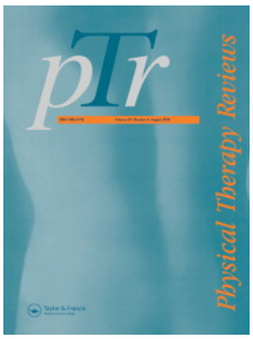

Physical Therapy Reviews

ISSN: 1083-3196 (Print) 1743-288X (Online) Journal homepage: www.tandfonline.com/journals/yptr20

The reliability, validity, and responsiveness of tests used to assess the effects of power training in older adults: a systematic review

Mohamed el Hadouchi, Henri Kiers, Brittany Boerstra, Ralf Berends, Emma Schreuder, Cindy Veenhof & Jaap van Dieën

To cite this article: Mohamed el Hadouchi, Henri Kiers, Brittany Boerstra, Ralf Berends, Emma Schreuder, Cindy Veenhof & Jaap van Dieën (2024) The reliability, validity, and responsiveness of tests used to assess the effects of power training in older adults: a systematic review, Physical Therapy Reviews, 29:4, 153-166, DOI: 10.1080/10833196.2024.2376439

To link to this article: https://doi.org/10.1080/10833196.2024.2376439

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

View supplementary material

Published online: 15 Jul 2024.

Submit your article to this journal

Article views: 1965

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=yptr20

PHYSICAL THERAPY REVIEWS
2024, VOL. 29, NO. 4, 153–166
https://doi.org/10.1080/10833196.2024.2376439

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

The reliability, validity, and responsiveness of tests used to assess the effects of power training in older adults: a systematic review

Mohamed el Hadouchi[a,b,c,d], Henri Kiers[a,b], Brittany Boerstra[d], Ralf Berends[a], Emma Schreuder[a], Cindy Veenhof[b,e] and Jaap van Dieën[c]

[a]Institute for Human Movement Studies, University of Applied Sciences Utrecht, Utrecht, the Netherlands; [b]Research Group Innovation of Human Movement Care, Expertise Center Healthy Urban Living, University of Applied Sciences Utrecht, Utrecht, the Netherlands; [c]Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands; [d]Sports & Science, Health Research Consultancy, Lochem, the Netherlands; [e]Department of Rehabilitation, Physical Therapy Science and Sport, Brain Center, University Medical Center Utrecht, Utrecht, the Netherlands

ABSTRACT

Background: Research shows that power training offers more potential for improving muscle power and physical performance in older adults than strength training. However, the measurement properties of the tests used to assess the effects of power training are unclear.

Objective: to review the reliability, validity, and responsiveness of tests used to measure the effects of power training in older adults.

Methods: A comprehensive literature search was conducted on 24 previously identified tests in PubMed, Embase, CINAHL, PsycInfo, and SPORTDiscus until April 29, 2024. The methodological quality of the studies was assessed using the COSMIN Risk of Bias checklist. Tests were categorized according to the International Classification of Functioning, Disability, and Health (ICF) and evaluated using Terwee's Modified Quality Criteria for Rating the Results of Measurement Properties.

Results: The search yielded a total of 74 articles, of which a majority had 'doubtful' or 'inadequate' methodological quality. Research on reliability was abundant and was considered high for a majority of tests, while validity and responsiveness were studied less. None of the included tests satisfied all criteria for Terwee's Checklist.

Conclusions: Aiming to cover each of the ICF domains, this review suggests the 1RM bench press, 1RM leg press, and CMJ for the function domain; and the 6-MWT, 10-MWT, timed stair climb, 5-STS, 30-seconds Sit to Stand, and TUG for the activities domain. No recommendations can be made for the participation domain at this time.

ARTICLE HISTORY
Received 4 May 2023
Accepted 1 July 2024

KEYWORDS
Power training; older adults; reliability; validity; responsiveness

Introduction

Muscle power is the product of muscle force and contraction velocity, and determines whether movements, like rising from a chair, can be performed [1–3]. Research shows that muscle power is a determinant of physical performance, and a critical variable in predicting mobility and physical functioning in older adults [4–7]. Aging is characterized by a loss of muscle power, which can make activities of daily life, such as rising from a chair or climbing a flight of stairs, more difficult to perform. Muscle power is best increased through power training [7], a form of exercise that involves moving resistance at a higher velocity than traditional strength training [5,6,8,9]. A recent systematic review suggests that power training offers more potential to improve muscle power and physical performance in older adults than strength training [10].

The effects of power training are commonly evaluated using performance-based tests administered by physical therapists or trainers, as these provide more accurate measures of performance than patient-reported outcome measures (PROMs) [11]. Tests can be categorized within the International Classification of Functioning, Disability, and Health (ICF) [12] muscle power can be considered a function, while physical performance belongs to the activity domain where a distinction can be made between generic tests and tests with an emphasis on movement speed. Lastly, physical functioning in daily life belongs to the participation domain [10]. A recent systematic review reported that while no

CONTACT Mohamed el Hadouchi mohamed.elhadouchi@hu.nl Institute for Human Movement Studies, University of Applied Sciences Utrecht, 3501 AA Utrecht, P.O. Box 12011, Utrecht, the Netherlands

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10833196.2024.2376439.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

reports on the effects of power training in the participation domain of the ICF were found, a wide variety of tests were used to evaluate the effects of power training in the function and activity domains of the ICF [10]. Overall, the results indicated a benefit of power training over strength training. However, the measurement properties of the tests used were unclear, making a comprehensive assessment of these tests warranted.

Therefore, the objective of this study was to review the reliability, validity, and responsiveness of tests used to measure the effects of power training in older adults. We specifically investigated the tests identified in a previous systematic review and meta-analysis of randomized controlled trials comparing the effectiveness of power training to strength training in older adults [10]. This research will support the development of a standardized testing protocol to assess the effects of power training in older adults, based on a set of reliable, valid, and responsive measurement tools.

Methods

This systematic review was developed in accordance with the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [13] and adhered to the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) [14] and the COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) [15–17]. This review was also registered in the International Prospective Register of Systematic Reviews (PROSPERO 2021: CRD42021273812).

Data sources and searches

A previous systematic review [10] identified 24 tests that were used to evaluate the effects of power training in older adults (Figure 1). Our aim was to assess the measurement properties of the tests used to measure the effects of power training in older adults in the above-mentioned review. Therefore, a comprehensive literature search was conducted in the bibliographic databases PubMed, Embase, CINAHL Plus (Ebsco), APA PsycInfo, and SPORTDiscus (Ebsco) from inception to April 29, 2024 in collaboration with a medical information specialist. The following terms were used (including synonyms and closely related words) as index terms of free-text words: 'functional strength measurements', 'muscular strength measurements', 'aged', 'tests', 'reliability', 'validity', 'responsiveness'. The search strategy for PubMed (Supplementary Appendix 1) was developed first and then translated for the remaining databases. Duplicate

