trunk flexion sustained test demonstrated high test–retest reliability in four studies according to coefficients of correlation (ICCs = 0.93–0.95 [92, 98], *r*s = 0.93–0.95 [90, 94]) and low test–retest reliability in three (ICCs = 0.51–0.71 [100, 103], r = 0.71 [106]). In a further study, the trunk flexion sustained test showed a mean test–retest difference of 7.9 s (95% CI − 5.7 to 21.5) and a low CV (3.7%); but with low inter-rater reliability (ICC = 0.76) and high dispersion of inter-rater measurement errors (SEM = 19%) [105]. In adults with low back pain aged 18–58 years, the trunk flexion sustained test returned results with high test–retest reliability in three studies (two of which using a modified version of this test, i.e., lying on the floor with legs elevated and arms folded across the chest) based on ICCs ranging from 0.91 to 0.97 or Pearson's coefficients of 0.91 [90, 91, 93], and the low dispersion of measurement errors between test–retest based on SEMs < 15% [91, 93]. In an additional study, the trunk flexion sustained test showed no significant mean test–retest differences ($p > 0.05$), an absolute mean test–retest difference of − 0.1 s, and limits of agreement between − 3.4 and 3.2 s (for inter-rater reliability) [89].

Finally, the *flexion-rotation trunk test* used to assess trunk flexor-rotator dynamic endurance was analyzed in two studies in young adults aged 19–27 years [107, 108]. The flexion-rotation trunk test provided results with moderate test–retest reliability based on ICCs ≥ 0.83 and SEMs ≤ 15% [107, 108].

Levels of Evidence: There was strong evidence indicating that: (i) the Sorensen test and its modified versions and (ii) the trunk flexion sustained test produce results with high test–retest reliability in healthy adults and adults with low back pain aged 18–64 years; and (iii) the prone and bilateral side bridge tests show moderate test–retest reliability in adults aged 18–64 years. There was moderate evidence indicating that: (i) the sit-up and partial curl-up tests offer results with moderate to high test–retest reliability in adults aged 18–57 years; and (ii) the flexion-rotation trunk test provides results with moderate test–retest reliability in adults aged 19–37 years. However, a prolonged familiarization period is needed before testing (at least three practice trials) (see Table 2).

4.2.2.2 Lower Body Endurance Strength Lower body endurance strength has been traditionally assessed using sit-to-stand and timed stair-ascent tests. The reliability of the *sit-to-stand test* was determined in seven high-quality studies [43, 45, 47, 58, 109–111]. Results of different versions of the sit-to-stand test were expressed as maximum repetitions executed in 60 s [58] and 30 s [111], and time spent on 5-reps [43, 45, 47, 109, 110] or 10-reps [43] in healthy adults and adults with low back pain [47]. The 5-reps sit-to-stand test showed moderate to high test–retest reliability based on coefficients of correlations in three studies (ICCs = 0.83–0.94 [45, 110], $r = 0.80$ [43]) in healthy adults aged 18–64 years. Overall, the SEMs observed in these studies were low (all < 15% [45, 47, 110]). However, one study showed a low test–retest reliability of the 5-reps sit-to-stand test in healthy adults aged 50–64 years (ICC = 0.72 [109]) and another one in adults with low back pain (ICC = 0.45 [47]) aged 21–63 years.

The reliability of the *timed stair ascent (12 steps) test* was analyzed in three high-quality studies [45, 89, 112]. In one study conducted in adults aged 18–43 years, the timed stair ascent test revealed high reliability (ICC = 0.90) [45]. In another study in adults of mean age 38.7 years, the timed

stair ascent test provided low and high reliability in unloaded and loaded versions, respectively (ICC = 0.79 and 0.94). Moreover, mean test–retest differences were 0.17 s and limits of agreement were between − 0.68 and 1.04 s in the unloaded version [0.21 s (− 0.25 and 0.66 s, for the loaded version)], indicating excellent agreement between tests [112]. A modified version test (18–20 steps) yielded high reliability in adults with back/neck pain aged 30–58 years and healthy adults aged 27–59 years, with no significant differences between test sessions ($p > 0.05$), a mean test–retest difference of 0.30 s, and limits of agreement of − 3.44 to 4.04 s [89]. Other lower body endurance strength field tests used in reliability studies were: single-leg squat (timed isometric wall squat (hip/knee 90°) [87], dynamic squat knee flex 90° [105] or 60° [93]) and step-up [95].

Levels of Evidence: There was strong evidence indicating that the 5-reps sit-to-stand test produces results with moderate to high test–retest reliability in adults aged 19–64 years. There was moderate evidence indicating that the timed stair ascent test offers high reliability in adults aged 18–58 years. However, comparisons between different testing protocols should be interpreted with caution. Due to the low number of studies, there was limited evidence indicating that: (i) the 30-s sit-to-stand test provides results with high reliability in healthy adults aged 18–55 years and the 60-s sit-to-stand test shows moderate reliability in adults with low back pain aged 55–64 years; (ii) the modified step-up test (20 kg load) produces high test–retest and inter-rater reliability in adults aged 19–46 years; and (iii) the single-leg squat test in isometric (i.e., timed wall squat hip/knee flex 90°) and dynamic positions (squat knee flex 90° or 60°) generates results with moderate reliability in adults aged 19–59 years. However, the timed isometric wall single-leg squat could be the safest test (see Table 2).

4.2.2.3 Upper Body Endurance Strength Upper body endurance strength has been traditionally assessed with push-up, pull-up, and bent-arm hang tests. The reliability of the *push-up test* was analyzed in four high-quality studies [87, 105, 113, 114]. The push-up test demonstrated moderate to high test–retest reliability in two studies in adults aged 18–45 years (ICCs ≥ 0.87 [87, 114]) and low test–retest reliability in one study in young adults of mean age 21.2 years (ICCs = 0.25–0.52, Kappa = 0.14–0.52 [113]). A modified version of the push-up test provided results with moderate inter-rater reliability (ICC = 0.88) and low test–retest reliability (a mean test–retest difference of 3 repetitions with a CI between 2.1 and 3.9 repetitions and SEM of 22%) in adults aged 25–59 years [105].

The reliability of the *pull-up test* and its modified version (i.e., in horizontal position) was analyzed in two high-quality studies showing a high test–retest reliability based on ICCs ranging from 0.94 to 0.98 and the low dispersion of measurement errors between test–retest based on low SEMs (< 15%) [95, 114] in adults aged 19–45 years. The reliability of the modified version of the *bent-arm hang test* (i.e., timed until 90° of elbow extension) was investigated in two high-quality studies. In adults aged 19–36 years, the bent-arm hang test provided results with moderate to high test–retest reliability (ICCs = 0.89–0.99 [31, 115]; SEM = 8.5% [31]). Finally, the reliability of the *lift and lower task test* (as a measure of upper body function, consisting of lifting a weighted basket) was investigated in two high-quality studies [58, 112]. One of them showed low reliability based on an ICC of 0.66 in adults aged 55–65 years [58], and the other generated results with a high reliability based on an ICC of 0.93 in men of an average age of 37.8 years [112].