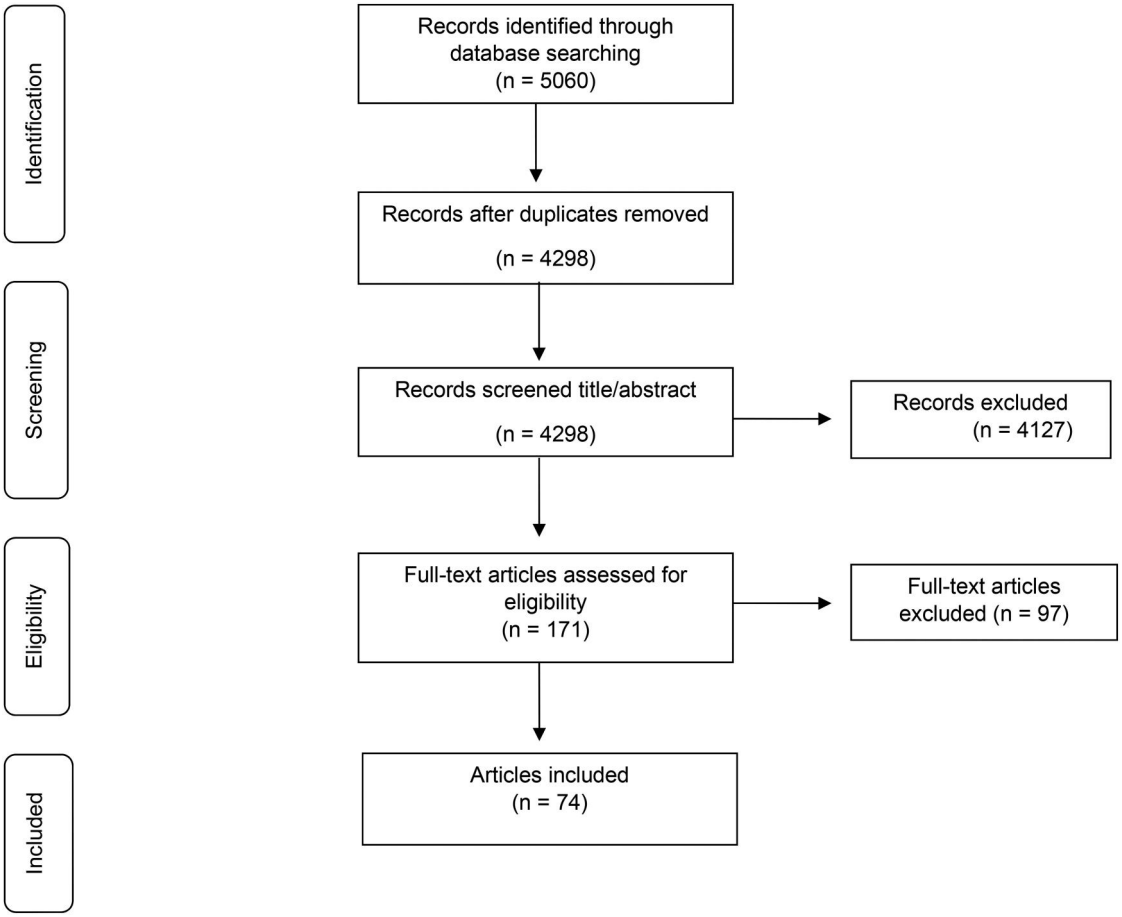

Figure 1. PRISMA flowchart of search and selection process.

articles were excluded by a medical information specialist using Endnote X21.0.1 (Clarivatetm), following the Amsterdam EfficientDeduplication (AED)-method [18] and the Bramer-method [19]. The references of the identified articles were also searched for relevant publications.

Study selection

Study selection was independently performed by four researchers (MeH, BB, ES, RB). First, all potentially relevant titles and abstracts were screened using the online software Rayyan [20]. Second, the remaining articles were read in entirety to ensure eligibility criteria were met. Differences in judgement were resolved through a consensus procedure.

Studies that contained data on the reliability, validity, or responsiveness of measurement of the instrument of interest were included if: (1) the study population consisted of older adults (mean age >65 years) recruited from a healthy population, regardless of their level of disability, physical functioning, or fitness. We defined healthy using the World Health Organization's (WHO) definition for health, in which individuals can be considered healthy despite the presence of (chronic) disease [21]; (2) studies evaluated the performance test or measurement tool itself, and not the equipment used to support the measurement (e.g. stopwatch or force plate); (3) data on measurement properties were available in the form of a statistical summary; (4) the article was written in English; and (5) a full-text version of the article was available.

Methodological quality of selected studies

The methodological quality of studies evaluating structural validity, internal constency, construct validity, or responsiveness were assessed via the COSMIN RoB checklist [15–17]. Box 3 was used for structural validity, box 4 for internal consistency, box 9a for construct validity, and box 10 for responsiveness. Although the COSMIN RoB checklist was primarily developed for patient-reported outcome measures (PROMs), it can also be used to evaluate performance-based outcome measurement instruments (PerFOMs) [14].

The methodological quality of studies evaluating reliability and measurement error was assessed *via* the extended COSMIN RoB designed for performance-based measurement isntruments (PerfOMs) [14]. In this extended COSMIN RoB Tool, both Part A: Elements of a comprehensive research question and Part B: Assessing risk of bias were applied independently by 3 researchers (BB, ES, RB). If a consensus could not be reached, a fourth researcher (MeH) was consulted whom made the final decision.

The assessment procedure for the COSMIN Risk of Bias checklist involves a 4-point scale scoring each item as very good, adequate, doubtful, or inadequate. The 'worst score counts' method was applied, meaning that the overall rating of the quality of each study is determined by the lowest rating on any standard in the box [15–17].

Measurement properties definitions

In the COSMIN taxonomy [22], reliability refers to the proportion of the total variance in the measurements which is due to true differences between individuals, and measurement error refers to the systematic and random error of an individual's score that is not attributed to true changes in the construct to be measured. Internal consistency evaluates the degree of interrelatedness among items for tests composed of several subscales. Within the validity domain, structural validity and construct validity were considered. Structural validity is the degree to which the test results are an adequate reflection of the construct to be measured, while construct validity, reflects how well a test compares with another test measuring the same construct. Construct validity was used instead of criterion validity, the degree to which the score on a test is an adequate reflection of a gold standard, since there is no generally accepted standard for evaluating power training in older adults. Finally, responsiveness is the ability of a measurement instrument to detect a change over time in the construct to be measured.

Data extraction

Data and study characteristics were extracted from each of the included studies and manually entered into Microsoft Excel by three reviewers (BB, ES, RB). Data for reliability and responsiveness are summarized in Supplementary Appendix 2 and validity results are summarized in Supplementary Appendix 3. The results of the studies were then classified as evidence for reliability, validity, and responsiveness in Table 1. For the sake of clarity and consistency, measurement properties were recategorized to what was factually measured according to the COSMIN Taxonomy of Measurement Properties. Measurement properties that did not fit any of the COSMIN definitions for measurement properties were excluded.

Rating of measurement properties

Each test was given a quality rating based on the number and methodological quality of included

Table 1. Characteristics of the included studies.

Study #	Author and year	COSMIN domain	Measurement property	COSMIN RoB rating
1	Adell et al. (2013) [23]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
2	Alfonso-Rosa et al. (2014) [24]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
3	Amarante do Nascimento et al. (2013) [25]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
4	Barbalho et al. (2018) [26]	Reliability	Reliability Measurement error	Doubtful
5	Batting et al. (2019) [27]	Reliability	Measurement error	Inadequate
		Validity	Construct validity: comparison with another instrument	Inadequate
6	Beauchamp et al. (2012) [28]	Reliability	Reliability Measurement error	Adequate
7	Beninato et al. (2016) [29]	Validity	Structural validity	Adequate
8	Bergamin et al. (2017) [30]	Reliability	Reliability	Doubtful
9	Bodilsen et al. (2015) [31]	Reliability	Reliability	Adequate
			Measurement error	Adequate
10	Brooks et al. (2006) [32]	Validity	Construct validity: comparison with another instrument	Inadequate
		Responsiveness	Responsiveness	Doubtful
11	Chan et al. (2020) [33]	Reliability	Reliability	Doubtful
		Validity	Construct validity: comparison with another instrument	Inadequate
12	Choi et al. (2014) [34]	Reliability	Reliability	Very good
			Measurement error	Very good
13	Connelly et al. (1996) [35]	Reliability	Reliability	Inadequate
14	Cress et al. (1996) [36]	Reliability	Reliability	Inadequate
			Internal consistency	Very good
		Validity	Construct validity: comparison with another instrument	Doubtful
15	Ditroilo et al. (2011) [37]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
16	Ellis et al. (2019) [38]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
17	Fonseca et al. (2014) [39]	Reliability	Reliability	Very good
			Measurement error	Adequate
18	Freire et al. (2012) [40]	Reliability	Reliability	Doubtful
19	Frykholm et al. (2019) [41]	Reliability	Reliability	Very good
			Measurement error	Very good
20	Galhardas et al. (2020) [42]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
21	Germanou et al. (2007) [43]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
22	Goldberg et al. (2012) [44]	Reliability	Reliability	Inadequate
			Measurement error	Inadequate
		Validity	Construct validity: comparison with another instrument	Very good
23	Gómez et al. (2013) [45]	Reliability	Reliability	Doubtful
		Validity	Structural validity	Doubtful
24	Hansen et al. (2018) [46]	Reliability	Reliability	Very good
			Measurement error	Very good
25	Hernandes et al. (2011) [47]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
26	Hoeymans et al. (1997) [48]	Reliability	Reliability	Doubtful
27	Hwang et al. (2016) [49]	Reliability	Reliability	Inadequate
			Measurement error	Inadequate
		Validity	Construct validity: comparison with another instrument	Inadequate
28	Işik et al. (2015) [50]	Reliability	Reliability	Inadequate
		Validity	Construct validity: comparison with another instrument	Inadequate
29	Jenkins et al. (2017) [51]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
30	Kristensen et al. (2019) [52]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
31	Labadessa et al. (2016) [53]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
32	LeBrasseur et al. (2008) [54]	Reliability	Reliability	Inadequate
			Measurement error	Inadequate
33	Lin et al. (2004) [55]	Reliability	Reliability	Inadequate
			Construct validity: comparison with another instrument	Adequate
		Responsiveness	Responsiveness	Doubtful
34	Madsen et al. (1996) [56]	Reliability	Measurement error	Inadequate
35	Marques et al. (2016) [57]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
36	Marques et al. (2021) [58]	Validity	Structural validity	Adequate
			Construct validity: comparison with another instrument	Doubtful
37	McLay et al. (2020) [59]	Validity	Construct validity: comparison with another instrument	Adequate
38	Medina-Mirapeix et al. (2016) [60]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
39	Medina-Mirapeix et al. (2018) [61]	Reliability	Reliability	Inadequate
			Measurement error	Inadequate
		Validity	Construct validity: comparison with another instrument	Inadequate
40	Mkacher et al. (2017) [62]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
41	Morris-Chatta et al. (1994) [63]	Reliability	Reliability	Inadequate
42	Nunes et al. (2020) [64]	Reliability	Reliability	Inadequate