Levels of Evidence: There was moderate evidence indicating that: (i) the pull-up test and its modified version show high reliability in adults aged 18–45 years; and (ii) the bent-arm hang test and its modified version produce results with moderate and high reliability in adults aged 19–45 years. However, we need to be cautious when comparing the reliability of an original test with that of its modified versions. Finally, there was inconclusive evidence about the reliability of the push-up test and its modified versions and the lift and reach task test (see Table 2). The pull-up test and the bent-arm hang test are highly influenced by body weight. An alternative test to assess upper body endurance strength in individuals with overweight or even a low physical fitness level could be the push-up test or the modified pull-up test (i.e., horizontal). In addition, the bent-knee push-up test could be used instead of the full-body push-up test for adults with a low physical fitness level.

4.2.3 Explosive Strength

4.2.3.1 Lower and Upper Body Explosive Strength Lower body explosive strength has been traditionally assessed with jump tests. The reliability of the standing broad jump test was assessed in two high-quality studies [31, 87]. The standing broad jump test demonstrated moderate to high test–retest reliability, with ICCs of 0.89 to 0.98 and low dispersion of test–retest measurement errors (SEMs < 15%) in adults aged 18–45 years [31, 87]. The reliability of the *hop sequence test* (i.e., single leg hop for distance, 6-m timed hop, triple hop for distance, and crossover hop for distance) was analyzed in two high-quality studies in young adults aged 18–24 years [116, 117]. Based on coefficients of correlation, this test showed moderate to high test–retest reliability (ICCs = 0.80–0.95). SEM values based on test–retest data were reported as 4.5–7.9 cm for the single leg hop, 0.06–0.13 s for the timed hop, 11.2–23.2 cm for the triple hop, and 15.9–21.2 cm for the crossover hop [116, 117]. Learning effects were observed for all hops (SEMs between

trials 1–5 were < 15% vs. < 2.3% between trials 4–5 [117]). Other field-based fitness tests to assess low body explosive strength used in reliability studies were: countermovement vertical jump [118], jump and reach [105], figure-eight hop, up-and-down hop, and t side-to-side hop [119]. Field-based fitness tests to assess trunk and upper body explosive strength used in reliability studies were: modified versions of the medicine ball put [120] and front/side abdominal power [121].

Levels of Evidence: There was moderate evidence indicating that the standing broad jump and the hop sequence tests produce results with moderate to high test–retest reliability in adults aged 18–45 years. Due to the low number of studies, there was limited evidence indicating that: (i) the countermovement vertical jump test provides results with moderate to high reliability in young adults; (ii) the jump and reach test offers high reliability in adults aged 25–59 years; (iii) the figure-eight hop and up-and-down hop tests produce results with high reliability and the side-to-side hop test shows moderate reliability in young adults; (iv) the modified version of the medicine ball put test (i.e., utilizing a 45° incline bench to facilitate the optimal trajectory of the shot-put ball) returns results with high test–retest reliability in young adults; and (v) the front/side abdominal power test demonstrates high test–retest reliability in young adults (see Table 2).

4.2.4 Flexibility

Sit-and-reach and toe-touch tests are widely used to assess hamstring and lower back flexibility. The reliability of the *sit-and-reach test* and its modified versions was investigated in 13 high-quality studies [31, 93, 122–132]. Several protocols of the sit-and-reach test are described in the scientific literature. Ten studies analyzed the reliability of the sit-and-reach test, one the modified sit-and-reach test, four the back-saver sit-and-reach test, three the modified back-saver sit-and-reach test, three the V sit-and-reach test, and one the chair sit-and-reach test. All sit-and-reach tests have the same testing procedure involving maximal trunk flexion, but with differences with respect to position (unilateral or bilateral) and equipment (standard sit-and-reach box, chair, bench or floor). Nine studies in adults aged 18–64 years concluded that the sit-and-reach test provided high test-rest reliability based on coefficients of correlation (ICCs = 0.91–0.99 [31, 93, 122, 123, 126, 128–130, 132], *r*s = 0.95–0.98 [124, 125]) and one study found moderate reliability (ICC = 0.83 [127]). In four of these studies, the sit-and-reach test showed the low dispersion of measurement errors between test–retest (SEMs < 11% [31, 93, 122, 128]). Test–retest measurements in the V sit-and-reach test [126, 129, 130], the back-saver sit-and-reach test [126, 129–131] and the modified back-saver sit-and-reach test [129, 130, 132] demonstrated moderate to high test–retest agreement (ICCs ≥ 0.80) in adults aged 18–64 years. Finally, the chair sit-and-reach test showed moderate to high test–retest reliability based on ICCs of 0.89–0.97 [132] in adults aged 18–48 years. The reliability of the *toe-touch test* was examined in ten studies (eight of high quality [94, 128, 130, 132–136]) in adults aged 18–64 years. The toe-touch showed high reliability based on coefficients of correlation (ICCs = 0.93–0.99 [130, 133, 135], *r*s = 0.88–0.97 [94, 134]) in five of these studies and moderate reliability in two (ICCs ≥ 0.89 [132, 137]). The SEM was calculated in three of these studies, revealing the low dispersion of measurement errors between test–retest in two (SEMs < 15% [128, 135]).

The reliability of the *trunk lift test* used to measure trunk flexibility and endurance strength was analyzed in two high-quality studies in adults aged 18–28 years [124, 138]. The trunk lift test demonstrated high test–retest reliability (ICC = 0.96, SEM < 15% [138], *r* = 0.96 [124]) in young adults. Finally, the reliability of the *side-bending of the trunk test* used to assess lateral trunk flexibility was investigated in two studies in adults aged 25–59 years [94, 105]. One study found high coefficients of correlation in test–retest and inter-rater reliability (*r*s = 0.82–0.88), but significant differences between test–retest ($p < 0.001$) [94]. In the other study, the side-bending of the trunk test demonstrated high inter-rater (ICC = 0.92) and test–retest reliability (mean test–retest difference of − 0.5 cm and CI of − 1.3–0.3 cm and a CV of 4.7%) [105]. In both studies the side-bending of the trunk test showed the low dispersion of measurement errors between test–retest (SEMs < 5% [94, 105]). Other musculoskeletal field-based fitness tests to assess trunk or upper body flexibility used in reliability studies were: hand-behind-back [139] and shoulder–neck mobility [105].

Levels of Evidence: Overall, there was strong evidence indicating that the sit-and-reach test and its modified versions and the toe-touch test offer high test–retest reliability in adults aged 18–64 years. There was moderate evidence indicating that the trunk lift test produces results with high reliability in adults aged 18–28 years. Due to the low number of studies, there was limited evidence that: (i) the hand-behind-back test shows high reliability; and (ii) the shoulder–neck mobility test provides results with low agreement. Finally, there was inconclusive evidence regarding the level of reliability of the side-bending test (see Table 2).

4.3 Motor Fitness

Motor fitness refers to any component of physical fitness that enables a person to successfully perform a particular motor task, game, or activity [140]. Specific motor fitness

components include agility, balance, coordination, power, reaction time, and speed [29]. The reliability of tests assessing motor fitness was investigated in 22 studies (all classified as high quality).