(*continued*)

Table 1. Continued.

Study #	Author and year	COSMIN domain	Measurement property	COSMIN RoB rating
			Measurement error	Inadequate
		Validity	Construct validity: comparison with another instrument	Very good
43	Ordway et al. (2006) [65]	Reliability	Measurement error	Doubtful
44	Östchega et al. (2000) [66]	Reliability	Reliability	Inadequate
45	Özden et al. (2020) [67]	Reliability	Reliability	Inadequate
			Measurement error	Inadequate
		Validity	Construct validity: comparison with another instrument	Inadequate
46	Parraca et al. (2022) [68]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
47	Peel et al. (2001) [69]	Reliability	Reliability	Doubtful
48	Phillips et al. (2004) [70]	Reliability	Measurement error	Doubtful
49	Rikli et al. (1998) [71]	Reliability	Reliability	Doubtful
		Validity	Construct validity: comparison with another instrument	Very good
50	Rockwood et al. (2000) [72]	Reliability	Reliability	Inadequate
		Validity	Construct validity: comparison with another instrument	Inadequate
51	Rodrigues et al. (2023) [73]	Reliability	Reliability	Inadequate
52	Rolenz et al. (2016) [74]	Validity	Construct validity: comparison with another instrument	Adequate
53	Rydwik et al. (2007) [75]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
54	Saint-Maurice et al. (2020) [76]	Reliability	Reliability	Doubtful
55	Schaubert et al. (2005) [77]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
56	Schaubert et al. (2005) [78]	Reliability	Reliability	Doubtful
		Validity	Construct validity: comparison with another instrument	Adequate
57	Schroeder et al. (2007) [79]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
58	Schwenk et al. (2012) [80]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
59	Sherwood et al. (2019) [81]	Validity	Construct validity: comparison with another instrument	Inadequate
60	Suwit et al. (2020) [82]	Reliability	Reliability	Very good
			Measurement error	Very good
61	Suzuki et al. (2019) [83]	Reliability	Reliability	Doubtful
			Measurement error	Inadequate
62	Symons et al. (2004) [84]	Reliability	Reliability	Inadequate
			Measurement error	Inadequate
63	Symons et al. (2005) [85]	Reliability	Reliability	Inadequate
			Measurement error	Inadequate
64	Tolk et al. (2019) [86]	Reliability	Reliability	Doubtful
		Validity	Construct validity: comparison with another instrument	very good
		Responsiveness	Responsiveness	Very good
65	Tolk et al. (2019) [87]	Reliability	Reliability	Doubtful
		Validity	Construct validity: comparison with another instrument	Inadequate
		Responsiveness	Responsiveness	Inadequate
66	Topp et al. (1994) [88]	Reliability	Reliability	Doubtful
67	Uszko-Lencer et al. (2017) [89]	Reliability	Reliability	Doubtful
		Validity	Construct validity: comparison with another instrument	Inadequate
68	Van Driessche et al. (2018) [90]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
69	van Iersel et al. (2008) [91]	Reponsiveness	Responsiveness	Doubtful
70	Verdijk et al. (2009) [92]	Validity	Construct validity: comparison with another instrument	Adequate
71	Wallmann et al. (2013) [93]	Reliability	Reliability	Doubtful
72	Webber et al. (2010) [94]	Reliability	Reliability	Doubtful
			Measurement error	Doubtful
73	Whitney et al. (2005) [95]	Validity	Construct validity: comparison with another instrument	Adequate
74	Wrisley et al. (2010) [96]	Validity	Construct validity: comparison with another instrument	Very good

studies using Terwee et al.'s Modified Quality Criteria for Rating the Results of Measurement Properties checklist [97]. This checklist was modified from its original purpose of evaluating health status questionnaires in collaboration with its authors, to provide explicit criteria to critically appraise and score measurement properties of performance tests and tools. This tool rates criteria related to the validity, reliability, and responsiveness as positive (+), negative (−), or undetermined (?), which are then used to determine the overall level of evidence for the quality of measurement properties. The level of evidence was rated as 'strong' when a performance test or tool scored three positives (+++) or three negatives (−−), as 'moderate' with 2 positives (++) or 2 negatives (-), as 'limited' with one positive (+) or negative (−), as 'conflicting' with a mixture of positives (+) and negatives (−), and as 'unknown' when no studies were available.

Within the validity domain of Terwee's checklist, columns for content validity and cross-cultural validity were deemed irrelevant as these forms of validity do not apply to performance-based outcome

measurements, and thus scored as 'not applicable'. Criterion validity was also scored as 'not applicable', because there is no accepted 'golden standard' to which the tests can be compared.

Results

Search and study selection

The literature search generated a total of 5060 references: 1266 in PubMed, 1393 in Embase, 1903 in CINAHL, 295 in APA PsycInfo and 203 in SPORTDiscus. Reference checking did not lead to any further unique articles. Following the removal of duplicates, 4298 references were screened based on title and abstract, of which 171 references continued to full-text screening. During this phase, 97 articles were excluded on the basis of eligibility criteria, leaving a total of 74 articles included in the present review. Figure 1 summarizes the search and selection process in a PRISMA flow diagram.

The 74 included studies evaluated measurement properties for 15 of the pre-identified 24 tests, as shown in Figure 2. Nine of the 15 tests showed promising results and are covered in this review in greater detail. Within the reliability domain, 60 of the included studies evaluated reliability, 42 studies evaluated measurement error, and only one study evaluated the internal consistency of tests. Within the validity domain, 24 studies evaluated construct validity and five studies evaluated structural validity. The responsiveness of tests was evaluated in five studies.

Methodological quality of selected studies

An overview of the methodological quality of the included studies according to the COSMIN Risk of Bias checklist is given in Table 1. Of the 60 studies that reported reliability, five studies had very good methodological quality, two studies were adequate, 38 studies were doubtful, and 15 studies were inadequate. Of the 42 studies that reported on measurement error, four studies had very good methodological quality, three studies were adequate, 24 studies were doubtful, and 11 studies were inadequate. The high number of 'doubtful' ratings was mainly due to a lack of reporting on whether the individuals administering and scoring the performance test were blinded to previous results (standards 4 and 5 in the extended COSMIN RoB tool to assess the quality of studies on reliability and measurement error). The single study that evaluated internal consistency had very good methodological quality [36].