4.3.1 Balance

The reliability of static and dynamic balance tests was analyzed in nine studies [31, 43, 58, 105, 109, 141–144]. Among the static balance tests, reliability studies analyzed the single-leg stand test using different protocols (i.e., eyes open, closed or with the head turned on the floor or on a bar) and also with the variations feet together, semi-tandem, and tandem stand [31, 43, 58, 105, 109, 141]. The *single-leg stand test* (eyes open) demonstrated low test–retest reliability in three studies (ICCs = 0.60–0.73; SEMs = 10–18% [31, 109], r = 0.69 [43]) in adults aged 19–64 years. A fourth study analyzed the single-leg stand test with eyes open, closed, or head turned in adults aged 25–59 years, and obtained results with low inter-rater reliability in all tests (ICCs = 0.18–0.76 and SEMs = 7–38%) and high variability between test–retest for standing on one leg with eyes open (mean differences between testing days of 3.7 s with a confidence interval of − 2.2 to 9.6 s), eyes closed (mean differences of 0.6 s, from − 1.6 to 2.8 s) and head turned (mean differences of 2.6 s, from − 0.7 to 5.9 s), and low CVs ranging from 11 to 5% [105]. However, the single-leg stand test (on a narrow bar) showed high reliability based on ICCs > 0.90 and SEMs ≤ 5.5% in adults aged 36–64 years [141]. The tests semi-tandem stand with eyes open and tandem stand with eyes open and closed demonstrated low test–retest reliability according to ICCs of 0.27–0.58, but the dispersion of test–retest measurement errors was low based on SEMs ≤ 8.5% in adults aged 50–64 years [109]. The Romberg test (eyes open) showed high reliability based on a percent agreement for the parallel stance of 94.7% and low reliability for the semi-tandem and tandem positions (< 75%) in adults aged 55–70 years [58].

The reliability of the *star excursion balance test* was analyzed in three studies in adults aged 18–50 years [142, 143, 145]. Overall, this test demonstrated a moderate test–retest reliability for the dominant leg (ICCs > 0.80) and low test–retest reliability for the non-dominant leg (ICCs < 0.75). In two of these studies, this test showed the low dispersion of measurement errors between test–retest (SEMs < 5% [142, 145]). Other dynamic balance field-based fitness tests used in reliability studies were: functional balance (i.e., walk forwards and backwards across a wooden plank) [144] and tandem walking (forwards and backwards) [141].

Levels of Evidence: There was strong evidence indicating that the single-leg stand test and its modified versions show low reliability in adults aged 19–64 years. There was limited evidence, mainly due to a limited number of studies, indicating that the tandem walking and functional balance tests provide results showing moderate to high agreement in adults aged 33–64 years. Finally, there was inconclusive evidence regarding the level of reliability of the parallel, semi-tandem and tandem stand tests and star excursion balance test (see Table 2).

4.3.2 Gait Speed-Agility

The *timed up-and-go test* is a multidimensional test that measures mobility skills (agility), combining gait speed, balance, and functional capacity [146]. The reliability of this test was investigated in two high-quality studies. The timed up-and-go test (3 m) returned high test–retest reliability based on coefficients of correlation (ICCs = 0.90–0.98) and the low dispersion of measurement errors between test–retest (SEMs ≤ 6%) in healthy adults and adults with low back pain aged 21–64 years [47, 147]. The *T*-test is a measure of four-directional agility and body control that assesses the ability to change direction quickly and maintain balance without reducing speed. Its reliability was analyzed in three studies in adults aged 18–39 years. The *T*-test [117, 148] and its modified version [149] (i.e., reducing the total distance covered) provided results with moderate to high reliability with ICCs ranging between 0.82 and 0.98 and a low dispersion of measurement errors between test–retest (SEMs ≤ 3%) [117, 148, 149]. In addition, the modified *T*-test demonstrated no systematic differences between test sessions (p > 0.05), a mean test–retest difference of 0.03 s, a limit of agreement of ± 0.33 s for females and ± 0.37 s for males, and low CVs (≤ 2.7%) [149]. The reliability of the *hexagon agility test,* a measure of agility and foot quickness involving balance and coordination, was analyzed in two studies in adults aged 19–30 years [119, 150]. The hexagon agility test and its modified version (i.e., hopping single-legged instead of double-legged) showed moderate to high test–retest reliability (ICCs = 0.84–0.94) and the low dispersion of measurement errors between test–retest (SEMs ≤ 0.7 s [119] ranging from 1 to 9.4% between sessions [150]). Thus, a practice trial is recommended prior to recording scores to minimize any possible learning effect. The reliability of *the gait speed test* was investigated in five studies. The gait speed test (4 m/3 m) demonstrated low reliability in two studies in adults aged 35–64 years (ICC = 0.56; SEM = 2% [109], *r* = 0.57 [43]). However, in one study in healthy adults and adults with low back pain aged 21–63 years, the gait speed test [15.2 m (50 ft) fast walk or preferred speed walk] showed high reliability (ICCs = 0.91–0.99; SEMs ≤ 4.6% [47]). In another study also in healthy adults and those with low/high back pain aged 27–59 years, the gait speed test (2 × 20-m walk, with and without load) showed no significant mean test–retest

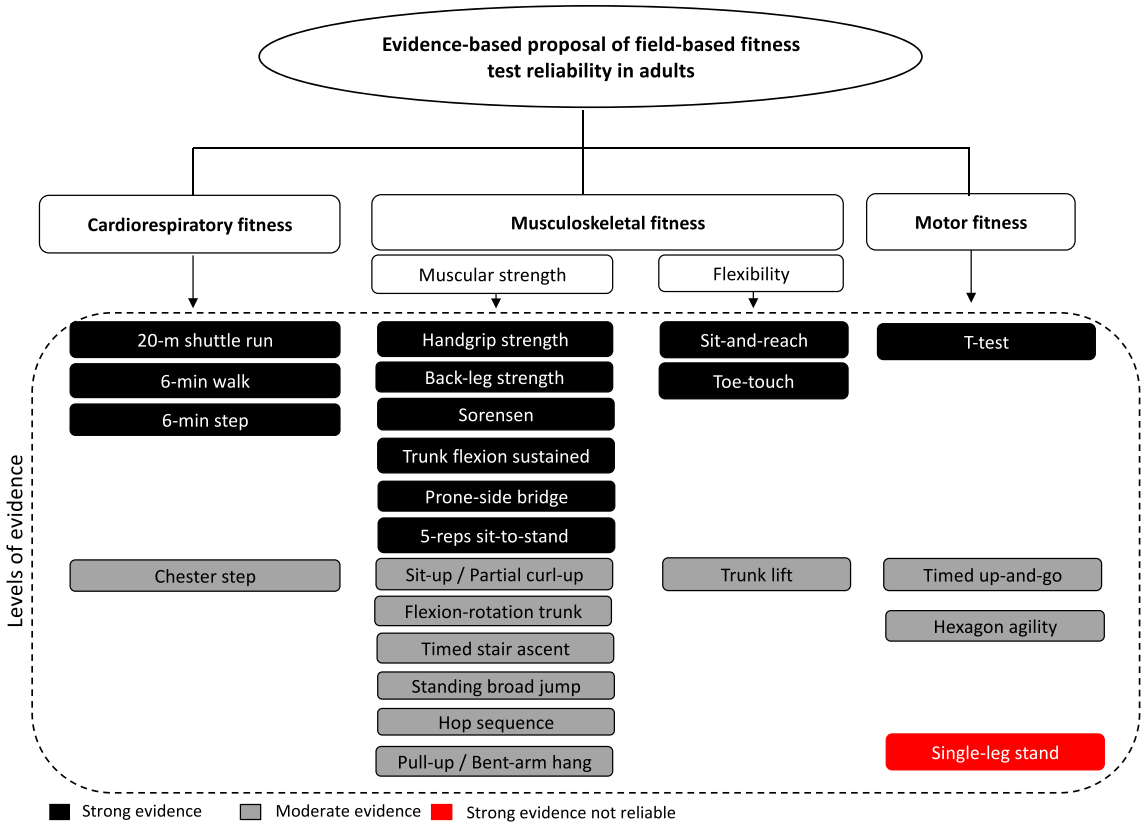

Fig. 2 Evidence-based proposal of field-based fitness test reliability in adults

differences ($p > 0.05$) [89]. Finally, in one study in adults of mean age 39 years, the gait speed test (50 m fast walk with and without load) returned results showing a high reliability based on ICCs of 0.98–0.99 [112]. Other gait speed and agility field-based fitness tests used in reliability studies were: 4-square step [45], 10×5 shuttle run [31], pro-agility [87], agility track [151], zig-zag run [119], Illinois agility, Edgren side step [148], and plate tapping [31].

Levels of Evidence: There was strong evidence indicating that the *T*-test shows moderate to high test–retest and inter-rater reliability in adults aged 18–39 years. There was moderate evidence indicating that: (i) the timed up-and-go test demonstrates high reproducibility in adults aged 19–64 years with or without low back pain; and (ii) the hexagon agility test and its modified version show moderate to high reliability in adults aged 19–64 years. A large number of other motor fitness tests provided limited evidence, mainly due to a low number of studies: (i) the 10×5 shuttle run and zig-zag tests provide results with moderate and high reliability, respectively, in young adults; (ii) the agility track and pro-agility tests show high test–retest reliability in adults aged 28–55 years; (iii) the 4-square step test returned results with moderate test–retest reliability in adults aged 19–43 years; (iv) the Edgren side step and Illinois agility tests show low test–retest reliability but high inter-rater reliability in adults aged 19–39 years; and (v) the plate tapping test provides results with low reliable values in young adults of mean age 19.5 years. Finally, there was inconclusive evidence regarding the level of reliability of the gait speed tests. The fact that different versions of the gait speed test were used involving sample groups with different characteristics makes it difficult to compare their reliability data (see Table 2).

5 Conclusions

To our knowledge, this is the first systematic review to examine and compare reliability data from studies analyzing field-based fitness tests including tests of cardiorespiratory fitness, musculoskeletal fitness and motor fitness in adults aged 19–64 years. Based on quality assessment and established levels of evidence for each study, our findings indicate: (i) a strong level of evidence exists for the high reliability (see Fig. 2) of the cardiorespiratory fitness tests: *20-m shuttle run, 6-min step* and *6-min* walk. The 20-m shuttle run is a maximal test that offers high reliability in young people,

while the 6-min step and the 6-min walk tests emerged as possible alternative submaximal tests to assess cardiorespiratory fitness in adults who are not very physically fit and/or find it difficult to run (e.g., obese individuals); (ii) among the musculoskeletal fitness tests, the *handgrip strength test* using a JAMAR dynamometer offers high test–retest reliability yet this dynamometer does not allow for adapting grip span to the individual's hand size with the same precision as the TKK dynamometer; (iii) the *back-leg strength, Sorensen* and its modified versions, *trunk flexion sustained, back-leg strength* and *5-reps sit-to-stand* tests show high reliability; (iv) the *bilateral side* and *prone bridge tests* provide results with moderate reliability; (v) the *sit-and-reach test* and its modified versions, and the *toe-touch test* show high reliability; and finally, among the motor fitness tests (vi) the *single-leg stand test* and its modified versions demonstrate low reliability; and (vii) the *T*-test and its modified version provide results demonstrating moderate reliability. There is inconclusive evidence regarding the reliability of balance and gait speed tests. Due to the low number of studies, evidence for the reliability of a large number of other field-based fitness tests was limited.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40279-021-01635-2.

Acknowledgements Language and editorial assistance was provided by Ana Burton.

Declarations

Funding This project was supported by the Ministry of Economy, Industry and Competitiveness in the 2017 call for R&D Projects of the State Program for Research, Development and Innovation Targeting the Challenges of the Company; National Plan for Scientific and Technical Research and Innovation 2013–2016 (DEP2017-88043-R); and the Regional Government of Andalusia and University of Cadiz: Research and Knowledge Transfer Fund (PPIT-FPI19).

Conflict of interest Magdalena Cuenca-Garcia, Nuria Marin-Jimenez, Alejandro Perez-Bey, David Sanchez-Oliva, Daniel Camiletti-Moiron, Inmaculada C. Alvarez-Gallardo, Francisco B. Ortega and Jose Castro-Piñero declare that they have no conflicts of interest relevant to the content of this review.

Ethics approval Not applicable.

Authors' contributions MCG and JCP conceived the study idea. MCG led the writing of the review and carried out methodological procedures with APB, DSO and NMJ. All authors discussed the results and contributed to the final manuscript, and agreed upon the order of presentation of the authors. All authors have read and approved the final manuscript.

Data availability statement The authors declare that all relevant data are included in the article and/or its supplementary information files.

References

1. Carbone S, Kirkman DL, Garten RS, Rodriguez-Miguelez P, Artero EG, Lee D-C, et al. Muscular strength and cardiovascular disease: an updated state-of-the-art narrative review. J Cardiopulm Rehabil Prev. 2020;40(5):302–9.
2. Kaminsky LA, Arena R, Ellingsen Ø, Harber MP, Myers J, Ozemek C, et al. Cardiorespiratory fitness and cardiovascular disease—the past, present, and future. Prog Cardiovasc Dis. 2019;62(2):86–93.
3. Lavie CJ, Ozemek C, Carbone S, Katzmarzyk PT, Blair SN. Sedentary behavior, exercise, and cardiovascular health. Circ Res. 2019;124(5):799–815.
4. Kodama S, Saito K, Tanaka S, Maki M, Yachi Y, Asumi M, et al. Cardiorespiratory fitness as a quantitative predictor of all-cause mortality and cardiovascular events in healthy men and women: a meta-analysis. JAMA. 2009;301(19):2024–35.
5. Ruiz JR, Castro-Pinero J, Artero EG, Ortega FB, Sjostrom M, Suni J, et al. Predictive validity of health-related fitness in youth: a systematic review. Br J Sports Med. 2009;43(12):909–23.
6. García-Hermoso A, Ramírez-Campillo R, Izquierdo M. Is muscular fitness associated with future health benefits in children and adolescents? a systematic review and meta-analysis of longitudinal studies. Sports Med. 2019;49(7):1079–94.
7. Harber MP, Kaminsky LA, Arena R, Blair SN, Franklin BA, Myers J, et al. Impact of cardiorespiratory fitness on all-cause and disease-specific mortality: advances since 2009. Prog Cardiovasc Dis. 2017;60(1):11–20.
8. García-Hermoso A, Cavero-Redondo I, Ramírez-Vélez R, Ruiz JR, Ortega FB, Lee D-C, et al. Muscular strength as a predictor of all-cause mortality in an apparently healthy population: a systematic review and meta-analysis of data from approximately 2 million men and women. Arch Phys Med Rehabil. 2018;99(10):2100-2113.e5.
9. Barry VW, Caputo JL, Kang M. The joint association of fitness and fatness on cardiovascular disease mortality: a meta-analysis. Prog Cardiovasc Dis. 2018;61(2):136–41.
10. Canadia CSFEPT. Physical activity, fitness and lifestyle appraisal. Canada H, editor. Ottawa, ON; 1996.
11. Suni JH, Oja P, Miilunpalo SI, Pasanen ME, Vuori IM, Bos K. Health-related fitness test battery for adults: associations with perceived health, mobility, and back function and symptoms. Arch Phys Med Rehabil. 1998;79(5):559–69.
12. Oja P, Tuxworth BE, editor. Eurofit for adults: assessment of health-related fitness [Internet]. Finland: Council of Europe Publishing; 1995. http://www.ukkinstituutti.fi/filebank/500-ALPHA_FIT_Testers_Manual.pdf. Accessed 22 Apr 2021.
13. Suni J, Husu P, Rinne M. Fitness for health: the ALPHA-FIT test battery for adults aged 18–69 [Internet]. 2009. https://ukkinstituutti.fi/filebank/500-ALPHA_FIT_Testers_Manual.pdf. Accessed 25 Apr 2021.
14. Ruiz JR, Castro-Piñero J, España-Romero V, Artero EG, Ortega FB, Cuenca MM, et al. Field-based fitness assessment in young people: the ALPHA health-related fitness test battery for children and adolescents. Br J Sports Med. 2011;45(6):518–24.
15. Suni JH, Miilunpalo SI, Asikainen TM, Laukkanen RT, Oja P, Pasanen ME, et al. Safety and feasibility of a health-related fitness test battery for adults. Phys Ther. 1998;78(2):134–48.
16. Drake D, Kennedy R, Wallace E. The validity and responsiveness of isometric lower body multi-joint tests of muscular strength: a systematic review. Sport Med Open. 2017;3(1):23.
17. Ortega FB, Cadenas-Sánchez C, Sánchez-Delgado G, Mora-González J, Martínez-Téllez B, Artero EG, et al. Systematic review and proposal of a field-based physical fitness-test