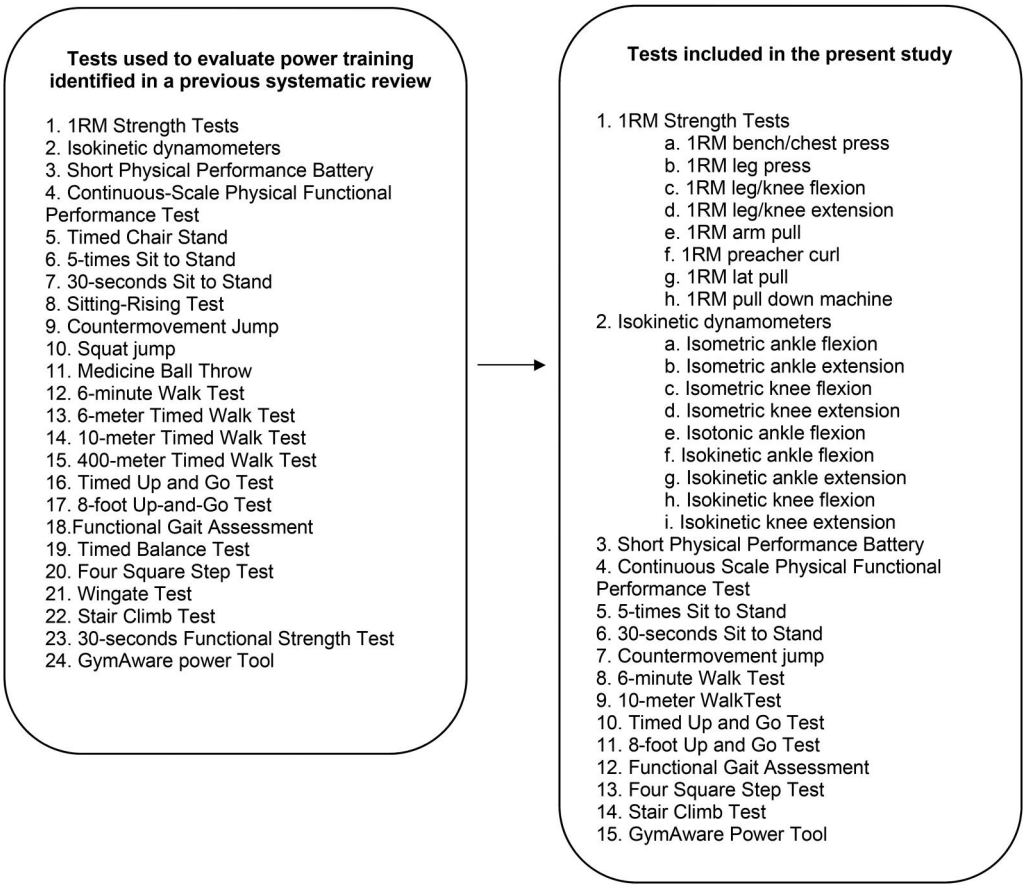

Figure 2. Previously-identified tests used to evaluate power training and tests included in the present study.

Within the validity domain, five studies evaluated structural validity, of which two studies had adequate methodological quality and three studies were rated as doubtful as a result of unclear rotation methods in the factor analysis. Of the 24 studies that evaluated construct validity, five studies had very good methodological quality, six studies were adequate, two studies were doubtful, and 11 studies were inadequate. All of the studies that were declared inadequate lacked information on the measurement properties of the comparative test. Of the five studies that reported responsiveness, one study had good methodological quality, three studies were doubtful, and one study was inadequate.

Rating of measurement properties

The overall rating of each test according to the Terwee criteria [97] is summarized in Table 2. On the basis of these findings, the 1RM bench press, 1RM leg press, 5-times Sit to Stand (5-STS), 30-seconds Sit to Stand, countermovement jump, 6-min Walk Test (6-MWT), 10-meter walk test (10-MWT), Timed Up and Go (TUG) test, and Stair Climb Test are recommended for use in the evaluation of power training effects in older adults. The transfer to activities of daily life may be more problematic for isolated movements than for functional movements as the latter more closely mimic real-life movements. Even so, isolated movement testing, such as muscle strength testing for specific muscle groups, had relatively high reliability and validity. In contrast, functional testing can be more difficult to design and may have more varied results. For example, a simple functional test like a chair stand test can be limited by lower body strength or power, but it does predict physical performance in older adults. The Short Physical Performance Battery (SPPB) is excluded from the recommendation despite its positive rating because this test has a ceiling effect.

Reliability

All of the tests recommended to evaluate the effects of power training in older adults met Terwee's criteria for strong positive reliability (3 or more studies with ICC/weighted kappa >0.70 or Pearson's $r > 0.80$) [97] Despite this, three studies were found in which several recommended tests showed moderate reliability. In Ostchega et al. [66], the 5-STS had a test-retest ICC of 0.64, and in Rockwood et al. [72] the TUG had a test-retest ICC of 0.56. Moreover, the peak power/body mass (W/Kg) estimated from a countermovement jump showed moderate reliability between sessions (ICC: 0.62) for men [37]

None of the recommended tests met Terwee's criteria for measurement error (MIC > SDC or MIC outside of LOA) as the minimal important important change (MIC) was not reported in any of the studies. The studies quantified measurement error using the standard error of measurement (SEM), smallest detectable change (SDC), limits of agreement (LoA), and the coefficient of variation (CV). Note that we also used the term SDC where original studies used minimal detectable change (MDC), minimal detectable difference (MDD), and smallest real difference (SRD), as these are calculated in the same way. As a result, the rating for measurement error was 'unknown' for all tests, but in an effort to be more forthcoming in the findings, these statistical results will be discussed below. The TUG had consistently low measurement errors for both test-retest (SEM: 0.04–1.20; SDC: 0.37–3.01; LoA: 0.02–0.05) and inter-rater measurement error (LoA: 0.06). Measurement errors were also small but less consistent for the 10-MWT (SEM: 0.14 s, SDC95: 0.13–0.40 s), timed stair climb (SEM: 1.06–1.98 s), and 5STS (SEM: 0.76–1.98; SDC95: 2.10–2.91). Values were slightly higher for the CMJ (SDC: 3.2–444.7; SDC%: 14.9%–19.3%) and the 30-sec STS for both test-retest (SEM: 0.85–1.32; SEM%: 18%; SDC: 2.68–3.67; SDC%: 27.0%–49.0%) and inter-rater measurement error (SDC: 567; SDC%: 16%). The 6-MWT had conflicting results, with some studies reporting small measurement errors (SEM: 9.88–14.84 s, 2.5%–5%), while other studies reported large measurement errors (SEM: 31.0–34.0 s; LoA: −67.0 to 120.0). Measurement errors were also conflicting for 1RM strength tests, ranging from small errors for the bench press (SEM: 1.71; LoA: −48.34 to 36.45; %CV: 4.9–7.9), to relatively large errors for the leg press (LoA: −371.42 to 236.10; %CV: 4.2–6.3).

Internal consistency was not assessed in any of the recommended tests, only in the CS-PFP which had a limited positive rating for internal consistency (Cronbach's alpha: 0.96–0.97).

Validity

Structural validity was not assessed in any of the recommended tests, only in the Functional Gait Assessment (FGA) test and the SPPB. The FGA had a limited positive rating for structural validity because a single study explained 53.3% of variance through a factor analysis [58], while the rating for SPPB was unknown because the explained variance was not reported.

Construct validity was assessed more frequently. Of the recommended tests, the 1RM bench press, 1RM leg press, 6-MWT, and 10-MWT had a limited positive rating for construct validity (1 study in which correlation with an instrument measuring the

Table 2. Rating of measurement properties using terwee's criteria.