battery in preschool children: the PREFIT battery. Sports Med. 2015;45(4):533–55.
18. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535.
19. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev. 2012;1:2.
20. Artero EG, España-Romero V, Castro-Piñero J, Ortega FB, Suni J, Castillo-Garzon MJ, et al. Reliability of field-based fitness tests in youth. Int J Sports Med. 2011;32(3):159–69.
21. Castro-Pinero J, Artero E, Espana-Romero V, Ortega F, Sjostrom M, Suni J, et al. Criterion-related validity of fi eld-based fitness tests in youth: a systematic review. Br J Sport Med. 2010;2010(44):934–43.
22. Hopkinsm WG. Measures of reliability in sports medicine and science. Sport Med. 2000;30(1):1–15.
23. Olds T. Five errors about error. J Sci Med Sport. 2002;5:336–40.
24. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998;26(4):217–38.
25. Bruton A, Conway JH, Holgate ST. Reliability: what is it and how is it measured? Physiotherapy. 2000;86(2):94–9.
26. Vincent-Smith B, Gibbons P. Inter-examiner and intra-examiner reliability of the standing flexion test. Man Ther. 1999;4(2):87–93.
27. Weir JP. Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231–40.
28. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–82.
29. ACSM. American College of Sports Medicine, ACSM's guidelines for exercise testing and prescription. 9th ed. Philadelphia: Lippincott Williams & Williams; 2013.
30. Sproule J, Kunalan C, McNeill M, Wright H. Validity of 20-MST for predicting VO2max of adult Singaporean athletes. Br J Sports Med. 1993;27(3):202–4.
31. Tsigilis N, Douda H, Tokmakidis SP. Test–retest reliability of the Eurofit test battery administered to university students. Percept Mot Skills. 2002;95(3 Pt 2):1295–300.
32. Cooper S-M, Baker JS, Tong RJ, Roberts E, Hanford M. The repeatability and criterion related validity of the 20 m multistage fitness test as a predictor of maximal oxygen uptake in active young men. Br J Sports Med. 2005;39(4):e19.
33. Lamb KL, Rogers L. A re-appraisal of the reliability of the 20 m multi-stage shuttle run test. Eur J Appl Physiol. 2007;100(3):287–92.
34. Aandstad A, Holme I, Berntsen S, Anderssen SA. Validity and reliability of the 20 meter shuttle run test in military personnel. Mil Med. 2011;176(5):513–8.
35. Kim J, Jung SH, Cho HC. Validity and reliability of shuttle-run test in Korean adults. Int J Sports Med. 2011;32(8):580–5.
36. Metsios GS, Flouris AD, Koutedakis Y, Nevill A. Criterion-related validity and test–retest reliability of the 20m square shuttle test. J Sci Med Sport. 2008;11(2):214–7.
37. Penry JT, Wilcox AR, Yun J. Validity and reliability analysis of Cooper's 12-minute run and the multistage shuttle run in healthy adults. J Strength Cond Res. 2011;25(3):597–605.
38. Leger LA, Mercier D, Gadoury C, Lambert J. The multistage 20 metre shuttle run test for aerobic fitness. J Sports Sci. 1988;6(2):93–101.
39. Flouris AD, Metsios GS, Famisis K, Geladas N, Koutedakis Y. Prediction of VO2max from a new field test based on portable indirect calorimetry. J Sci Med Sport. 2010;13(1):70–3.
40. Mikawa K, Yano Y, Senjyu H. Development of a field test for evaluating aerobic fitness. Int J Sports Med. 2012;33(5):346–50.
41. Taylor S, Frost H, Taylor A, Barker K. Reliability and responsiveness of the shuttle walking test in patients with chronic low back pain. Physiother Res Int J Res Clin Phys Ther. 2001;6(3):170–8.
42. Jurgensen SP, Trimer R, Dourado VZ, Di Thommazo-Luporini L, Bonjorno-Junior JC, Oliveira CR, et al. Shuttle walking test in obese women: test–retest reliability and concurrent validity with peak oxygen uptake. Clin Physiol Funct Imaging. 2015;35(2):120–6.
43. Curb JD, Ceria-Ulep CD, Rodriguez BL, Grove J, Guralnik J, Willcox BJ, et al. Performance-based measures of physical function for high-function populations. J Am Geriatr Soc. 2006;54(5):737–42.
44. Reuter SE, Massy-Westropp N, Evans AM. Reliability and validity of indices of hand-grip strength and endurance. Aust Occup Ther J. 2011;58(2):82–7.
45. Wilken JM, Darter BJ, Goffar SL, Ellwein JC, Snell RM, Tomalis EA, et al. Physical performance assessment in military service members. J Am Acad Orthop Surg. 2012;20(Suppl 1):S42–7.
46. Larsson UE, Reynisdottir S. The six-minute walk test in outpatients with obesity: reproducibility and known group validity. Physiother Res Int. 2008;13(2):84–93.
47. Simmonds MJ, Olson SL, Jones S, Hussein T, Lee CE, Novy D, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine (Phila Pa 1976). 1998;23(22):2412–21.
48. Laukkanen RMT, Kukkonen-Harjula TK, Oja P, Pasanen ME, Vuori IM. Prediction of change in maximal aerobic power by the 2-km walk test after walking training in middle-aged adults. Int J Sports Med. 2000;21(2):113–6.
49. Gabriel KKP, Rankin RL, Lee C, Charlton ME, Swan PD, Ainsworth BE, et al. Test–retest reliability and validity of the 400-meter walk test in healthy, middle-aged women. J Phys Act Health. 2010;7(5):649–57.
50. Larsen GE, George JD, Alexander JL, Fellingham GW, Aldana SG, Parcell AC. Prediction of maximum oxygen consumption from walking, jogging, or running. Res Q Exerc Sport. 2002;73(1):66–72.
51. Léger L, Bouchet R. An indirect continuous running multistage field test: the Université de Montréal Track Test. Can J Appl Sport Sci. 1980;5(2):77–84.
52. Beriault K, Carpentier AC, Gagnon C, Menard J, Baillargeon J-P, Ardilouze J-L, et al. Reproducibility of the 6-minute walk test in obese adults. Int J Sports Med. 2009;30(10):725–7.
53. Carvalho LP, Di Thommazo-Luporini L, Aubertin-Leheudre M, Bonjorno Junior JC, de Oliveira CR, Luporini RL, et al. Prediction of cardiorespiratory fitness by the six-minute step test and its association with muscle strength and power in sedentary obese and lean young women: a cross-sectional study. PLoS One. 2015;10(12):e0145960.
54. Arcuri JF, Borghi-Silva A, Labadessa IG, Sentanin AC, Candolo C, Pires Di Lorenzo VA. Validity and reliability of the 6-minute step test in healthy individuals: a cross-sectional study. Clin J Sport Med. 2016;26(1):69–75.
55. Giacomantonio N, Morrison P, Rasmussen R, MacKay-Lyons MJ. Reliability and validity of the 6-minute step test for clinical assessment of cardiorespiratory fitness in people at risk of cardiovascular disease. J Strength Cond Res. 2018;34(5):1376–82.
56. Buckley JP, Sim J, Eston RG, Hession R, Fox R. Reliability and validity of measures taken during the Chester step test to predict aerobic power and to prescribe aerobic exercise. Br J Sports Med. 2004;38(2):197–205.