		Reliability			Validity					Responsiveness
#	Measurement instrument	Reliability	Measurement error	Internal consistency	Content validity	Structural validity	Construct validity*	Cross-cultural validity	Criterion validity	Responsiveness
1a	1RM bench/chest press	+ + +	?	N/A	N/A	?	+	N/A	N/A	?
1b	1RM leg/knee flexion	+ + +	?	N/A	N/A	?	?	N/A	N/A	?
1c	1RM leg/knee extension	+	?	N/A	N/A	?	+ +	N/A	N/A	?
1d	1RM leg press	+ + +	?	N/A	N/A	?	+	N/A	N/A	?
1e	1RM arm curl	+	?	N/A	N/A	?	?	N/A	N/A	?
1f	1RM preacher curl	+	?	N/A	N/A	?	+	N/A	N/A	?
1g	1RM lat pull	+	?	N/A	N/A	?	?	N/A	N/A	?
1h	1RM pull down machine	?	?	N/A	N/A	?	?	N/A	N/A	?
2a	Isometric ankle flexion	+	?	N/A	N/A	?	?	N/A	N/A	?
2b	Isometric ankle extension	?	?	N/A	N/A	?	?	N/A	N/A	?
2c	Isometric knee flexion	?	?	N/A	N/A	?	?	N/A	N/A	?
2d	Isometric knee extension	+ + +	?	N/A	N/A	?	?	N/A	N/A	?
2e	Isotonic ankle flexion	+	?	N/A	N/A	?	?	N/A	N/A	?
2f	Isokinetic ankle flexion	+ + −	?	N/A	N/A	?	?	N/A	N/A	?
2g	Isokinetic ankle extension	?	?	N/A	N/A	?	?	N/A	N/A	?
2h	Isokinetic leg/knee flexion	+ + +	?	N/A	N/A	?	?	N/A	N/A	?
2i	Isokinetic leg/knee extension	+ + +	?	N/A	N/A	?	+	N/A	N/A	?
3	Short Physical Performance Battery	+ + +	?	N/A	N/A	?	?	N/A	N/A	?
4	Continuous Scale Physical Functional Performance	+	?	+	N/A	?	?	N/A	N/A	?
5	5-times Sit to Stand	+ + +	?	N/A	N/A	?	?	N/A	N/A	?
6	30-sec Sit to Stand	+ + +	?	N/A	N/A	?	?	N/A	N/A	−
7	Countermovement jump	+ + +	?	N/A	N/A	?	?	N/A	N/A	?
8	6-min Walk Test	+ + +	?	N/A	N/A	?	+	N/A	N/A	?
9	10-meter Walk Test	+ + +	?	N/A	N/A	?	+	N/A	N/A	?
10	Timed Up and Go	+ + +	?	N/A	N/A	?	+ + -	N/A	N/A	?
11	8-foot Up and Go	+	?	N/A	N/A	?	?	N/A	N/A	?
12	Functional Gait Assessment	?	?	N/A	N/A	+	+ +	N/A	N/A	?
13	Four Square Step Test	+	?	N/A	N/A	?	+ −	N/A	N/A	?
14	Stair Climb Test	+ + +	?	N/A	N/A	?	?	N/A	N/A	−
15	Gymaware power tool	?	?	N/A	N/A	?	+	N/A	N/A	?

The level of evidence was rated as 'strong' when a performance test or tool scored three positives (+++) or three negatives (−−), as 'moderate' with 2 positives (++) or 2 negatives (−), as 'limited' with one positive (+) or negative (−), as 'conflicting' with a mixture of positives (+) and negatives (−), and as 'unknown' when no studies were available. Abbreviations used: 1RM: 1-repetition maximum.

same construct was >0.5). The TUG had a conflicting rating for construct validity, meaning that some studies found correlations >0.5 and some did not.

Responsiveness

The Terwee criterion for responsiveness was: (a) correlation with an instrument measuring the same construct is >0.5; or (b) at least 75% of the results are in accordance with the hypotheses; or (c) AUC >0.7 and correlation with related constructs is higher than with unrelated constructs. None of the recommended tests met this criterion. In fact, responsiveness was only assessed in the 30-seconds Sit to Stand, the Stair Climb Test, and the TUG. The 30-seconds Sit to Stand fulfilled 44% of a priori hypotheses, while the Stair Climb Test fulfilled 36% of a priori hypotheses. With regard to the TUG, responsiveness was quantified using a standardized response mean (SRM) and responsiveness index (RI): Brooks et al. [32] reported an SRM of 1.1 while van Iersel et al. [91] found an SRM of 0.9 (1.1%), and RI of 1.9 (2.6%), and an effect size of 0.1. As a result, the 30-seconds Sit to Stand, the Stair Climb Test, and the TUG all received a limited negative rating for responsiveness.

Discussion

A previous systematic review showed that a wide variety of different tests and outcome measures have been used interchangeably without distinguishing between various domains [10]. This review categorized 24 previously-identified tests using the ICF framework to provide a structured and in-depth analysis of the tests used to measure the effects of power training. Of the 24 previously identified tests, literature on measurement properties was available for only 15 tests. Our findings suggest that while a majority of tests exhibited good reliability (ICC or correlation >0.7) [97] the validity and responsiveness of each of the tests were poorly described or not investigated. Thus, it remains unclear whether the tests truly measure the construct or outcome they are intended to measure and whether the tests are able to pick up a change in performance over time. The methodological quality of the majority of the studies reporting on measurement properties was doubtful or insufficient as well, possibly because several studies had been conducted prior to the publication of the guidelines on conducting and reporting such studies. Based on these findings, a recommendation of tests used to measure the effects of power training in older adults can be made for the time being. Aiming to cover each of the ICF domains, this review suggests the 1RM bench press, 1RM leg press, and the CMJ for the function domain; and the 6-MWT, 10-MWT, the Stair Climb Test, 5-STS, 30-seconds Sit to Stand, and the TUG for the activity domain. No recommendations can be provided for the participation domain based on this review. Several of these tests have been recommended in other populations as well. In patients with hip and knee arthritis, the 30-seconds Sit to Stand and TUG were the best rated Sit to Stand tests [98] and the 6-MWT was recommended to assess the functional status of patients with chronic obstructive pulmonary disorder (COPD) [99] Overall, further research is needed to provide information on the tests used to measure the effects of power training in older adults.

Measurement properties are not the sole factor that determine which tests are appropriate when evaluating the effects of power training in older adults. Specific to measuring muscle power, it is crucial to critically assess whether the tests reflects the construct of muscle power. Muscle power is the product of muscle force and contraction velocity, and thus, tests challenging the force and velocity component of the movement simultaneously should be utilized. Commonly used strength tests, such as isometric or 1RM strength tests, capture muscle force but do not take the velocity of movement into account. While force or velocity alone can be used as a proxy for power, proxies are rarely as accurate as actual measurements and effect estimates based on proxies may limit our understanding of improvements in power and resulting effects on performance with training.

Additionally, in tests that utilize a scoring system, it is important to be aware of a ceiling effect that may prohibit an accurate effect estimation. The SPPB, for example, is a generic measurement instrument with scores ranging from 0 to 12 or 16 points, and has been used to make an inference about physical performance or functional status in previous research. However, it is difficult to measure the effect of the intervention if a participant already scores a 9 at baseline. For this reason, the SPPB was not included in the list of recommended tests.

To the best of our knowledge, this is the first systematic review to summarize the measurement properties of tests used to evaluate the effects of power training in older adults. We included a wide range of studies while rigorous assessment criteria in the COSMIN RoB ensured that high methodological quality was pursued throughout this review. This review evaluated tests that were categorized within the function and activities domain of the ICF, which provided a more complete picture of the tests and a more in-depth understanding of how these tests were used. However, for results to be

most relevant to older adults, it is important that changes also occur within the participation domain. As such, the effects of power training have been measured by performance tests that evaluate an individual's ability to perform an activity or movement within a clinical or research setting. Yet, researchers are often actually interested whether the effect of power training translates into increased participation and increased independence. In short, there is a discrepancy between the outcome that is currently being measured by tests (function and activity) and the outcome that we are actually interested in (participation). Future research into the effects of power training should employ methods for outcome assessment within a participant's home environment to see how improved capacity translate to increased participation in daily tasks and activities. Tests or devices that are able to measure the change of velocity, such as accelerometers, can provide a solution for measuring within the participation domain. Accelerometers were not included in the present review because these devices have not yet been used to evaluate the effects of power training in older adults. Covering each of the ICF domains will make it possible to more accurately determine whether power training is effective and to which ICF domain this effect relates. Further research is needed to gain expert consensus on the categorization of tests within ICF domains and evaluate whether there are other or additional tests that should be use to measure the effects of power training in older adults.

Limitations

This review encountered several limitations. While the use of the ICF framework facilitated specific insights into the effects of power training on various domains of the ICF, categorizing tests into ICF domains can be considered artificial as some tests may cover multiple domains. Additionally, some tests that are relevant for evaluating the effects of power training may not fully be covered by the ICF categories. These limitations may infleunce the interpretation and generalization of study findings. With regard to the availability of relevant literature, tests that performed best were studied most often while tests that performed poorly were studies less often. This may not necessarily reflect the measurement properties of the test, but could also be the result of publication bias. It is important to differentiate between tests that scored low as a result of inadequate measurement properties and tests that scored low as a result of insufficient research. In the data extraction process, the type and amount of data on measurement properties varied largely between studies, making comparibility between studies more difficult. For example, SEM estimates were not providedly consistently in each study, and if they were reported, confidence intervals were often omitted. Furthermore, the present study focused on the tests used to evaluate power training in older adults and only included literature in which measurement properties were evaluated in older adults as well. This may have contributed to a lack of research for some tests. It is likely that more literature is available on measurement properties of the included tests applied to younger populations, but the question remains how transferable these results are to older populations. Lastly, we acknowledge that the PROSPERO registration of this study occurred after the first systematic search, which ideally, would have taken place before the systematic search.

Conclusions

Aiming to cover each of the ICF domains, this review suggest the 1RM bench press, 1RM leg press, and CMJ for the function domain; and the 6-MWT, 10-MWT, timed stair climb, 5-STS, 30-seconds Sit to Stand, and TUG for the activities domain. No recommendations can be made for the participation domain at this time.

Acknowledgements

We would like to thank Annemarie van der Velden and Sjoerd Beelen from the University of Applied Sciences Utrecht, and Ralph de Vries from Vrije Universiteit Amsterdam for their efforts and expertise regarding the search strategy. Additionally, we would like to recognize all supporting staff at Health Research Consultancy (HRC) in Lochem, the Netherlands.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Dutch Research Council under a grant (NWO 2017/BOO/00279639).

ORCID

Mohamed el Hadouchi http://orcid.org/0000-0001-9501-6408

Data availability statement

data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. Strotmeyer ES, Winger ME, Cauley JA, et al. Normative values of muscle power using force plate jump tests in men aged 77-101 years: the osteoporotic fractures in men (MROS) study. J Nutr Health Aging. 2018;22(10):1167–1175. doi: 10.1007/s12603-018-1081-x.
2. Strollo SE, Caserotti P, Ward RE, et al. A review of the relationship between leg power and selected chronic disease in older adults. J Nutr Health Aging. 2015;19(2):240–248. doi: 10.1007/s12603-014-0528-y.
3. Stephenson ML, Smith DT, Heinbaugh EM, et al. Total and lower extremity lean mass percentage positiveky correlates with jump performance. J Strength Cond Res. 2015;29(8):2167–2175. doi: 10.1519/JSC.0000000000000851.
4. de Vos NJ, Singh NA, Ross DA, et al. Optimal load for increasing muscle power during explosive resistance training in older adults. J Gerontol A Biol Sci Med Sci. 2005;60(5):638–647. doi: 10.1093/gerona/60.5.638.
5. Balachandran A, Krawczyk SN, Potiaumpai M, et al. High-speed circuit training vs hypertrophy training to improve physical function in sarcopenic obese adults: a randomized controlled trial. Exp Gerontol. 2014;60:64–71. doi: 10.1016/j.exger.2014.09.016.
6. Marsh AP, Miller ME, Rejeski WJ, et al. Lower extremity muscle function after strength or power training in older adults. J Aging Phys Act. 2009;17(4):416–443. doi: 10.1123/japa.17.4.416.
7. Fielding R A, Lebrasseur NK, Cuoco A, et al. High-velocity resistance training increases skeletal muscle peak. J Am Geriatr Soc. 2002;50:655–662.
8. Miszko T A, Cress ME, Slade JM, et al. Effect of strength and power training on physical function in community-dwelling older adults. J Gerontol A Biol Sci Med Sci. 2003;58(2):171–175. doi: 10.1093/gerona/58.2.m171.
9. Tiggemann CL, Dias CP, Radaelli R, et al. Effect of traditional resistance and power training using rated perceived exertion for enhancement of muscle strength, power, and functional performance. Age. 2016;38(2):42. doi: 10.1007/s11357-016-9904-3.
10. el Hadouchi M, Kiers H, de Vries R, et al. Effectiveness of power training compared to strength training in older adults: a systematic review and meta-analysis. Eur Rev Aging Phys Act. 2022;19(1):18. doi: 10.1186/s11556-022-00297-x.
11. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. New Jersey: Pearson/Prentice Hall; 2009.
12. World Health Organization. International classification of functioning, disability and health. World report on child injury prevention. WHO: Geneva; 2001.
13. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi: 10.1136/bmj.n71.
14. Mokkink LB, Boers M, van der Vleuten CPM, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol. 2020;20(1):293. doi: 10.1186/s12874-020-01179-5.
15. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1171–1179. doi: 10.1007/s11136-017-1765-4.
16. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147–1157. doi: 10.1007/s11136-018-1798-3.
17. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159–1170. doi: 10.1007/s11136-018-1829-0.
18. Otten R, de Vries R, Schoonmade L. Amsterdam Efficient Deduplication (AED) method. Zenodo. 2019.
19. Bramer WM, Giustini D, De Jonge GB, et al. De-duplication of database search results for systematic reviews in EndNote. JMLA. 2016;104(3):240–243. doi: 10.5195/jmla.2016.24.
20. Ouzzani M, Hammady H, Fedorowicz Z, et al. Rayyan. Syst Rev. 2016;5.
21. International Health Conference. Constitution of the World Health Organization. 1946. Bull World Health Organ. 2002;80(12):983–984.
22. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–745. doi: 10.1016/j.jclinepi.2010.02.006.
23. Adell E, Wehmhörner S, Rydwik E. The test-retest reliability of 10 meters maximal walking speed in older people living in a residential care unit. J Geriatr Phys Ther. 2013;36(2):74–77. doi: 10.1519/JPT.0b013e318264b8ed.
24. Alfonso-Rosa RM, Del Pozo-Cruz B, Del Pozo-Cruz J, et al. Test-retest reliability and minimal detectable change scores for fitness assessment in older adults with type 2 diabetes. Rehabil Nurs. 2014;39(5):260–268. doi: 10.1002/rnj.111.
25. Nascimento MADO, Januario RSB, Gerage AM, et al. Familiarization and reliability of one repetition maximum strength testing in older women. J Aging Health. 2013;27:1636–1642.
26. Barbalho M, Gentil P, Raiol R, et al. High 1RM tests reproducibility and validity are not dependent on training experience, muscle group tested or strength level in older women. Sports. 2018;6(4):171. doi: 10.3390/sports6040171.
27. Batting M, Barker KL. Reliability and validity of the Four Square Step Test in patients with hip osteoarthritis before and after total hip replacement. Physiotherapy. 2019;105(2):244–253. doi: 10.1016/j.physio.2018.07.014.
28. Beauchamp MK, Hao Q, Kuspinar A, et al. Reliability and minimal detectable change values for performance-based measures of physical functioning in the canadian longitudinal study on aging. J Gerontol A Biol Sci Med Sci. 2021;76(11):2030–2038. doi: 10.1093/gerona/glab175.
29. Beninato M, Ludlow LH. The functional gait assessment in older adults: validation through Rasch modeling. Phys Ther. 2016;96(4):456–468. doi: 10.2522/ptj.20150167.

30. Bergamin M, Gobbo S, Bullo V, et al. Reliability of a device for the knee and ankle isometric and isokinetic strength testing in older adults. Muscles Ligaments Tendons J. 2017;7(2):323–330. doi: 10.11138/mltj/2017.7.2.323.
31. Bodilsen AC, Juul-Larsen HG, Petersen J, et al. Feasibility and inter-rater reliability of physical performance measures in acutely admitted older medical patients. PLOS One. 2015;10(2):e0118248. doi: 10.1371/journal.pone.0118248.
32. Brooks D, Davis AM, Naglie G. Validity of 3 physical performance measures in inpatient geriatric rehabilitation. Arch Phys Med Rehabil. 2006;87(1):105–110. doi: 10.1016/j.apmr.2005.08.109.
33. Chan WLS, Pin TW. Reliability, validity and minimal detectable change of 2-min walk test and 10-m walk test in frail older adults receiving day care and residential care. Aging Clin Exp Res. 2020;32(4):597–604. doi: 10.1007/s40520-019-01255-x.
34. Choi YM, Dobson F, Martin J, et al. Interrater and intrarater reliability of common clinical standing balance tests for people with hip osteoarthritis. Phys Ther. 2014;94(5):696–704. doi: 10.2522/ptj.20130266.
35. Connelly DM, Stevenson TJ, Vandervoort AA. Between- and within-rater reliability of walking tests in a frail elderly population. Physiotherapy Canada. 1996;48:47–51.
36. Cress ME, Buchner DM, Questad KA, et al. Continuous-scale physical functional performance in healthy older adults: a validation study. Arch Phys Med Rehabil. 1996;77(12):1243–1250. doi: 10.1016/s0003-9993(96)90187-2.
37. Ditroilo M, Forte R, McKeown D, et al. Intra- and inter-session reliability of vertical jump performance in healthy middle-aged and older men and women. J Sports Sci. 2011;29(15):1675–1682. doi: 10.1080/02640414.2011.614270.
38. Ellis R, Holland AE, Dodd K, et al. Reliability of one-repetition maximum performance in people with chronic heart failure. Disabil Rehabil. 2019;41(14):1706–1710. doi: 10.1080/09638288.2018.1443160.
39. F AA, G CLA, C MCB, et al. Reproducibility of a test for the functional evaluation of dynamic balance and agility in elderly people. Iatreia. 2014;27:290–298.
40. Cremonese C, Freire C, Meyer A, et al. Pesticide exposure and adverse pregnancy events, Southern Brazil, 1996-2000. Cad Saude Publica. 2012;28(7):1263–1272. doi: 10.1590/s0102-311x2012000700005.
41. Frykholm E, Géphine S, Saey D, et al. Inter-day test–retest reliability and feasibility of isokinetic, isometric, and isotonic measurements to assess quadriceps endurance in people with chronic obstructive pulmonary disease: a multicenter study. Chron Respir Dis. 2019;16:1479973118816497. doi: 10.1177/1479973118816497.
42. Galhardas L, Raimundo A, Marmeleira J. Test-retest reliability of upper-limb proprioception and balance tests in older nursing home residents. Arch Gerontol Geriatr. 2020;89:104079. doi: 10.1016/j.archger.2020.104079.
43. Germanou E, Beneka A, Malliou P, et al. Reproducibility of concentric isokinetic strength of the knee extensors and flexors in individuals with mild and moderate osteoarthritis of the knee. IES. 2007;15(3):151–164. doi: 10.3233/IES-2007-0261.
44. Goldberg A, Chavis M, Watkins J, et al. The five-times-sit-to-stand test: validity, reliability and detectable change in older females. Aging Clin Exp Res. 2012;24(4):339–344. doi: 10.1007/BF03325265.
45. Gómez Montes JF, Curcio CL, Alvarado B, et al. Validity and reliability of the short physical performance battery (SPPB): a pilot study on mobility in the Colombian Andes. Colomb Med. 2013;44:165–171. doi: 10.25100/cm.v44i3.1181.
46. Hansen H, Beyer N, Frølich A, et al. Intra-and inter-rater reproducibility of the 6-minute walk test and the 30-second sit-to-stand test in patients with severe and very severe COPD. Int J Chron Obstruct Pulmon Dis. 2018;13:3447–3457. doi: 10.2147/COPD.S174248.
47. Hernandes NA, Wouters EFM, Meijer K, et al. Reproducibility of 6-minute walking test in patients with COPD. Eur Respir J. 2011;38(2):261–267. doi: 10.1183/09031936.00142010.
48. Hoeymans N, Wouters ERCM, Feskens EJM, et al. Reproducibility of performance-based and self-reported measures of functional status. J Gerontol A. 1997;52:363–368.
49. Hwang R, Morris NR, Mandrusiak A, et al. Timed up and go test: a reliable and valid test in patients with chronic heart failure. J Card Fail. 2016;22(8):646–650. doi: 10.1016/j.cardfail.2015.09.018.
50. Isik EI, Altug F, Cavlak U. Reliability and validity of four step square test in older adults. Turk J Geriatrics. 2015;18:151–155.
51. Jenkins NDM, Cramer JT. Reliability and minimum detectable change for common clinical physical function tests in sarcopenic men and women. J Am Geriatr Soc. 2017;65(4):839–846. doi: 10.1111/jgs.14769.
52. Kristensen MT, Bloch ML, Jønsson LR, et al. Interrater reliability of the standardized Timed Up and Go Test when used in hospitalized and community-dwelling older individuals. Physiotherapy Res Intl. 2019;24(2):1–6. doi: 10.1002/pri.1769.
53. Labadessa IG, Arcuri JF, Sentanin AC, et al. Should the 6-minute walk test be compared when conducted by 2 different assessors in subjects with COPD? Respir Care. 2016;61(10):1323–1330. doi: 10.4187/respcare.04500.
54. LeBrasseur NK, Bhasin S, Miciek R, et al. 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):2118–2123. doi: 10.1111/j.1532-5415.2008.01953.x.
55. Lin MR, Hwang HF, Hu MH, et al. Psychometric comparisons of the timed up and go, one-leg stand, functional reach, and Tinetti balance measures in community-dwelling older people. J Am Geriatr Soc. 2004;52(8):1343–1348. doi: 10.1111/j.1532-5415.2004.52366.x.
56. Madsen OR, Brot C. Assessment of extensor and flexor strength in the individual gonarthrotic patient: interpretation of performance changes. Clin Rheumatol. 1996;15(2):154–160. doi: 10.1007/BF02230333.
57. Marques A, Cruz J, Quina S, et al. Reliability, agreement and minimal detectable change of the Timed Up & Go and the 10-meter walk tests in older patients with COPD. COPD. 2016;13(3):279–287. doi: 10.3109/15412555.2015.1079816.
58. Marques LBF, Moreira B de S, Ocarino JdM, et al. Construct and criterion validity of the functional gait assessment–Brazil in community-dwelling older

adults. Braz J Phys Ther. 2021;25(2):186–193. doi: 10.1016/j.bjpt.2020.05.008.
59. McLay R, Kirkwood RN, Kuspinar A, et al. Validity of balance and mobility screening tests for assessing fall risk in COPD. Chron Respir Dis. 2020;17:1479973120922538. doi: 10.1177/1479973120922538.
60. Medina-Mirapeix F, Bernabeu-Mora R, Llamazares-Herrán E, et al. Interobserver reliability of peripheral muscle strength tests and short physical performance battery in patients with chronic obstructive pulmonary disease: a prospective observational study. Arch Phys Med Rehabil. 2016;97(11):2002–2005. doi: 10.1016/j.apmr.2016.05.004.
61. Medina-Mirapeix F, Vivo-Fernández I, López-Cañizares J, et al. Five times sit-to-stand test in subjects with total knee replacement: reliability and relationship with functional mobility tests. Gait Posture. 2018;59:258–260. doi: 10.1016/j.gaitpost.2017.10.028.
62. Mkacher W, Tabka Z, Trabelsi Y. Minimal detectable change for balance measurements in patients with COPD. J Cardiopulm Rehabil Prev. 2017;37(3):223–228. doi: 10.1097/HCR.0000000000000240.
63. Morris-Chatta R, Buchner DM, de Lateur BJ, et al. Isokinetic testing of ankle strength in older adults: assessment of inter-rater reliability and stability of strength over six months. Arch Phys Med Rehabil. 1994;75(11):1213–1216. doi: 10.1016/0003-9993(94)90007-8.
64. Nunes JP, Cunha PM, Antunes M, et al. The generality of strength: relationship between different measures of muscular strength in older women. Int J Exerc Sci. 2020;13:1638–1649.
65. Ordway NR, Hand N, Riggs G, et al. Reliability of knee and ankle strength measures in an older adult population. J Strength Cond Res. 2006;20:82–87.
66. Ostchega Y, Harris TB, Hirsch R, et al. Reliability and prevalence of physical performance examination assessing mobility and balance in older persons in the US: data from The Third National Health and Nutrition Examination Survey. J Am Geriatr Soc. 2000;48(9):1136–1141. doi: 10.1111/j.1532-5415.2000.tb04792.x.
67. Özden F, Coşkun G, Bakırhan S. The test-retest reliability and concurrent validity of the five times sit to stand test and step test in older adults with total hip arthroplasty. Exp Gerontol. 2020;142:111143. doi: 10.1016/j.exger.2020.111143.
68. Parraca JA, Adsuar JC, Domínguez-Muñoz FJ, et al. Test-retest reliability of isokinetic strength measurements in lower limbs in elderly. Biology. 2022;11(6):802. doi: 10.3390/biology11060802.
69. Peel C, Ballard D. Reproducibility of the 6-minute-walk test in older women. J Aging Phys Act. 2001;9(2):184–193. doi: 10.1123/japa.9.2.184.
70. Phillips WT, Batterham AM, Valenzuela JE, et al. Reliability of maximal strength testing in older adults. Arch Phys Med Rehabil. 2004;85(2):329–334. doi: 10.1016/j.apmr.2003.05.010.
71. Rikli RE, Jones CJ. The reliability and validity of a 6-minute walk test as a measure of physical endurance in older adults. J Aging Phys Act. 1998;6(4):363–375. doi: 10.1123/japa.6.4.363.
72. Rockwood K, Awalt E, Carver D, et al. Feasibility and measurement properties of the functional reach and the timed up and go tests in the Canadian study of health and aging. J Gerontol A. 2000;55:70–74.
73. Rodrigues F, Teixeira JE, Forte P. The reliability of the timed up and go test among portuguese elderly. Healthcare. 2023;11(7):928. doi: 10.3390/healthcare11070928.
74. Rolenz E, Reneker JC. Validity of the 8-Foot Up and Go, Timed Up and Go, and Activities- Specific Balance Confidence Scale in older adults with and without cognitive impairment. J Rehabil Res Dev. 2016;53(4):511–518. doi: 10.1682/JRRD.2015.03.0042.
75. Rydwik E, Karlsson C, Frändin K, et al. Muscle strength testing with one repetition maximum in the arm/shoulder for people aged 75+ – Test-retest reliability. Clin Rehabil. 2007;21(3):258–265. doi: 10.1177/0269215506072088.
76. Saint-Maurice PF, Sampson JN, Keadle SK, et al. Reproducibility of accelerometer and posture-derived measures of physical activity. Med Sci Sports Exerc. 2020;52(4):876–883. doi: 10.1249/MSS.0000000000002206.
77. Schaubert K, Bohannon RW. Reliability of the sit-to-stand test over dispersed test sessions. IES. 2005;13(2):119–122. doi: 10.3233/IES-2005-0188.
78. Schaubert KL, Bohannon RW. Reliability and validity of three strength measures obtained from community-dwelling elderly persons. J Strength Cond Res. 2005;19:717–720.
79. Schroeder ET, Wang Y, Castaneda-Sceppa C, et al. Reliability of maximal voluntary muscle strength and power testing in older men. J Gerontol A Biol Sci Med Sci. 2007;62(5):543–549. doi: 10.1093/gerona/62.5.543.
80. Schwenk M, Gogulla S, Englert S, et al. Test-retest reliability and minimal detectable change of repeated sit-to-stand analysis using one body fixed sensor in geriatric patients. Physiol Meas. 2012;33(11):1931–1946. doi: 10.1088/0967-3334/33/11/1931.
81. Sherwood JJ, Inouye C, Webb SL, et al. Reliability and validity of the sit-to-stand as a muscular power measure in older adults. J Aging Phys Act. 2020;28(3):455–466. doi: 10.1123/japa.2019-0133.
82. Suwit A, Rungtiwa K, Nipaporn T. Reliability and validity of the osteoarthritis research society international minimal core set of recommended performance-based tests of physical function in knee osteoarthritis in community-dwelling adults. Malays J Med Sci. 2020;27(2):77–89. doi: 10.21315/mjms2020.27.2.9.
83. Suzuki Y, Kamide N, Kitai Y, et al. Absolute reliability of measurements of muscle strength and physical performance measures in older people with high functional capacities. Eur Geriatr Med. 2019;10(5):733–740. doi: 10.1007/s41999-019-00218-9.
84. Symons TB, Vandervoort AA, Rice CL, et al. Reliability of isokinetic and isometric knee-extensor force in older women. J Aging Phys Act. 2004;12(4):525–537. doi: 10.1123/japa.12.4.525.
85. Symons TB, Vandervoort AA, Rice CL, et al. Reliability of a single-session isokinetic and isometric strength measurement protocol in older men. J Gerontol A Biol Sci Med Sci. 2005;60(1):114–119. doi: 10.1093/gerona/60.1.114.
86. Tolk JJ, Janssen RPA, Prinsen CAC, et al. The OARSI core set of performance-based measures for knee osteoarthritis is reliable but not valid and responsive. Knee Surg Sports Traumatol Arthrosc. 2019;27(9):2898–2909. doi: 10.1007/s00167-017-4789-y.
87. Tolk JJ, Janssen RPA, Prinsen CSAC, et al. Measurement properties of the OARSI core set of

performance-based measures for hip osteoarthritis: a prospective cohort study on reliability, construct validity and responsiveness in 90 hip osteoarthritis patients. Acta Orthop. 2019;90(1):15–20. doi: 10.1080/17453674.2018.1539567.
88. Topp R, Mikesky A. Reliability of isometric and isokinetic evaluations of ankle Dorsi/plantar strength among older adults. IES. 1994;4(4):157–163. doi: 10.3233/IES-1994-4407.
89. Uszko-Lencer NHMK, Mesquita R, Janssen E, et al. Reliability, construct validity and determinants of 6-minute walk test performance in patients with chronic heart failure. Int J Cardiol. 2017;240:285–290. doi: 10.1016/j.ijcard.2017.02.109.
90. Van Driessche S, Van Roie E, Vanwanseele B, et al. Test-retest reliability of knee extensor rate of velocity and power development in older adults using the isotonic mode on a Biodex System 3 dynamometer. PLoS One. 2018;13(5):e0196838. doi: 10.1371/journal.pone.0196838.
91. van Iersel MB, Munneke M, Esselink RAJ, et al. Gait velocity and the Timed-Up-and-Go test were sensitive to changes in mobility in frail elderly patients. J Clin Epidemiol. 2008;61(2):186–191. doi: 10.1016/j.jclinepi.2007.04.016.
92. Verdijk LB, Van Loon L, Meijer K, et al. One-repetition maximum strength test represents a valid means to assess leg strength in vivo in humans. J Sports Sci. 2009;27(1):59–68. doi: 10.1080/02640410802428089.
93. Wallmann HW, Evans NS, Day C, et al. Interrater reliability of the five-times-sit-to-stand test. Home Health Care Manag Pract. 2013;25(1):13–17. doi: 10.1177/1084822312453047.
94. Webber SC, Porter MM. Reliability of ankle isometric, isotonic, and isokinetic strength and power testing in older women. Phys Ther. 2010;90(8):1165–1175. doi: 10.2522/ptj.20090394.
95. Whitney SL, Wrisley DM, Marchetti GF, et al. Clinical measurement of sit-to-stand performance in people with balance disorders: Validity of data for the Five-Times-Sit-to-Stand Test. Phys Ther. 2005;85(10):1034–1045. doi: 10.1093/ptj/85.10.1034.
96. Wrisley DM, Kumar NA. Functional gait assessment: concurrent, discriminative, and predictive validity in community-dwelling older adults. Phys Ther. 2010;90(5):761–773. doi: 10.2522/ptj.20090069.
97. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42. doi: 10.1016/j.jclinepi.2006.03.012.
98. Dobson F, Hinman RS, Hall M, et al. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthritis Cartilage. 2012;20(12):1548–1562. doi: 10.1016/j.joca.2012.08.015.
99. Liu Y, Li H, Ding N, et al. Functional status assessment of patients with COPD: a Systematic review of performance-based measures and patient-reported measures. Medicine. 2016;95(20):e3672–e3672. doi: 10.1097/MD.0000000000003672.