57. Sykes K, Roberts A. The Chester step test—a simple yet effective tool for the prediction of aerobic capacity. Physiotherapy. 2004;90:183–8.
58. Ritchie C, Trost SG, Brown W, Armit C. Reliability and validity of physical fitness field tests for adults aged 55 to 70 years. J Sci Med Sport. 2005;8(1):61–70.
59. Santa Maria DL, Kinnear GR, Kearney JT, Martin P. The objectivity, reliability, and validity of the osu step test for college males. Res Q Am Alliance Health Phys Educ Recreat. 1976;47(3):445–52.
60. Gerodimos V. Reliability of handgrip strength test in basketball players. J Hum Kinet. 2012;31:25–36.
61. Savva C, Karagiannis C, Rushton A. Test–retest reliability of grip strength measurement in full elbow extension to evaluate maximum grip strength. J Hand Surg Eur. 2013;38(2):183–6.
62. Bohannon RW. Test–retest reliability of the five-repetition sit-to-stand test: a systematic review of the literature involving adults. J Strength Cond Res. 2011;25(11):3205–7.
63. van Meeteren J, van Rijn RM, Selles RW, Roebroeck ME, Stam HJ. Grip strength parameters and functional activities in young adults with unilateral cerebral palsy compared with healthy subjects. J Rehabil Med. 2007;39(8):598–604.
64. Boissy P, Bourbonnais D, Carlotti MM, Gravel D, Arsenault BA. Maximal grip force in chronic stroke subjects and its relationship to global upper extremity function. Clin Rehabil. 1999;13(4):354–62.
65. Chkeir A, Jaber R, Hewson DJ, Duchene J. Reliability and validity of the grip-ball dynamometer for grip-strength measurement. Conf Proc Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf. 2012;2012:1996–9.
66. Lusardi MBR. Hand Grip strength: comparability of measurements obtained with a jamar dynamometer and a modified sphygmomanometerle. J Hand Ther. 1991;4(4):117–22.
67. Hamilton GF, McDonald C, Chenier TC. Measurement of grip strength: validity and reliability of the sphygmomanometer and jamar grip dynamometer. J Orthop Sports Phys Ther. 1992;16(5):215–9.
68. Cadenas-Sanchez C, Sanchez-Delgado G, Martinez-Tellez B, Mora-Gonzalez J, Lof M, Espana-Romero V, et al. Reliability and validity of different models of TKK hand dynamometers. Am J Occup Ther. 2016;70(4):7004300010.
69. Bohannon RW. Test–retest reliability of the MicroFET 4 handgrip dynamometer. Physiother Theory Pract. 2006;22(4):219–21.
70. Reijnierse EM, de Jong N, Trappenburg MC, Blauw GJ, Butler-Browne G, Gapeyeva H, et al. Assessment of maximal handgrip strength: how many attempts are needed? J Cachexia Sarcopenia Muscle. 2017;8(3):466–74.
71. Peolsson A, Hedlund R, Oberg B. Intra- and inter-tester reliability and reference values for hand strength. J Rehabil Med. 2001;33(1):36–41.
72. Plant CE, Parsons NR, Edwards AT, Rice H, Denninson K, Costa ML. A comparison of electronic and manual dynamometry and goniometry in patients with fracture of the distal radius and healthy participants. J Hand Ther. 2016;29(1):73–80.
73. Hamilton A, Balnave R, Adams R. Grip strength testing reliability. J Hand Ther. 1994;7(3):163–70.
74. Shechtman O, MacKinnon L, Locklear C. Using the BTE Primus to measure grip and wrist flexion strength in physically active wheelchair users: an exploratory study. Am J Occup Ther. 2001;55(4):393–400.
75. Shechtman O, Davenport R, Malcolm M, Nabavi D. Reliability and validity of the BTE-Primus grip tool. J Hand Ther. 2003;16(1):36–42.
76. Shechtman O, Gestewitz L, Kimble C. Reliability and validity of the DynEx dynamometer. J Hand Ther. 2005;18(3):339–47.
77. Coldham F, Lewis J, Lee H. The reliability of one vs. three grip trials in symptomatic and asymptomatic subjects. J Hand Ther. 2006;19(3):318–27.
78. Svantesson U, Norde M, Svensson S, Brodin E. A comparative study of the Jamar (R) and the Grippit (R) for measuring handgrip strength in clinical practice. Isokinet Exerc Sci. 2009;17(2):85–91.
79. Mathiowetz V. Comparison of Rolyan and Jamar dynamometers for measuring grip strength. Occup Ther Int. 2002;9(3):201–9.
80. Shechtman O. The coefficient of variation as a measure of sincerity of effort of grip strength, part I: the statistical principle. J Hand Ther. 2001;14(3):180–7.
81. Espana-Romero V, Artero EG, Santaliestra-Pasias AM, Gutierrez A, Castillo MJ, Ruiz JR. Hand span influences optimal grip span in boys and girls aged 6 to 12 years. J Hand Surg Am. 2008;33:378–84.
82. Ruiz JR, Espana-Romero V, Ortega FB, Sjöström M, Castillo MJ, Gutierrez A. Hand span influences optimal grip span in male and female teenagers. J Hand Surg Am. 2006;31:1367–72.
83. Ruiz-Ruiz J, Mesa JL, Gutiérrez A, Castillo MJ. Hand size influences optimal grip span in women but not in men. J Hand Surgery, Am. 2002;27:897–901.
84. Sanchez-Delgado G, Adenas-Sanchez C, Mora-Gonzalez J, Martinez-Tellez B, Chillón P, Löf M, et al. Assessment of handgrip strength in preschool children aged 3 to 5 years. J Hand Surg Eur. 2015;40:966–72.
85. Ten Hoor GA, Musch K, Meijer K, Plasqui G. Test–retest reproducibility and validity of the back-leg-chest strength measurements. Isokinet Exerc Sci. 2016;24(3):209–16.
86. Coldwells A, Atkinson G, Reilly T. Sources of variation in back and leg dynamometry. Ergonomics. 1994;37(1):79–86.
87. Whitehead PN, Schilling BK, Peterson DD, Weiss LW. Possible new modalities for the Navy physical readiness test. Mil Med. 2012;177(11):1417–25.
88. Keller A, Hellesnes J, Brox JI. Reliability of the isokinetic trunk extensor test, Biering–Sorensen test, and Astrand bicycle test: assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. Spine (Phila Pa 1976). 2001;26(7):771–7.
89. Ljungquist T, Harms-Ringdahl K, Nygren A, Jensen I. Intra- and inter-rater reliability of an 11-test package for assessing dysfunction due to back or neck pain. Physiother Res Int. 1999;4(3):214–32.
90. Ito T, Shirado O, Suzuki H, Takahashi M, Kaneda K, Strax TE. Lumbar trunk muscle endurance testing: an inexpensive alternative to a machine for evaluation. Arch Phys Med Rehabil. 1996;77(1):75–9.
91. del Pozo-Cruz B, Mocholi MH, del Pozo-Cruz J, Parraca JA, Adsuar JC, Gusi N. Reliability and validity of lumbar and abdominal trunk muscle endurance tests in office workers with nonspecific subacute low back pain. J Back Musculoskelet Rehabil. 2014;27(4):399–408.
92. Reiman MP, Krier AD, Nelson JA, Rogers MA, Stuke ZO, Smith BS. Reliability of alternative trunk endurance testing procedures using clinician stabilization vs. traditional methods. J Strength Cond Res. 2010;24(3):730–6.
93. Kahraman BO, Sengul YS, Kahraman T, Kalemci O, Ozcan Kahraman B, Salik Sengul Y, et al. Developing a reliable core stability assessment battery for patients with nonspecific low back pain. Spine (Phila Pa 1976). 2016;41(14):E844–50.
94. Hyytiäinen K, Salminen J, Suvitie T, Wickström G, Pentti J. Reproducibility of nine tests to measure spinal mobility and trunk muscle strength. Scan J Rehab Med. 1991;23:3–10.
95. Larsson H, Tegern M, Monnier A, Skoglund J, Helander C, Persson E, et al. Content validity index and intra- and inter-rater

reliability of a new muscle strength/endurance test battery for Swedish soldiers. PLoS ONE. 2015;10(7):e0132185.
96. Latimer J, Maher CG, Refshauge K, Colaco I. The reliability and validity of the Biering–Sorensen test in asymptomatic subjects and subjects reporting current or previous nonspecific low back pain. Spine (Phila Pa 1976). 1999;24(20):2085–9.
97. Gruther W, Wick F, Paul B, Leitner C, Posch M, Matzner M, et al. Diagnostic accuracy and reliability of muscle strength and endurance measurements in patients with chronic low back pain. J Rehabil Med. 2009;41(8):613–9.
98. Evans K, Refshauge KM, Adams R. Trunk muscle endurance tests: reliability, and gender differences in athletes. J Sci Med Sport. 2007;10(6):447–55.
99. Greene PF, Durall CJ, Kernozek TW. Intersession reliability and concurrent validity of isometric endurance tests for the lateral trunk muscles. J Sport Rehabil. 2012;21(2):161–6.
100. Durall CJ, Greene PF, Kernozek TW. A comparison of two isometric tests of trunk flexor endurance. J Strength Cond Res. 2012;26(7):1939–44.
101. De Blaiser C, De Ridder R, Willems T, Danneels L, Vanden Bossche L, Palmans T, et al. Evaluating abdominal core muscle fatigue: assessment of the validity and reliability of the prone bridging test. Scand J Med Sci Sports. 2018;28(2):391–9.
102. Diener M, Golding L, Diener D. Validity and reliability of a one-minute half sit-up test of abdominal strength and endurance. Sport Med Train Rehabil Int J. 1995;6:105–19.
103. Moreland J, Finch E, Stratford P, Balsor B, Gill C. Interrater reliability of six tests of trunk muscle function and endurance. J Orthop Sports Phys Ther. 1997;26(4):200–8.
104. Robertson LD, Magnusdottir N. Evaluation of criteria associated with abdominal fitness testing. Res Q Exerc Sport. 1987;58(3):355–9.
105. Suni JH, Oja P, Laukkanen RT, Miilunpalo SI, Pasanen ME, Vuori IM, et al. Health-related fitness test battery for adults: aspects of reliability. Arch Phys Med Rehabil. 1996;77(4):399–405.
106. Vincent WJ, Britten SD. Evaluation of the curl up asubstitute for the bent knee sit up. J Phys Educ Recreat. 1980;51(2):74–5.
107. Brotons-Gil E, Garcia-Vaquero MP, Peco-Gonzalez N, Vera-Garcia FJ. Flexion-rotation trunk test to assess abdominal muscle endurance: reliability, learning effect, and sex differences. J Strength Cond Res. 2013;27(6):1602–8.
108. Juan-Recio C, Lopez-Plaza D, Barbado Murillo D, Pilar Garcia-Vaquero M, Vera-Garcia FJ. Reliability assessment and correlation analysis of 3 protocols to measure trunk muscle strength and endurance. J Sports Sci. 2018;36(4):357–64.
109. Wolinsky FD, Miller DK, Andresen EM, Malmstrom TK, Miller JP. Reproducibility of physical performance and physiologic assessments. J Aging Health. 2005;17(2):111–24.
110. Bohannon RW, Bubela DJ, Magasi SR, Gershon RC. Relative reliability of three objective tests of limb muscle strength. Isokinet Exerc Sci. 2011;19(2):77–81.
111. Kahraman T, Ozcan Kahraman B, Salik Sengul Y, Kalemci O. Assessment of sit-to-stand movement in nonspecific low back pain: a comparison study for psychometric properties of field-based and laboratory-based methods. Int J Rehabil Res Int Zeitschrift fur Rehabil Rev Int Rech Readapt. 2016;39(2):165–70.
112. LeBrasseur NK, Bhasin S, Miciek R, Storer TW. Tests of muscle strength and physical function: reliability and discrimination of performance in younger and older men and older men with mobility limitations. J Am Geriatr Soc. 2008;56(11):2118–23.
113. Fielitz L, Coelho J, Horne T, Brechue W. Inter-rater reliability and intra-rater reliability of assessing the 2-minute push-up test. Mil Med. 2016;181(2):167–72.
114. Negrete RJ, Hanney WJ, Kolber MJ, Davies GJ, Ansley MK, McBride AB, et al. Reliability, minimal detectable change, and normative values for tests of upper extremity function and power. J Strength Cond Res. 2010;24(12):3318–25.
115. Clemons JM. Construct validity of a modification of the flexed arm hang test. J Strength Cond Res. 2014;28(12):3523–30.
116. 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(12):947–54.
117. Munro AG, Herrington LC. Between-session reliability of four hop tests and the agility T-test. J Strength Cond Res. 2011;25(5):1470–7.
118. Moir G, Shastri P, Connaboy C. Intersession reliability of vertical jump height in women and men. J Strength Cond Res. 2008;22(6):1779–84.
119. Ortiz A, Olson SL, Roddey TS, Morales J. Reliability of selected physical performance tests in young adult women. J Strength Cond Res. 2005;19(1):39–44.
120. Clemons JM, Campbell B, Jeansonne C. Validity and reliability of a new test of upper body power. J Strength Cond Res. 2010;24(6):1559–65.
121. Cowley PM, Swensen TC. Development and reliability of two core stability field tests. J Strength Cond Res. 2008;22(2):619–24.
122. Bozic PR, Pazin NR, Berjan BB, Planic NM, Cuk ID. Evaluation of the field tests of flexibility of the lower extremity: reliability and the concurrent and factorial validity. J Strength Cond Res. 2010;24(9):2523–31.
123. Minkler S, Patterson P. The validity of the modified sit-and-reach test in college-age students. Res Q Exerc Sport. 1994;65(2):189–92.
124. Wear. Relationship of flexibility measurements to length of body segments title. Res Q. 1963;Vol. 34, N.
125. Liemohn WP, Sharpe GL, Wasserman JF. Lumbosacral movement in the sit-and-reach and in Cailliet's protective-hamstring stretch. Spine (Phila Pa 1976). 1994;19(18):2127–30.
126. Hui SC, Yuen PY, Morrow JRJ, Jackson AW. Comparison of the criterion-related validity of sit-and-reach tests with and without limb length adjustment in Asian adults. Res Q Exerc Sport. 1999;70(4):401–6.
127. Shephard RJ, Berridge M, Montelpare W. On the generality of the "sit and reach" test: an analysis of flexibility data for an aging population. Res Q Exerc Sport. 1990;61(4):326–30.
128. Ayala F, de Baranda RS, De Ste CM, 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(4):219–26.
129. Hui SS, Yuen PY. Validity of the modified back-saver sit-and-reach test: a comparison with other protocols. Med Sci Sports Exerc. 2000;32(9):1655–9.
130. Lopez Minarro PA, de Baranda Andujar PS, Rodriguez Garcia PL, Ortega TE. A comparison of the spine posture among several sit-and-reach test protocols. J Sci Med Sport. 2007;10(6):456–62.
131. Leard JS, Crane BA, Ball KA. Intrarater and interrater reliability of 22 clinical measures associated with lower quarter malalignment. J Manip Physiol Ther. 2009;32(4):270–6.
132. Atamaz F, Ozcaldiran B, Ozdedeli S, Capaci K, Durmaz B. Inter-observer and intraobserver reliability in lower-limb flexibility measurements. J Sports Med Phys Fitn. 2011;51(4):689–94.
133. Perret C, Poiraudeau S, Fermanian J, Colau MML, Benhamou MAM, Revel M. Validity, reliability, and responsiveness of the fingertip-to-floor test. Arch Phys Med Rehabil. 2001;82(11):1566–70.
134. Kippers V, Parker AW. Toe-touch test: a measure of its validity. Phys Ther. 1987;67(11):1680–4.

135. Gauvin MG, Riddle DL, Rothstein JM. Reliability of clinical measurements of forward bending using the modified fingertip-to-floor method. Phys Ther. 1990;70(7):443–7.
136. Dijkstra PU, de Bont LG, van der Weele LT, Boering G. Joint mobility measurements: reliability of a standardized method. Cranio. 1994;12(1):52–7.
137. Ayala F, Sainz de Baranda P, 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(9):2372–82.
138. Jackson AW, Morrow JRJ, Jensen RL, Jones NA, Schultes SS. Reliability of the prudential FITNESSGRAM trunk lift test in young adults. Res Q Exerc Sport. 1996;67(1):115–7.
139. van den Dolder PA, Ferreira PH, Refshauge K. Intra- and inter-rater reliability of a modified measure of hand behind back range of motion. Man Ther. 2014;19(1):72–6.
140. Kent M. The Oxford dictionary of sports science and medicine. 3rd ed. Oxford: Oxford University Press; 2007.
141. Rinne MB, Pasanen ME, Miilunpalo SI, Oja P. Test–retest reproducibility and inter-rater reliability of a motor skill test battery for adults. Int J Sports Med. 2001;22(3):192–200.
142. Kinzey SJ, Armstrong CW. The reliability of the star-excursion test in assessing dynamic balance. J Orthop Sports Phys Ther. 1998;27(5):356–60.
143. Gribble PA, Kelly SE, Refshauge KM, Hiller CE. Interrater reliability of the star excursion balance test. J Athl Train. 2013;48(5):621–6.
144. Punakallio A. Trial-to-trial reproducibility and test–retest stability of two dynamic balance tests among male firefighters. Int J Sports Med. 2004;25(3):163–9.
145. López-Plaza D, Juan-Recio C, Barbado D, Ruiz-Pérez I, Vera-Garcia FJ. Reliability of the star excursion balance test and two new similar protocols to measure trunk postural control. PM R. 2018;10(12):1344–52.
146. Podsiadlo D, Richardson S. The timed "Up & Go": a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc. 1991;39(2):142–8.
147. Spagnuolo DL, Jurgensen SP, Iwama AM, Dourado VZ. Walking for the assessment of balance in healthy subjects older than 40 years. Gerontology. 2010;56(5):467–73.
148. Raya MA, Gailey RS, Gaunaurd IA, Jayne DM, Campbell SM, Gagne E, et al. Comparison of three agility tests with male servicemembers: Edgren Side Step Test, T-Test, and Illinois Agility Test. J Rehabil Res Dev. 2013;50(7):951–60.
149. 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 Cond Res. 2009;23(6):1644–51.
150. Beekhuizen KS, Davis MD, Kolber MJ, Cheng M-SS. Test–retest reliability and minimal detectable change of the hexagon agility test. J Strength Cond Res. 2009;23(7):2167–71.
151. Manderoos SA, Vaara ME, Maki PJ, Malkia EA, Aunola SK, Karppi S-L. A new agility test for adults: its test–retest reliability and minimal detectable change in untrained women and men aged 28–55. J Strength Cond Res. 2016;30(8):2226–34.

Authors and Affiliations

Magdalena Cuenca-Garcia[1,2] · Nuria Marin-Jimenez[1,2] · Alejandro Perez-Bey[1,2] · David Sánchez-Oliva[1,2,3] · Daniel Camiletti-Moiron[1,2] · Inmaculada C. Alvarez-Gallardo[1,2] · Francisco B. Ortega[4,5,6] · Jose Castro-Piñero[1,2]

1 GALENO Research Group, Department of Physical Education, Faculty of Education Sciences, School of Education, University of Cádiz, Puerto Real, Avenida República Saharaui S/N, 11519 Puerto Real, Cádiz, Spain

2 Instituto de Investigación e Innovación Biomédica de Cádiz (INiBICA), Cadiz, Spain

3 ACAFYDE Research Group, Faculty of Sport Sciences, University of Extremadura, Caceres, Spain

4 PROFITH "PROmoting FITness and Health Through Physical Activity" Research Group, Sport and Health University Research Institute (iMUDS), Department of Physical and Sports Education, Faculty of Sport Sciences, University of Granada, 18071 Granada, Spain

5 Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

6 Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden