ELSEVIER

The Spine Journal ■■ (2017) ■■–■■

THE SPINE JOURNAL

Review Article

Reliability of physical functioning tests in patients with low back pain: a systematic review

Lenie Denteneer, MT, PT[a,*], Ulrike Van Daele, PhD, MT, PT[a], Steven Truijen, PhD, MSc[a], Willem De Hertogh, PhD, MT, PT[a], Jill Meirte, PhD, PT[a], Gaetane Stassijns, PhD, MD[b,c]

[a]*Faculty of Medicine and Health Sciences, Rehabilitation and Physiotherapy, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium*
[b]*Faculty of Medicine and Health Sciences, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium*
[c]*Physical Medicine and Rehabilitation, Antwerp University Hospital, Wilrijkstraat 10, 2650 Edegem, Belgium*

Received 9 May 2017; revised 21 August 2017; accepted 29 August 2017

Abstract

PURPOSE: The aim of this study was to provide a comprehensive overview of physical functioning tests in patients with low back pain (LBP) and to investigate their reliability.

DATA SOURCES: A systematic computerized search was finalized in four different databases on June 24, 2017: PubMed, Web of Science, Embase, and MEDLINE.

STUDY SELECTION: Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines were followed during all stages of this review. Clinical studies that investigate the reliability of physical functioning tests in patients with LBP were eligible. The methodological quality of the included studies was assessed with the use of the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist. To come to final conclusions on the reliability of the identified clinical tests, the current review assessed three factors, namely, outcome assessment, methodological quality, and consistency of description.

DATA SYNTHESIS: A total of 20 studies were found eligible and 38 clinical tests were identified. Good overall test-retest reliability was concluded for the extensor endurance test (intraclass correlation coefficient [ICC]=0.93–0.97), the flexor endurance test (ICC=0.90–0.97), the 5-minute walking test (ICC=0.89–0.99), the 50-ft walking test (ICC=0.76–0.96), the shuttle walk test (ICC=0.92–0.99), the sit-to-stand test (ICC=0.91–0.99), and the loaded forward reach test (ICC=0.74–0.98). For inter-rater reliability, only one test, namely, the Biering-Sörensen test (ICC=0.88–0.99), could be concluded to have an overall good inter-rater reliability. None of the identified clinical tests could be concluded to have a good intrarater reliability.

CONCLUSIONS: Further investigation should focus on a better overall study methodology and the use of identical protocols for the description of clinical tests. The assessment of reliability is only a first step in the recommendation process for the use of clinical tests. In future research, the identified clinical tests in the current review should be further investigated for validity. Only when these clinimetric properties of a clinical test have been thoroughly investigated can a final conclusion regarding the clinical and scientific use of the identified tests be made.

Keywords: Low back pain; Lumbar vertebrae; Physical examination; Physical fitness; Reproducibility of results; Social participation

FDA device/drug status: Not applicable.

Author disclosures: ***LD:*** Nothing to disclose. ***UVD:*** Nothing to disclose. ***ST:*** Nothing to disclose. ***WDH:*** Nothing to disclose. ***JM:*** Nothing to disclose. ***GS:*** Nothing to disclose.

The disclosure key can be found on the Table of Contents and at www.TheSpineJournalOnline.com.

To the knowledge of the authors, there are no conflicts of interest.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

* Corresponding author. Faculty of Medicine and Health Sciences, Rehabilitation and Physiotherapy, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium. Tel.: (32) 494884189.

E-mail address: lenie.denteneer@uantwerpen.be (L. Denteneer)

https://doi.org/10.1016/j.spinee.2017.08.257
1529-9430/

Background

Patients with low back pain (LBP) suffer from a wide range of problems that can all be allocated within the International Classification of Functioning, Disability and Health (ICF) [1]. The level of physical functioning (ICF category "activities and participation") and its relevance for patients with LBP has been emphasized repeatedly [2–4]. When patients with LBP are asked to report treatment goals before starting their rehabilitation, 96% of those reported goals can be classified in the ICF category activities and participation [5]. This phenomenon increases the importance to assess physical functioning in patients with LBP to set appropriate treatment goals and to make an estimation of their prognosis. Physical functioning is best defined as an individual's ability to perform activities required in their daily lives [6–8]. The term "activity" can be defined as bodily movement resulting from the contraction of the skeletal muscles that results in an increase in energy expenditure above resting levels [9,10].

The assessment of physical functioning is often performed with the use of patient-reported outcome measurements (PROMs) [11–14]. It is, however, important to note that there are some limitations to the use of these measurements, such as inaccuracies caused by recall bias, social desirability bias, and errors in self-observation [15–17]. Therefore, it seems to be important that one complements the use of PROMs with clinical testing in which patients actually have to perform a certain task. The growing interest toward physical functioning in specific patient populations has led to the development of clinical tests that measure physical functioning of standardized tasks such as walking, balancing, reaching, rising from a chair, and climbing stairs [18,19].

As a clinical test is developed, testing for reliability is considered before testing for validity because a test cannot be considered valid if the results are inconsistent [20]. To determine physical functioning with the use of clinical tests, reliable outcome measures need to be available [21]. Once sufficiently reliable outcome measures are identified, they can form a base for further research regarding validity. Hence, the aim of the present study was to provide a comprehensive overview of all physical functioning tests in patients with LBP and to investigate their test-retest, inter-rater, or intrarater reliability.

Methods

The reporting of this systematic literature review was done in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [22] and has been registered in the PROSPERO database as 2016:CRD42016046215.

Study selection

This review included studies that investigate the reliability of physical functioning tests in patients with LBP. Studies were included if they met following inclusion criteria: (1) test-retest, intrarater, or inter-rater reliability had to be conducted on a population with LBP aged between 18 and 70 years; (2) the identified clinical tests assessed physical functioning; (3) only simple measurement devices were allowed (field-based); and finally, (4) the studies were written in English or Dutch.

Studies were excluded if they met following exclusion criteria: (1) the patients had pathologies other than LBP or the patients had LBP not primarily originating from the lumbar spine (eg, sacroiliac or pelvic pathology, pregnancy-related LBP, and malignity); (2) the clinical tests required high technological medical devices (eg, magnetic resonance imaging, 3D motion analysis system); and finally, (3) the studies were reviews, practice guidelines, pilot studies, case reports, commentaries, editorials, letters, study protocols, and books.

Data sources and search strategy

A systematic computerized search was finalized by one author on June 24, 2017, in four different databases: PubMed (1972–), Web of Science (1955–), Embase (1947–), and MEDLINE (1946–). A brief search in ClinicalTrials.gov was conducted to identify any possible ongoing studies. The keywords for the search strategy are presented as a supplementary file. Identified studies were uploaded into Endnote (Thomas Reuters), and duplicates were removed. Two independent reviewers selected studies for inclusion in a two-step process. First, studies were screened based on the title and the abstract. If there was no consensus between the decisions of the two reviewers, the study was included into the second stage without deliberation. The second stage, screening on full text, was also conducted independently by both reviewers. In case of disagreement on the inclusion of a study, the two reviewers came to a consensus during a deliberation session.

Data extraction

Relevant data were extracted as follows: (1) description and scoring of the clinical tests; (2) population characteristics; (3) inclusion and exclusion criteria; (4) description of the used procedures; (5) results for test-retest, intrarater, and inter-rater reliabilities; and (6) the statistical test.

Reliability

In line with Rousson et al. [23], we assessed the following three aspects of reliability: intrarater reliability (when presenting repeatedly the same observations to one rater); inter-rater reliability (when presenting the same observations to two or more raters), and test-retest reliability (when presenting the same task to the same subjects two or more times). Results from one subject examined by the same observer (intrarater or test-retest reliability) or by several observers examining the same subject (inter-rater reliability) should stay

consistent [20]. The reliability of nominal and ordinal unpaired data is often analyzed with the Cohen or the weighted kappa (κ) coefficient [24]. κ-Values may vary between −1 and 1. Agreement can be interpreted as κ<0.20=poor, κ: 0.21–0.40=fair, κ: 0.41–0.60=moderate, κ: 0.61–0.80=good, and κ: 0.81–1.0=excellent [25]. Continuous data are often analyzed with intraclass correlation coefficients (ICCs) [24]. The most common ICC model for reliability is the two-way random ICC model (ICC 2.1), which means that reliability is calculated from a single measurement and that each subject is measured by each rater [26]. Whatever the type of ICC that is calculated, it is suggested that an ICC close to 1 indicates "excellent" reliability. An ICC exceeding 0.70 indicates good reliability and an ICC below 0.70 indicates moderate to poor reliability [27].

Quality of the evidence

To come to final conclusions on the reliability of the identified clinical tests, the current review will take the following factors into account:

1) Outcome assessment of the identified clinical tests: Only tests with at least "good" ICC values (ICC ≥0.70) and, if applicable, at least good κ-values (κ ≥0.81) were taken into account to come to final conclusions on reliability. Also, if clinical tests are assessed within multiple independent studies, these results will be given higher values to come to final conclusions on their reliability.
2) Consistency in the description of the identified clinical tests: When clinical tests were described in multiple studies, it is important that the description of that test is consistent throughout the different studies. Clinical tests that were investigated in studies with an inconsistency in the description of execution or interpretation of a test were not handled as a merit to come to final conclusions regarding the reliability of a clinical test.
3) Methodological quality of the included articles: This was assessed independently by two authors with the use of the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist [28]. The COSMIN checklist assesses the subitem "reliability," for which 14 questions need to be answered [29]. The COSMIN four-point scale was used ("excellent," "good," "fair," or "poor") to assess the methodological quality for each question [30]. The overall assessment of the methodological quality is obtained by taking the lowest rating (namely, the "worse score counts" algorithm). For each article, an overall methodological quality was rated as excellent, good, fair, or poor [30]. Studies with a higher overall methodological quality were given a higher value to come to final conclusions on reliability.

Results

A total of 20 studies [31–50] were included in this review. The flowchart is presented in the Figure. In total, 38 clinical tests were identified, representing a total sample of 798 patients (Table 1). A detailed description of the interpretation and the execution of each clinical test is provided in Table 2. The most reported clinical test was the *Biering-Sörensen test*, which was assessed within a total of four independent studies. This clinical test was then followed by the *50-ft walking test*, the *sit-to-stand test*, the *loaded forward reach test*, the *extensor endurance test*, the *floor-to-waist lift test*, and the *repeated trunk flexion test*, which were all assessed within three independent studies.

Table 3 presents all outcome results for test-retest, inter-rater, and intrarater reliabilities. Within this table, clinical tests were subdivided according to the number of times they were assessed in different studies and if the tests were consistently described. Finally, Table 3 also points out the studies that were assessed as having a poor overall methodological quality.

Test-retest reliability

A total of 21 clinical tests were assessed for test-retest reliability [31,33–36,38,40,43–46]. Out of these tests, a total of 11 clinical tests were investigated in a single study. Of the remaining 10 clinical tests that were investigated in at least two independent studies, 8 clinical tests showed an overall good test-retest reliability and 2 clinical tests did not. The clinical tests that were investigated in multiple studies and were in agreement with a good test-retest reliability were the *extensor endurance test* (ICC=0.93–0.97 [31,35]), the *flexor endurance test* (ICC=0.90–0.97 [31,35]), the *5-minute walking test* (ICC=0.89–0.99 [43,46]),the *50-ft walking test* (ICC=0.76–0.96 [43,46]), the *shuttle walk test* (ICC=0.92–0.99 [40,45]), the *sit-to-stand test* (ICC=0.91–0.99 [43,46]), the *loaded forward reach* test (ICC=0.74–0.98 [43,46]), and the *progressive isoinertial lifting evaluation test* (ICC=0.92–0.94 [34,43]). The other two tests that were not in agreement were the *Biering-Sörensen test* (ICC=0.59–0.95 [33,36]) and the *repeated trunk flexion test* (percentage agreement [PA]=62% [44], ICC=0.99 [46]).

Inter-rater reliability

A total of 33 clinical tests were assessed for inter-rater reliability [32,37–39,41,42,44,47,48,50]. In total, 27 clinical tests were investigated in a single study. Of the remaining six clinical tests that were investigated in multiple studies, two were identified as having an overall good inter-rater reliability and four were not. The clinical tests that were investigated in at least two studies and were in agreement with a good inter-rater reliability were the *Biering-Sörensen test* (ICC=0.88–0.99 [37,42]) and the *floor-to-waist lift test* (κ=0.62–1.00

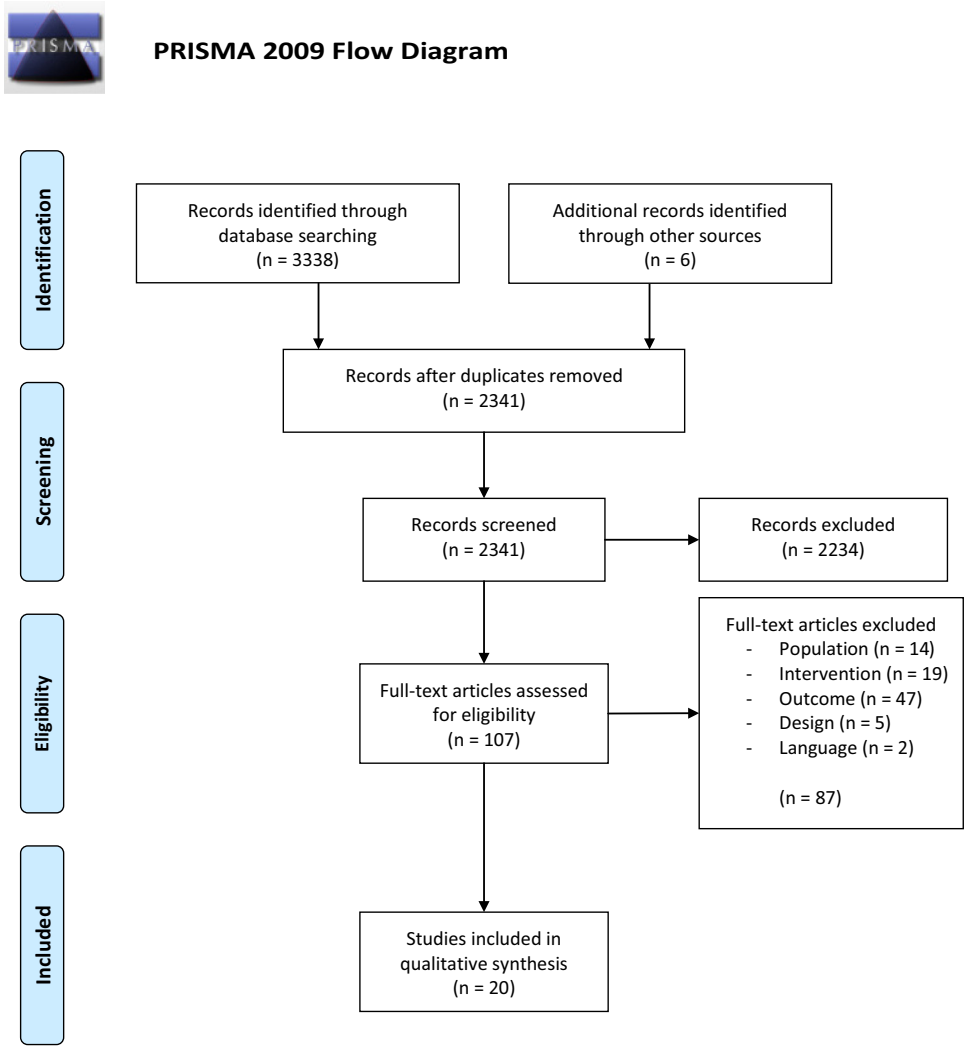

Figure. Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) flow diagram.

[32,38,50]). The other four clinical tests that were not in agreement were the *repeated trunk flexion* (PA=65% [44], ICC=0.99 [42]), *single squat* (κ=0.55–1.00 [32,39]), *sit-up* (κ=0.48–0.62 [47,48]), and the *bilateral active straight leg raise* (PA=65% [44], κ=0.77 [48]).

Intrarater reliability

A total of 12 clinical tests were assessed for inter-rater reliability [39,42,49,50]. However, all of these clinical tests were investigated in a single study. Good reliability was identified for the *Biering-Sörensen test* (ICC=0.91 [42]), the *50-ft walking test* (ICC=0.95–0.99 [42]), the *loaded forward reach test* (ICC=0.99 [42]), the *time up and go test* (ICC=0.98 [42]), the *floor-to-waist lift test* (κ=0.73 [50]), the *single-squat test* (κ=1.0 [39]), the *30-second chair stand test* (ICC=0.94 [49]), and the *unloaded reach test* (ICC=0.98 [42]).

Further results for either test-retest, inter-rater, and intrarater reliabilities are shown in Table 3.

Methodological quality

None of the 20 studies scored excellent or good for the overall methodological quality according to the COSMIN checklist. The overall methodological quality was rated as fair in 11 studies [31,33,35,36,38,42,43,45,46,48,49] and as poor in the remaining 9 studies [32,34,37,39–41,44,47,50]. The main reasons for the downgrade of the methodological quality caused by a poor or fair rating were, first, an inadequate sample size (17/20) [31–34,36–42,44–47,49,50]; second,

Table 1
Included articles

	Author	Reported clinical tests	Demographic values	Inclusion and exclusion criteria
1	Del Pozo-Cruz et al. 2014 [31]	Extensor endurance test Flexor endurance test	N=31 (71% women) Mean age, men: 45.9 y (SD 9.17) Mean age, women: 46.0 y (SD 8.18) Mean ODI, men: 29.93% (SD 1.49) Mean ODI, women: 28.12% (SD 2.52) Mean RMDQ, men: 11.21 (SD 2.22) Mean RMDQ, women: 12.04 (SD 2.40)	Inclusion: subacute nonspecific LBP in office workers, either the first episode of LBP or a recurrence, current episode having lasted between 6 and 12 wk, age between 18–65 y, physical inactivity, employee status, more than 6 h/d at a computer workstation Exclusion: a specific diagnosed cause of LBP, chronic backache, other major disease, lack of fluency in Spanish
2	Durand et al. 2004 [32]	Walking ability Stair climbing Floor-to-waist lift test Repeated trunk rotation (seated or standing) Squatting (repeated or single) Recline reach Carrying (unilateral or bilateral) Push Pull Kneeling Crawling Ladder climbing	N=40 (n=9 women) Mean age: 40.9 y (SD 9.9) Mean VAS: 4.5 (SD 2.1) Mean duration of LBP: 13 mo (SD 18.6)	Inclusion: back pain, age between 18 and 65 y, French speaking Exclusion: specific back pain, back surgery, pregnancy
3	Gruther et al. 2009 [33]	Biering-Sörensen test	N=32 Mean age: 43 y (SD 10)	Inclusion: age 18–60 y, BMI <30 kg/m [2], LBP at least 3 mo, pain no more than 3 VAS, no headache history for more than 5 d in the previous year and no headache in the last 6 wk Exclusion: none
4	Hodselmans et al. 2007 [34]	Progressive isoinertial lifting evaluation	N=20 (n=12 women) Mean age: 33.8 y (SD 8.6)	Inclusion: nonspecific chronic LBP, motivation to participate in rehabilitation. Exclusion: small amount of suffering caused by low back pain, conflicts with employer or insurance company, any medical condition that would interfere with psychophysical performance tests, major surgery within the last year, infections, cancer, and neuralgic or cardiovascular disease
5	Ito et al. 1996 [35]	Extensor endurance test Flexor endurance test	N=100 (n=60 women) Mean age: 45.3 y (range 33–48 y)	Inclusion: primary low back pain Exclusion: sciatica and neurologic deficits at least 6 mo, history of low back surgery, radiological findings of slightly or moderately degenerative changes without any gross spinal pathology such as tumor, infection, osteoporosis, spondylolysis and spondylolisthesis, involvement with workers' compensation, litigation or disability insurance
6	Kahraman et al. 2016 [49]	30-s Chair stand test	N=38 (37% women) Mean age: 35 y (SD 10) Mean ODI: 22.6 (SD 14.2)	Inclusion: patients aged 18–55 y old with chronic nonspecific low back pain Exclusion: pregnancy, received physiotherapy in the last 6 mo, taking analgesics during the study, presence of major musculoskeletal or neurologic disorders
7	Keller et al. 2001 [36]	Biering-Sörensen test	N=31 (n=24 women) Mean age: 36 y (range 30.0–42.5) Mean duration of LBP: 4 y (range 1.4–10.0)	Inclusion: LBP at least 3°mo Exclusion: inflammatory diseases, spinal deformity, structural damage caused by fractures, idiopathic scoliosis, congenital malformations, any surgery involving the low back
8	Latimer et al. 1999 [37]	Biering-Sörensen test	N=23 (n=13 women) Mean age: 35.9 y (SD 15.7) Mean RMDQ: 3.7 (SD 2.8)	Inclusion: nonspecific low back pain Exclusion: diagnosed cardiovascular disease, neurologic disease, pain caused by a specific disease process
9	Magnussen et al. 2004 [38]	Floor-to-waist lift test Sock test Pick-up test Roll-up test	N=41 LBP (54% women) Mean age: 39.8 y (SD 10.7) Mean RMDQ: 10.5 (SD 4.4) Mean FABQ: 38.6 (SD 16.3)	Inclusion: patients with LBP attending an outpatient rehabilitation program Exclusion: none

(*Continued*)

Table 1
(*Continued*)

	Author	Reported clinical tests	Demographic values	Inclusion and exclusion criteria
10	Pratt et al. 2002 [40]	Shuttle walk test	N=29 (n=12 women,) Mean age: 69 y (range 49–82)	Inclusion: patients with diagnosis of lumbar spinal stenosis established on clinical history of neurogenic claudication and positive MRI Exclusion: any comorbidity that limits walking distance
11	Paatelma et al. 2010 [39]	Extensor endurance test Walking ability Single squat Single-leg stance	N=15 (74% women) Mean age: 37.9 y (SD 4.5)	Inclusion: LBP lasting less than 3 mo Exclusion: pregnancy, use of psychogenic medication, diagnosed osteoporosis
12	Reneman et al. 2005 [41]	Lifting capacity	N=16 (n=4 women) Mean age: 39.6 y (SD 7.1) Mean RMDQ: 11.1 (SD 4.7)	Inclusion: chronic LBP Exclusion: comorbidity (additional diagnoses unrelated to low back pain), specific diagnoses related to LBP
13	Smeets et al. 2006 [43]	5-min Walking test 50-ft Walking test Stair climbing Sit-to-stand test Progressive isoinertial lifting evaluation Loaded forward reach	N=53 (n=28 women) Mean age: 43.19 y (SD 9.27) Mean RMDQ: 13.7 (SD 4.21)	Inclusion: patients with nonspecific chronic LBP, RMDQ >3 Exclusion: lumbar spondylolisthesis, spondylosis, fracture, severe psychopathology
14	Smith 1994 [50]	Floor-to-waist lift test	N=21 (n=5 women,) Mean age: 40 y (range 22–61) Average duration of symptoms 24 mo (range 5–57)	Inclusion: low back pain Exclusion: none
15	Spratt et al. 1990 [44]	Repeated trunk flexion Partial push-up (single or repeated) Bilateral active straight leg raise	N=42 (n=23 women) Mean age: 38.9 y (range 20.6–59.0)	Inclusion: age between 18 and 60 y, LBP more than 4 wk Exclusion: LBP related to pregnancy, neurologic deficit, excessive obesity, gross psychologic disturbance, congenital pathologic or traumatic spondylolisthesis, tumor, vertebral fracture or collapse, discitis, osteomyelitis, scoliosis or ankylosing spondylitis
16	Simmonds et al. 1998 [42]	Biering-Sörensen test 50-ft Walking test Sit-to-stand test Repeated trunk flexion Forward reach (loaded or unloaded)	N=44 (n=28 women) Mean age: 42.6 y (range 21–63) Mean duration of LBP: 12.4 mo (SD 20.8) Mean RMDQ: 8.34 SD 6.02)	Inclusion: nonspecific mechanic LBP Exclusion: any medical condition besides LBP in the LBP group, major surgery within the last year, current infectious disease, cancer, neurologic or cardiovascular disease
17	Taylor et al. 2001 [45]	Shuttle walk test	N=44 (n=26 women) Mean age: 48.2 y (SD 14.2) Mean ODI: 28.5	Inclusion: LBP at least 6 mo, mechanical LBP with or without sciatica Exclusion: none
18	Teixeira da Cunha-Filho et al. 2010 [46]	5-min Walking test 50-ft Walking test Time up and go Sit-to-stand test Repeated trunk flexion Loaded forward reach	N=30 (n=24 women) Mean age: 39.37 y (SD 12.3) Mean RMDQ: 9.8 (SD 5.8)	Inclusion: complaint of chronic LBP, independent of additional bodily distribution, which led the patient to seek medical assistance Exclusion: any medical condition that could influence or interfere with the performance of the tests, spine surgery, infectious diseases, cancer, uncontrolled neurologic diseases, uncontrolled cardiovascular diseases
19	Viikari-Juntura et al. 1998 [47]	Sit-up	n=27 (no further info available)	Inclusion: workers with LBP Exclusion: none
20	Waddell et al. 1992 [48]	Bilateral active straight leg raise Sit-up	N=120 (n=61 women) Mean age: 35.3 y (SD 9.9) Total duration of LBP: 83.5 mo (SD 84.7)	Inclusion: chronic LBP with or without referred pain Exclusion: neurologic symptoms or signs, serious spinal complications, previous spinal surgery, spinal fracture, spinal deformity

LBP, low back pain; ODI, Oswestry Disability Index; RMDQ, Roland-Morris Disability Questionnaire; MODI, Modified Oswestry Disability Index; ROM, range of motion; BMI, body mass index; SD, standard deviation; MRI, magnetic resonance imaging; VAS, visual analog scale.

Table 2
Specifications of indentified clinical tests

Identified clinical tests	Author	Specification of test
Biering-Sörensen test	Gruther et al. 2009 [33]	The subject was stabilized with the lower extremity by two belts at the level of the hips and below the knees. The trunk was raised to the horizontal position with hands crossed over the chest. The posture was continued until the participant could no longer hold the horizontal posture, or until he or she reached the limit of fatigue pain.
	Keller et al. 2001 [36]	A roman chair device with a padded pelvic support, padded adjustable height, dorsal calf supports, and handlebars was used to stabilize subject entry. The participant initially hung flexed 90° with the head down from the waist, having the pelvis supported by the pubic symphysis and anterior iliac prominence. The trunk was then raised to the horizontal position with hands crossed over the chest. The test was continued until the participant could no longer control the horizontal posture, or until he or she reached the limit for fatigue or pain.
	Latimer et al. 1999 [37]	The patient lies prone on the treatment couch with the lower half of the body below the level of the anterior superior iliac spines strapped to the couch at three positions. Arms are crossed over the chest and the trunk should be maintained in neutral alignment as long as possible.
	Simmonds et al. 1998 [42]	Subjects laid prone on a standard treatment table, positioned with their hips at the edge of the table. Straps were placed across the thighs and calves to stabilize the subject. With the arms against the body, the subjects were required to lift the upper body so that it was in a horizontal plane with the table and to hold that position for as long as possible.
Extensor endurance test	Del Pozo-Cruz et al. 2014 [31]	Patients were asked to adopt a prone position while keeping the breastbone off the floor. A small pillow was placed under the lower abdomen to decrease the lumbar lordosis. Patients had to maintain this positions as long as possible with maximum flexion of the cervical spine and with pelvic stabilization through gluteal muscle contraction.
	Ito et al. 1996 [35]	Subjects were asked to lie prone while holding the sternum off the floor. A small pillow was placed under the lower abdomen to decrease the lumbar lordosis. Patients were asked to keep maximum cervical flexion. Subjects needed to maintain the position as long as possible and not exceeding a 5 min time limit.
	Paatelma et al. 2010 [39]	With the patient lying prone on the plinth, fixate the patient's feet above the malleoli firmly against the plinth. Ask him to extend his back as high as possible with hands behind his neck, and hold that position for 60 s. Mark down the angle at which the back has been lifted with an inclinometer 10 cm above the line of the spina iliaca posterior superior. Mark the finding if the patient cannot sustain that position for 60 s or if the back drops at 10 cm or more (2); otherwise, normal (1).
Flexor endurance test	Del Pozo-Cruz et al. 2014 [31]	Participants were requested to recline in a supine position and to elevate the lower extremities to a 90° flexion of hip and knee joints.
	Ito et al. 1996 [35]	Patients lie supine and raise lower extremities with 90° hip flexion and knee flexion. Patients were asked to keep maximum cervical flexion. Subjects needed to maintain the position as long as possible and not exceeding a 5-min time limit.
5-min Walking test	Smeets et al. 2006 [43]	Patients have to walk as fast and far as possible, no running for 5 min on a 30-m-long eight-shaped circuit. Participants were allowed to take a rest on a chair.
	Teixeira da Cunha-Filho et al. 2010 [46]	This test estimates the functional aerobic capacity during walking. The participant was asked to walk as fast as possible on a 60-m circular track for 5 min. There were marks on the ground at each 2 m. The chronometer was activated when the individual began walking, and when the given time was over, he or she stopped just where he or she was. Then the total distance completed was recorded in meter.
50-ft Walking test	Smeets et al. 2006 [43]	Patients walk as fast as possible (no running) until they got back to the starting point. No walking aids were allowed. The circuit was 15 m long, eight shaped
	Simmonds et al. 1998 [42]	For the distance walk, subjects walked 25 ft, turned around, and walked back to the starting position, once as fast as they could and once at their preferred walking speed.
	Teixeira da Cunha-Filho et al. 2010 [46]	This test provides an estimation of gait speed. The participant was asked to walk to the end of a 25-ft walkway previously marked by adhesive tape, turn around, and return to the starting point as fast as possible. The chronometer registered the total time in seconds needed to perform the test.
Time up and go	Simmonds et al. 1998 [42]	The subject started from seated position, stood, and walked forward to a line 3 m away, turned back to the chair, and sat down.
	Teixeira da Cunha-Filho et al. 2010 [46]	The individual began by sitting on a chair then got up, walked 3 m, turned around, and returned to sit down again, as fast as possible without using the chair's arms to stand. The total time in seconds was recorded.
Shuttle walk test	Pratt et al. 2002 [40]	A 10-m course is measured on flat ground. The patient must reach the end of the walkway within a specified time dedicated by a beep sounding from the tape. During the first minute of the test, beeps sound each 20 s, and three shuttles (30 m) are completed. During the second minute, four shuttles are competed; during the third minute, five shuttles are completed; and so on up to 14 transits in 12 min, with a maximum total distance of 1,020 m. The assessor counts the number of completed shuttles. The test ends when the patient can no longer complete a shuttle before the next beep sounds. If the patient is within 50 cm of the end of the shuttle when the beep sounds, he or she is given the opportunity to make up the distance during the next shuttle. The result of the test is given in meter (number of completed shuttles multiplied by 10).

(*Continued*)

Table 2
(*Continued*)

Identified clinical tests	Author	Specification of test
	Taylor et al. 2001 [45]	Subjects need to walk a 10-m course, identified at each end by two cones inset 0.5 m from either and to avoid the need for abrupt changes in direction. The speed at which the subjects walked was dictated by an audio signal played on the tape recorder. The speed and distance was increased each minute until the end of the test. The test is stopped when the subject failed to complete a shuttle in the time allocated or if the subject reported pain or discomfort from preventing to continue.
Walking ability	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability to the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
	Paatelma et al. 2010 [39]	Follow the patient's movements, if the movement is difficult (limping, slow), then make a finding (2); otherwise, normal (1).
Stair climbing	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
	Smeets et al. 2006 [43]	Patients had to walk a stair up and down for 1 min. The circuit was five stairs high and eight shaped.
Sit-to-stand test	Smeets et al. 2006 [43]	Patients stand up from a chair without arms five times as fast as possible. The test is performed twice and the average time was calculated
	Simmonds et al. 1998 [42]	The subject was required to rise to a standing position and return to sitting as fast as possible five times.
	Teixeira da Cunha-Filho et al. 2010 [46]	The participant, starting from the sitting position on an armless chair, stood upright and sat five times as fast as he or she could. The time to complete the test was recorded in seconds.
30-s chair to stand test	Kahraman et al. 2016 [49]	Patients have to stand up and sit down from a chair as many times as possible within 30 s. The starting position was standardized, including buttock placement, back support, and foot placement. The patients were instructed to look straight ahead and to rise after the command as fast as possible with their arms folded across their chest.
Progressive isoinertial lifting evaluation	Hodselmans et al. 2007 [34]	For a period of 20 s, the patient had to lift a box with weights, four times from the ground on to a table. Stepwise, after each session, during the 20-s rest, the weight of the box increased by 2.25 kg for women and 4.5 kg for man, respectively. The heart rate was measured with a heart rate equipment. The observer stopped the test when the patient had reached the heart rate safety limit (220–age×85%). The patient was instructed to stop the performance when an acceptable maximal effort was reached. Perceived exertion was assessed directly after reaching the acceptable maximal effort, the cardiac safety limit, or the end of the test calculated with the Waters formula. The maximum safety load for men was 402 N and that for women was 202 N. The psychophysical dynamic lifting capacity was calculated as the ration of the performance and Borg score, expressed in N/B.
	Smeets et al. 2006 [43]	Patients had to lift a box with a weight four times within 20 s from floor up to a 75-cm high table. The starting weight for women was 3.6 kg and that for men was 5.85 kg. After each completed cycle, the weight for women was increased by 2.25 kg and that for men was increased by 4.5 kg. Test was stopped when the patient could not lift the box four times within 20 s, experienced pain, or had a very high heart rate.
Floor-to-waist lift test	Durand et al. 2004 [32]	This test is scored by a six-category ordinal scale ranging from very heavy to unable. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
	Magnussen et al. 2004 [38]	The patient is standing on the floor in front of a table and is asked to lift a box containing a sandbag of 5 kg for 1 min from the floor to the table (height 76 cm) and back to the floor using an optional technique.
	Smith 1994 [50]	Each patient performed a series of these lifts, in which they lifted a 35×35 cm box from a table at waist height, turned 90° to the left, and lowered the box to the floor. The patient then lifted the box off of the floor, rotated 90° to the right, and returned the box to the original position on the table. Slots 5 cm wide and 12 cm long cut in the side of the wooden box, 8 cm from the bottom of the box, served as handles.
Sock test	Magnussen et al. 2004 [38]	The patient is sitting on a high, form bench, the feet not reaching the floor. One leg is tested at the time—the least reach scored.
Pick-up test	Magnussen et al. 2004 [38]	The patient is standing on the floor. A curled piece of paper is dropped on the floor and the patient is asked to pick the paper up.
Roll-up test	Magnussen et al. 2004 [38]	The patient is lying supine on a firm mattress and is asked to roll slowly into a long sitting position with the arms relaxed.

(*Continued*)

Table 2
(*Continued*)

Identified clinical tests	Author	Specification of test
Repeated trunk flexion	Spratt et al. 1990 [44]	No information
	Simmonds et al. 1998 [42]	From the natural position, the patient is asked to flex to the limit of range and to return to the upright position as fast as tolerable. The movement was repeated 10 times.
	Teixeira da Cunha-Filho et al. 2010 [46]	The speed in executing the following movement was assessed. The participant was asked to bend his or her trunk forward, reaching for the toes from the upright position, as far as possible and then to return to a standing position. This movement was repeated five times and, after a short pause, the patient repeated the task. A stopwatch was used to record the time in each trial. The average time between the two trials in seconds was registered for analysis.
Repeated trunk extension	Spratt et al. 1990 [44]	No information
Repeated trunk rotation seated	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Repeated trunk rotation standing	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability to the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Single partial push-up	Spratt et al. 1990 [44]	No information
Repeated partial push-up	Spratt et al. 1990 [44]	No information
Repeated squatting	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability to the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Single squat	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability to the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
	Paatelma et al. 2010 [39]	Patient stands with his feet apart and holds the therapist by the hands. The patient is asked to squat so that the heels are in contact with the ground at all times, and then the same when the patient is standing on toes. If this is difficult, make a finding (2); otherwise, normal (1).
Sit-up	Viikari-Juntura et al. 1998 [47]	This test assesses the inability to perform a sit-up.
	Waddell et al. 1992 [48]	No information
Single leg standing	Paatelma et al. 2010 [39]	Patient flexes his left hip and knee to 90 degrees and stands for 30 s. If the patient loses balance, make a finding (2); otherwise, normal (1). Report also if a pelvic belt makes a difference. Repeat on left.
Bilateral active straight leg raise	Spratt et al. 1990 [44]	No information
	Waddell et al. 1992 [48]	No information
Loaded forward reach	Smeets et al. 2006 [43]	Patients had to hold a stick with weight 4.5 kg at shoulder height and had to reach forward as far as possible without lifting heels off the floor.
	Simmonds et al. 1998 [42]	Subjects stood next to a wall on which a meter rule was mounted at shoulder height. They then reached forward at shoulder height, holding a weight that did not exceed 5 % of body weight or 4.5 kg. The maximum reach distance was recorded. The subjects had to keep their heels on the floor.
	Teixeira da Cunha-Filho et al. 2010 [46]	In this test, the participant was asked to stand next to a wall while holding a stick with a given weight in both hands at the shoulder level near the body. He or she raised the shoulder to 90° (with the elbows extended) and then reached forward, flexing his or her trunk without stepping forward. The maximum horizontal reaching distance was then recorded in centimeter. This test has been hypothesized to test the compressive forces on the lumbar spine, muscle force, endurance, and ability for static support. A plastic stick with a 2-kg weight was used.

(*Continued*)

Table 2
(*Continued*)

Identified clinical tests	Author	Specification of test
Unloaded forward reach	Simmonds et al. 1998 [42]	Subjects stood next to a wall on which a meter rule was mounted at shoulder height. They then reached forward at shoulder height. The maximum reach distance was recorded. Subjects had to keep their heels on the floor.
Recline reach	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Lifting capacity	Reneman et al. 2005 [41]	A receptacle is lifted from a 76 cm shelf, turning 90°, lowering the receptacle until it touches the floor, lifting toward an upright position, turning back 90°, and returning the receptacle to its original position. This is repeated five times with the same weight. Before every set, the heart rate is recorded. After each set heart rate, duration of the test and self-rating of effort is recorded. The weight in the receptacle is increased in four to seven steps until a maximum was reached.
Carrying bilateral	Durand et al. 2004 [32]	This test is scored by a six-category ordinal scale ranging from very heavy to unable. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Carrying unilateral	Durand et al. 2004 [32]	This test is scored by a six-category ordinal scale ranging from very heavy to unable. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Push	Durand et al. 2004 [32]	This test is scored by a six-category ordinal scale ranging from very heavy to unable. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability aith the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Pull	Durand et al. 2004 [32]	This test is scored by a six-category ordinal scale ranging from very heavy to unable. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Kneeling	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Crawling	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.
Ladder climbing	Durand et al. 2004 [32]	This test is scored by a four-category ordinal scale ranging from constantly to never. In addition to this physical performance score, a score indicating the patient's participation was obtained for each task. This score was calculated by comparing the patient's perception of his or her maximum ability with the therapists' observations of physical signs of maximum effort. The possible scores included "full participation" if there was agreement between the subject and rater, "self-limiting" if the subject did not fully participate, and "overextending" if the patient wanted to continue the task after reaching maximum effort.

Table 3
Reliability of identified clinical tests

			Test-retest reliability		Inter-rater reliability		Intrarater reliability	
Identified clinical tests	Author+year	Specification of test	κ or ICC value	Conclusion	κ or ICC value	Conclusion	κ or ICC value	Conclusion
Clinical tests reported in multiple studies +consistently reported								
Biering Sörensen test	Gruther et al. 2009 [33]		ICC=0.59	Moderate				
	Keller et al. 2001 [36]	Overall	ICC=0.93	Good				
		Men	ICC=0.84	Good				
		Women	ICC=0.95	Good				
	Latimer et al. 1999 [37]				ICC=0.88 (95% CI 0.73–0.95)	Good*		
	Simmonds et al. 1998 [42]				ICC=0.99	Good	ICC=0.91	Good
Extensor endurance test	Del Pozo-Cruz et al. 2014 [31]	Men	ICC=0.97 (95% CI 0.94–0.98)	Good				
		Women	ICC=0.96 (95% CI 0.94–0.98)	Good				
	Ito et al. 1996 [35]		ICC=0.93	Good				
	Paatelma et al. 2010 [39]				κ=0.46 (95% CI 0.13–0.78) PA=70%	Moderate*	κ=0.48 (95% CI 0.16–0.77) PA=67%	Moderate*
Flexor endurance test	Del Pozo-Cruz et al. 2014 [31]	Men	ICC=0.97 (95% CI 0.96–0.99)	Good				
		Women	ICC=0.96 (95% CI 0.92–0.99)	Good				
	Ito et al. 1996 [36]		ICC=0.90	Good				
5-min Walking test	Smeets et al. 2006 [43]		ICC=0.89 (95% CI 0.81–0.93)	Good				
	Teixeira da Cunha-Filho et al. 2010 [46]		ICC=0.99	Good				
50-ft Walking test	Smeets et al. 2006 [43]		ICC 0.76 (95% CI 0.61–0.85)	Good				
	Simmonds et al. 1998 [42]	Fast speed			ICC=0.99	Good	ICC=0.99	Good
		Preferred speed			ICC=0.98	Good	ICC=0.95	Good
	Teixeira da Cunha-Filho et al. 2010 [46]		ICC=0.96	Good				
Shuttle walk test	Pratt et al. 2002 [40]		ICC=0.92	Good*				
	Taylor et al. 2001 [45]		ICC=0.99	Good				
Sit-to-stand test	Smeets et al. 2006 [43]		ICC=0.91 (95% CI 0.84–0.94)	Good				
	Simmonds et al. 1998 [42]				ICC=0.99	Good	ICC=0.45	Moderate
	Teixeira da Cunha-Filho et al. 2010 [46]		ICC=0.99	Good				
Loaded forward reach	Smeets et al. 2006 [43]		ICC=0.74 (95% CI 0.59–0.84)	Good				
	Simmonds et al. 1998 [42]				ICC=0.99	Good	ICC=0.99	Good
	Teixeira da Cunha-Filho et al. 2010 [46]		ICC=0.98	Good				
Clinical tests reported in multiple studies+inconsistently reported								
Time up and go	Simmonds et al. 1998 [42]				ICC=0.99	Good	ICC=0.98	Good
	Teixeira da Cunha-Filho et al. 2010 [46]		ICC=0.92	Good				
Walking ability	Simmonds et al. 1998 [42]				ICC=0.99	Good	ICC=0.45	Moderate
	Teixeira da Cunha-Filho et al. 2010 [46]		ICC=0.99	Good				
Stair climbing	Durand et al. 2004 [32]	Level of work			κ=0.80 (95% CI 0.62–0.99) PA=90%	Good*		
		Subject participation			κ=0.66 (95% CI 0.04–1.00) PA=98%	Good*		
	Smeets et al. 2006 [43]		ICC 0.96 (95% CI 0.93–0.98)	Good				

(*Continued*)

Table 3
(*Continued*)

Identified clinical tests	Author+year	Specification of test	Test-retest reliability		Inter-rater reliability		Intrarater reliability	
			κ or ICC value	Conclusion	κ or ICC value	Conclusion	κ or ICC value	Conclusion
Progressive isoinertial lifting evaluation	Hodselmans et al. 2007 [34]		ICC=0.94 (95% CI 0.85–0.97)	Good*				
	Smeets et al. 2006 [43]		ICC 0.92 (95% CI 0.87–0.96)	Good				
Floor-to-waist lift test	Durand et al. 2004 [32]	Level of work			κ=0.72 (95% CI 0.55–0.89) PA=80%	Good*		
		Subject participation			κ=0.62 (95% CI 0.41–0.83) PA=75%	Good*		
	Magnussen et al. 2004 [38]		κ=0.55 (95% CI 0.51–0.59)	Moderate	κ=1.00 (95% CI 1.00–1.00)	Excellent		
	Smith 1994 [50]				κ=0.64 PA=82%	Good*	κ=0.73 PA=87%	Good*
Repeated trunk flexion	Spratt 1990 [44]		PA=62%	Moderate*	1 PA=65%	Moderate*		
	Simmonds et al. 1998 [42]				ICC=0.99	Good	ICC=0.45	Moderate[a]
	Teixeira da Cunha-Filho et al. 2010 [46]		ICC=0.99	Good				
Single squat	Durand et al. 2004 [32]	Level of work			κ=0.65 (95% CI 0.48–0.88) PA=82%	Good*		
		Subject participation			κ=0.55 (95% CI 0.34–0.81) PA=77%	Moderate*		
	Paatelma et al. 2010 [39]				κ=1.0 (95% CI 1.0–1.0) PA=100%	Excellent*	κ=1.0 (95% CI 1.0–1.0) PA=100%	Excellent*
Sit-up	Viikari-Juntura et al. 1998 [47]				κ=0.62	Good*		
	Waddell et al. 1992 [48]				κ=0.48 PA=85%	Moderate		
Bilateral active straight leg raise	Spratt et al. 1990 [44]		PA=71%	Good*	PA=65%	Moderate*		
	Waddell et al. 1992 [48]				κ=0.77 PA=95%	Good		
Clinical tests reported in a single study								
30-s Chair stand test	Kahraman et al. 2016 [49]						ICC=0.94 (95% CI 0.89–0.97)	Good
Sock test	Magnussen et al. 2004 [38]		κ=0.73 (95% CI 0.53–0.95)	Good	κ=0.95 (95% CI 0.86–1.04)	Excellent		
Pick-up test	Magnussen et al. 2004 [38]		κ=0.70 (95% CI 0.48–0.94)	Good	κ=0.90 (95% CI 0.78–1.04)	Excellent		
Roll-up test	Magnussen et al. 2004 [38]		κ=0.83 (95% CI 0.66–1.02)	Excellent	κ=1.00 (95% CI 1.00–1.00)	Excellent		
Unloaded reach	Simmonds et al. 1998 [42]				ICC=0.99	Good	ICC=0.98	Good
Recline reach	Durand et al. 2004 [32]	Level of work			κ=0.76 (95% CI 0.62–0.95) PA=87%	Good*		
		Subject participation			κ=0.64 (95% CI 0.37–1.00) PA=95%	Good*		
Repeated trunk extension	Spratt et al. 1990 [44]		PA=75%	Good*	PA=56%	Moderate*		
Repeated trunk rotation seated	Durand et al. 2004 [32]	Level of work			κ=0.37 (95% CI 0.05–0.70) PA=79%	Fair*		
		Subject participation			κ=0.65 (95% CI 0.39–1.00) PA=97%	Good*		
Repeated trunk rotation standing	Durand et al. 2004 [32]	Level of work			κ=0.54 (95% CI 0.25–0.83) PA=82%	Moderate*		
		Subject participation			κ=0.48 (95% CI −0.13–1.00) PA=95%	Moderate*		
Single partial push-up	Spratt et al. 1990 [44]		PA=78%	Good*	PA=94%	Good*		

(*Continued*)

Table 3
(*Continued*)

			Test-retest reliability		Inter-rater reliability		Intrarater reliability	
Identified clinical tests	Author+year	Specification of test	κ or ICC value	Conclusion	κ or ICC value	Conclusion	κ or ICC value	Conclusion
Repeated partial push-up	Spratt et al. 1990 [44]		PA=65%	Moderate*	PA=87%	Good*		
Repeated squatting	Durand et al. 2004 [32]	Level of work			κ=0.70 (95% CI 0.50–0.90) PA=83%	Good*		
		Subject participation			κ=0.63 (95% CI 0.36–0.90) PA=85%	Good*		
Single leg standing	Paatelma et al. 2010 [39]				κ=0.67 (95% CI 0.32–1.00) PA 84%	Good*	κ=0.59 (95% CI 0.04–0.89) PA 90%	Moderate*
Lifting capacity	Reneman et al. 2005 [41]	CR10 rating			ICC=0.76 (95% CI 0.69–0.83)	Good*		
		Categorical ranking			κ=0.50	Moderate*		
Carrying bilateral	Durand et al. 2004 [32]	Level of work			κ=0.84 (95% CI 0.70–0.98) PA=90%	Excellent*		
		Subject participation			κ=0.32 (95% CI 0.08–0.55) PA=55%	Fair*		
Carrying unilateral	Durand et al. 2004 [32]	Level of work			κ=0.73 (95% CI 0.56–0.91) PA=83%	Good*		
		Subject participation			κ=0.42 (95% CI 0.19–0.65) PA=63%	Moderate*		
Push	Durand et al. 2004 [32]	Level of work			κ=0.82 (95% CI 0.65–0.99) PA=90%	Excellent*		
		Subject participation			κ=0.25 (95% CI 0.02–0.49) PA=53%	Fair*		
Pull	Durand et al. 2004 [32]	Level of work			κ=0.79 (95% CI 0.60–0.98) PA=90%	Good*		
		Subject participation			κ=0.60 (95% CI 0.39–0.81) PA=75%	Moderate*		
Kneeling	Durand et al. 2004 [32]	Level of work			κ=0.83 (95% CI 0.68–0.98) PA=90%	Excellent*		
		Subject participation			κ=0.81 (95% CI 0.56–1.00) PA=95%	Excellent*		
Crawling	Durand et al. 2004 [32]	Level of work			κ=0.78 (95% CI 0.63–1.00) PA=	Good*		
		Subject participation			92% κ=0.64 (95% CI 0.38–1.00) PA=97%	Good*		
Ladder climbing	Durand et al. 2004 [32]	Level of work			κ=0.47 (95% CI 0.22–0.74) PA=72%	Moderate*		
		Subject participation			κ=0.38 (95% CI –0.13–1.00) P=95%	Fair*		

κ, kappa; ICC, intracorrelation coefficient; PA, percentage agreement; CI, confidence interval.
* Reliability reported in a study with an overall poor methodological quality.

the lack of description of how missing data were handled (17/20) [31,32,34–39,41–48,50]; third, the lack of an implementation of two measurement moments (3/20) [32,41,50]; fourth, issues regarding the time interval between the two measurements (2/20) [33,34]; and finally, issues regarding the statistical method (2/20) [32,44] (Table 4).

Discussion

To our knowledge, this is the first review that aimed to provide a complete overview of the reliability for physical functioning tests specifically in patients with LBP. A total of 20 studies were included representing 38 different clinical tests. Clinical tests with an overall good test-retest reliability in at least two independent studies were the *extensor endurance test*, the *flexor endurance test*, the *5-minute walking test*, the *50-ft walking test*, the *shuttle walk test*, the *sit-to-stand test*, the *loaded forward reach*, and the *progressive isoinertial lifting evaluation test*. Tests with an overall good inter-rater reliability were the *Biering-Sörensen test* and the *floor-to-waist lift test*. None of the identified clinical tests could be identified to have an overall good intrarater reliability in at least two independent studies.

Quality of the evidence

All included studies were scored as having either a fair or a poor overall methodological quality according to our scoring with the COSMIN checklist. These low scores might be due to a strict application of the COSMIN scoring method. The scoring method as it was used in this review has been proposed by Terwee et al. [30]. These researchers developed a specific scoring system that replaces the original dichotomous response options into four options representing excellent, good, fair, and poor methodological qualities. Subsequently, they provide an option to assess an overall methodological quality that is obtained by taking the lowest rating of any of the assessed questions ("worst score counts"). However, to prevent this method for being too strict, Terwee et al. defined a series of possible fatal flaws (such as an inadequate statistical analysis or inappropriate sample size), which may occur in a study design. Only these fatal flaws could be assessed as being a sign of a poor methodological quality, which was in this review the case for a total of nine studies. Terwee et al. also considered an alternative method that calculates a mean score instead of the worst score count method. However, a disadvantage of this method is that fatal flaws in the design or analyses can be compensated by other good design aspects. This method was finally considered as being undesirable according to the COSMIN Delphi panel. Hence, the COSMIN checklist grades the importance of the different subitems that it investigates, and therefore, we believe that the COSMIN checklist, together with the proposed scoring system, is an adequate tool for assessing the methodological quality of the included studies.

Based on the overall disappointing methodological quality of the included studies, the decision was made not to come to final conclusions on the reliability of a clinical test if it was assessed in one study. Hence, if a clinical test was assessed within multiple studies with at least one study being of a fair methodological quality, a final conclusion for the reliability of that specific could be made.

Reliability of the physical functioning tests

This review identified a total of 10 clinical tests (8 of which are investigated for test-retest reliability and 2 for inter-rater reliability) with an overall good outcome for reliability in at least 2 independent studies. Final conclusions for the reliability of these physical functioning tests are further discussed.

The *test-retest reliability* of the *progressive isoinertial lifting* evaluation was described within two studies [34,43] with one study having a fair [43] and the other a poor [34] methodological quality. Unfortunately, the execution and interpretation differed within these two articles, which makes it very difficult to come to a final conclusion for this test. However, this issue was not present for the remaining seven clinical tests, which were identified for having good outcome in test-retest reliability. Conclusively, good test-retest reliability was finally concluded for the *extensor endurance test*, the *flexor endurance test*, the *5-minute walking test*, the *50-ft walking test*, the *shuttle walk test*, the *sit-to-stand test*, and the *loaded forward reach test*.

The *inter-rater reliability* of the *floor-to-waist lift test* was rated as being good within three independent studies [32,38,50]. Out of these studies, one was identified with a fair [38] and the other two with a poor [32,50] overall methodological quality. However, when one compares the description of this test throughout the included studies, some major differences can be noted. Therefore, it is not possible to make a general conclusion for this test and it cannot be identified as having a good inter-rater reliability. Conclusively, the *Biering-Sörensen test* was the only test that could be concluded to have good inter-rater reliability.

Intrarater reliability was always investigated within a single study, which makes it impossible to come to a final conclusion. In this review, intrarater reliability was the least investigated form of reliability. A possible reason might be that, in the literature, clinical tests are often used in randomized controlled trials or multicenter trials and are therefore applied by more than one researcher. This implies the importance for the investigation of test-retest or inter-rater reliability because it seems to be more relevant for scientific research. In clinical practice, however, patients are often followed up by the same health-care worker and intrarater reliability gains importance. Conclusively, additional studies addressing the intrarater reliability are highly needed and recommended [51].

Finally, the *Biering-Sörensen test* was the most investigated test in this review. However, even though it has been

Table 4
Overview of the risk of bias assessment with the COSMIN checklist

	Del Pozo-Cruz et al. 2014 [31]	Durand et al. 2004 [32]	Gruther et al. 2009 [33]	Hodselmans et al. 2007 [34]	Ito et al. 1996 [35]	Kahraman et al. 2016 [49]	Keller et al. 2001 [36]	Latimer et al. 1999 [37]	Magnussen et al. 2004 [38]	Pratt et al. 2002 [40]	Paatelma et al. 2010 [39]	Reneman et al. 2005 [41]	Smeets et al. 2006 [43]	Smith 1994 [50]	Spratt et al. 1990 [44]	Simmonds et al. 1998 [42]	Taylor et al. 2001 [45]	Teixeira da Cunha-Filho et al. 2010 [46]	Viikari-Juntura et al. 1998 [47]	Waddell et al. 1992 [48]
Percentage of missing items described?	G	G	E	G	G	G	G	G	E	E	G	G	G	G	G	G	G	G	G	G
Description of handling missing items?	F	F	E	F	F	G	F	F	F	E	F	F	F	F	F	F	F	F	F	F
Was the sample size adequate?	F	F	F	P	E	F	F	P	F	P	P	P	G	P	F	F	F	F	P	E
At least two measurements?	E	P	E	E	E	E	E	E	E	E	E	P	E	P	E	E	E	E	E	E
Administrations independent?	E	N/A	G	G	G	G	G	G	G	G	G	N/A	G	N/A	G	G	G	G	G	G
Time interval stated?	E	N/A	E	E	E	E	E	E	E	E	E	N/A	E	N/A	E	E	E	E	E	E
Were patients stable in-between measurements?	G	N/A	F	G	G	E	G	G	E	G	G	N/A	E	N/A	G	G	G	G	G	G
Time interval appropriate?	E	N/A	F	F	E	E	E	E	E	E	E	N/A	E	N/A	E	G	E	E	E	E
Conditions similar for both measurements?	E	N/A	F	E	G	E	G	G	G	G	G	N/A	E	N/A	G	G	G	G	G	G
Any important flaws in design?	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
ICC for continuous scores?	G	N/A	G	E	G	E	E	E	E	E	N/A	E	E	N/A	N/A	E	E	E	E	E
Kappa for dichotomous, ordinal, or nominal scores?	E	E	N/A	N/A	N/A	N/A	N/A	N/A	E	N/A	E	E	N/A	E	P	N/A	N/A	N/A	E	E
Weighted kappa for ordinal scores?	N/A	F	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	E	N/A	N/A	P	N/A	N/A	N/A	N/A	N/A
Weighting scheme described for ordinal scores?	N/A	F	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	G	N/A	N/A	P	N/A	N/A	N/A	N/A	N/A
Final score	F	P	F	P	F	F	F	P	F	P	P	P	F	P	P	F	F	F	P	F

E, excellent; G, good; F, fair; P, poor; N/A, not applicable; ICC, intracorrelation coefficient.

investigated in four independent studies, only good interrater reliability could be concluded. The other forms of reliability were either inadequately investigated or showed heterogeneous outcome results. Actually, none of the 37 included tests could be concluded to have overall good test-retest, interrater and intrarater reliabilities. There seems to be an oversupply of physical functioning tests that are not adequately investigated for their clinimetric properties. This implies a lack of uniformity in outcome results when physical functioning is investigated with the use of clinical testing. It is important to downgrade the oversupply of clinical tests and to work toward the further investigation of some adequate tests. A first step in this process was completed in this review by providing some general conclusions for the reliability of physical functioning tests. These tests can then be further investigated and, finally, they might form a core set of clinical tests that are usable for both clinicians and researchers.

Clinical implications

Clinical testing should be an important factor in the assessment of physical functioning. In contrast to this review, which specifically investigated reliability in a patient population with LBP, many studies will investigate a clinical test in a healthy population sample. Therefore, reference data are often only established for healthy individuals, which makes it very difficult to make recommendations for these clinical tests in a specific patient population such as LBP. This review serves as a first step in the process of identifying a series of usable clinical tests specifically for patients with LBP. Moreover, an advantage for the use of these clinical tests is that they are easily applicable and that they can be performed without any additional (expensive) devices. Consequently, they can be considered as a cost-effective method to evaluate the physical functioning level in patients with LBP. All recommended tests are easy to execute and can be assessed within every clinical setting. Although this review serves as a starting point for further research, clinicians and researchers should already be encouraged to implement the recommended tests into practice to help them to provide a more comprehensive insight in the physical functioning level of a patient.

Directions for further research

The assessment of reliability is only the first step in the recommendation process for the use of clinical tests. In future research, the recommended clinical tests in the current review should be further investigated for their validity. Only then can a final conclusion regarding the clinical and scientific use be made.

As stated previously, physical functioning is best defined as an individual's ability to perform activities required in daily life [6–8]. These activities can be defined as bodily movement resulting from a contraction of the skeletal muscles that results in an increase in energy expenditure above resting levels [9,10]. To investigate the concurrent validity of physical functioning tests, one can compare its outcome with a direct measurement of the energy expenditure level [52]. The energy expenditure level can be, for example, calculated with the doubly labeled water method, which is known as an accurate field technique for the determination of free-living energy and is often used as a golden standard [53]. However, this test has a disadvantage in that it only provides a global measure of the energy expenditure level and it cannot detect, for example, a pattern of physical functioning [52]. If one wants to investigate other patterns of physical functioning (such as frequency, intensity, and duration), studies can use accelerometers and heart rate monitors or a combination of these two to assess the concurrent validity of a physical functioning test [52]. Unfortunately, a single perfect gold standard for the assessment of the validity of physical functioning tests does not exist, so directions for further research are based on the combined use of these instruments [52,54].

A second direction for further research is to assess the construct validity of the recommended physical functioning tests in this review by assessing their direct relationship [55–57]. If one can identify strong correlations between different clinical tests, it can be assumed that these tests measure the same construct, and therefore, these tests can be identified for having a good construct validity.

Finally, future research should also focus on investigating the correlations of popular PROMs in patients with LBP with the recommended clinical tests in this review [42,57,58]. The use of PROMs has some disadvantages, such as social desirability bias and errors in self-observation [15–17], which makes it interesting to assess how well these two measurement options are actually correlating with each other.

Study limitations and potential bias in the review process

First, none of the included studies could be rated as having overall good or excellent scores for all the subitems within the COSMIN reliability checklist. This finding makes the interpretation of the results in this review difficult and expresses the need for new studies, which have an appropriate study design, adequate sample size, and appropriate statistical analysis.

Second, based on these overall disappointing methodological quality scores, the decision was made not to make any final conclusions for clinical tests that were investigated within a single study. Therefore, it might be possible that some potential reliable tests were not identified as being reliable because they were not investigated appropriately. This phenomenon was also reported within a review regarding the reliability assessment of clinical tests associated with motor control impairment and lumbar instability [51].

Third, because of heterogeneous descriptions of the tests, we might have misinterpreted the final conclusions on these tests. When tests are interpreted and described inconsistently, it is very difficult to compare these study results.

Fourth, all studies in this review report the reliability of clinical tests specifically in patients with LBP. The rationale for this decision is that reliability outcome might differ for a healthy compared with a specific patient population. For example, Gruther et al. [33] state that there are reasons to believe that reliability will be lower if investigated in patient-specific populations compared with healthy subjects. It is important to note that many studies investigate clinical tests for reliability in a healthy population, which may have resulted in less studies meeting the inclusion criteria for this review.

Finally, apart from the brief search in the ClincalTrials.gov database, no additional search was performed for gray literature. The search terms and the inclusion criteria narrowed the search result, possibly missing relevant articles. Further, only studies in English or Dutch were included, also resulting in a potential bias. A more open search strategy may have identified more studies, giving a wider range of results. The present search terms were, however, considered relevant according to the objectives.

Conclusions

The present study is, to our knowledge, the first systematic review that assessed methodological quality and summarized the results of test-retest, inter-rater, and inter-rater reliabilities of physical functioning tests in subjects with LBP. Good test-retest reliability was concluded for the *extensor endurance test*, the *flexor endurance test*, the *5-minute walking test*, the *50-ft walking test*, the *shuttle walk test*, the *sit-to-stand test,* and the *loaded forward reach test*. For inter-rater reliability, only one test, namely, the *Biering-Sörensen test*, could be concluded to have an overall good inter-rater reliability. None of the identified clinical tests could be concluded to have good intrarater reliability. Further investigation should focus on a better overall study methodology, and it would be highly beneficial for authors to use identical protocols for the description of clinical tests to be able to generalize results and to compare them in-between different studies. The assessment of reliability is only the first step in the recommendation process for the use of clinical tests. In future research, the identified clinical tests in the current review should be further investigated for validity. Only when these clinimetric properties of a clinical test have been thoroughly investigated can a final conclusion regarding the clinical and scientific use of the identified tests be made.

Supplementary material

Supplementary material related to this article can be found at http://doi.org/10.1016/j.spinee.2017.08.257.

References

[1] Glocker C, Kirchberger I, Glassel A, et al. Content validity of the comprehensive International Classification of Functioning, Disability and Health (ICF) core set for low back pain from the perspective of physicians: a Delphi survey. Chronic Illn 2013;9:57–72.

[2] Hayden JA, Chou R, Hogg-Johnson S, et al. Systematic reviews of low back pain prognosis had variable methods and results: guidance for future prognosis reviews. J Clin Epidemiol 2009;62:781–96, e1.

[3] Hayden JA, Dunn KM, van der Windt DA, et al. What is the prognosis of back pain? Best Pract Res Clin Rheumatol 2010;24:167–79.

[4] Denteneer L, Van Daele U, De Hertogh W, et al. Identification of preliminary prognostic indicators for back rehabilitation in patients with nonspecific chronic low back pain: a retrospective cohort study. Spine 2016;41:522–9.

[5] Bagraith KS, Hayes J, Strong J. Mapping patient goals to the International Classification of Functioning, Disability and Health (ICF): examining the content validity of the low back pain core sets. J Rehabil Med 2013;45:481–7.

[6] Painter P, Stewart AL, Carey S. Physical functioning: definitions, measurement, and expectations. Adv Ren Replace Ther 1999;6:110–23.

[7] Stewart AL, Painter PL. Issues in measuring physical functioning and disability in arthritis patients. Arthritis Care Res 1997;10:395–405.

[8] Cieza A, Stucki G, Weigl M, et al. ICF Core Sets for low back pain. J Rehabil Med 2004;44(Suppl.):69–74.

[9] Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Health Rep 1985;100:126–31.

[10] Pinheiro Volp AC, Esteves de Oliveira FC, Duarte Moreira Alves R, et al. Energy expenditure: components and evaluation methods. Nutr Hosp 2011;26:430–40.

[11] Roland M, Fairbank J. The Roland-Morris disability questionnaire and the oswestry disability questionnaire. Spine 2000;25:3115–24.

[12] Beurskens AJ, de Vet HC, Koke AJ, et al. Measuring the functional status of patients with low back pain. Assessment of the quality of four disease-specific questionnaires. Spine 1995;20:1017–28.

[13] Beurskens AJ, de Vet HC, Koke AJ. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996;65:71–6.

[14] Schoppink LE, van Tulder MW, Koes BW, et al. Reliability and validity of the Dutch adaptation of the Quebec Back Pain Disability Scale. Phys Ther 1996;76:268–75.

[15] Stone AA, Shiffman S, Schwartz JE, et al. Patient compliance with paper and electronic diaries. Control Clin Trials 2003;24:182–99.

[16] Buer N, Linton SJ. Fear-avoidance beliefs and catastrophizing: occurrence and risk factor in back pain and ADL in the general population. Pain 2002;99:485–91.

[17] Heneweer H, Picavet HS, Staes F, et al. Physical fitness, rather than self-reported physical activities, is more strongly associated with low back pain: evidence from a working population. Eur Spine J 2012; 21:1265–72.

[18] Applegate WB, Blass JP, Williams TF. Instruments for the functional assessment of older patients. N Engl J Med 1990;322:1207–14.

[19] Painter P. Physical functioning in end-stage renal disease patients: update 2005. Hemodial Int 2005;9:218–35.

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

[21] O'Sullivan P. Diagnosis and classification of chronic low back pain disorders: maladaptive movement and motor control impairments as underlying mechanism. Man Ther 2005;10:242–55.

[22] Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264–9, w64.

[23] Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test-retest reliability of continuous measurements. Stat Med 2002;21:3431–46.

[24] Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil 1998;12:187–99.

[25] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[26] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
[27] Fleiss JL. Reliability of measurement. In: Fleiss JL, editor. The design and analysis of clinical experiments. New York: John Wiley & Sons, Inc; 1999. p. 1–32.
[28] Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res 2010;19:539–49.
[29] Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
[30] Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 2012;21:651–7.
[31] Del Pozo-Cruz B, Mocholi MH, Del Pozo-Cruz J, et al. Reliability and validity of lumbar and abdominal trunk muscle endurance tests in office workers with nonspecific subacute low back pain. J Back Musculoskeletal Rehabil 2014;27:399–408.
[32] Durand MJE, Loisel P, Poitras S, et al. The interrater reliability of a functional capacity evaluation: the physical work performance evaluation. J Occup Rehabil 2004;14:119–29.
[33] Gruther W, Wick F, Paul B, et al. Diagnostic accuracy and reliability of muscle strength and endurance measurements in patients with chronic low back pain. J Rehabil Med 2009;41:613–19.
[34] Hodselmans AP, Dijkstra PU, van der Schans C, et al. Test-retest reliability of psychophysical lift capacity in patients with non-specific chronic low back pain and healthy subjects. J Rehabil Med 2007;39:133–7.
[35] Ito T, Shirado O, Suzuki H, et al. Lumbar trunk muscle endurance testing: an inexpensive alternative to a machine for evaluation. Arch Phys Med Rehabil 1996;77:75–9.
[36] Keller A, Hellesnes J, Brox JI. Reliability of the isokinetic trunk extensor test, Biering-Sørensen test, and Åstrand bicycle test: assessment of intraclass correlation coefficient and critical difference in patients with chronic low back pain and healthy individuals. Spine 2001;26:771–7.
[37] Latimer J, Maher CG, Refshauge K, et al. The reliability and validity of the Biering-Sorensen test in asymptomatic subjects and subjects reporting current or previous nonspecific low back pain. Spine 1999;24:2085–9, discussion 2090.
[38] Magnussen L, Strand LI, Lygren H. Reliability and validity of the back performance scale: observing activity limitation in patients with back pain. Spine 2004;29:903–7.
[39] Paatelma M, Karvonen E, Heinonen A. Inter- and intra-tester reliability of selected clinical tests in examining patients with early phase lumbar spine and sacroiliac joint pain and dysfunction. Adv Physiother 2010;12:74–80.
[40] Pratt RK, Fairbank JCT, Virr A. The reliability of the Shuttle Walking Test, the Swiss Spinal Stenosis Questionnaire, the Oxford Spinal Stenosis Score, and the Oswestry Disability Index in the assessment of patients with lumbar spinal stenosis. Spine 2002;27:84–91.
[41] Reneman MF, Fokkens AS, Dijkstra PU, et al. Testing lifting capacity: validity of determining effort level by means of observation. Spine 2005;30:E40–6.
[42] Simmonds MJ, Olson SL, Jones S, et al. Psychometric characteristics and clinical usefulness of physical performance tests in patients with low back pain. Spine 1998;23:2412–21.
[43] Smeets RJ, Hijdra HJ, Kester AD, et al. The usability of six physical performance tasks in a rehabilitation population with chronic low back pain. Clin Rehabil 2006;20:989–97.
[44] Spratt KF, Lehmann TR, Weinstein JN, et al. A new approach to the low-back physical examination. Behavioral assessment of mechanical signs. Spine 1990;15:96–102.
[45] Taylor S, Frost H, Taylor A, et al. Reliability and responsiveness of the shuttle walking test in patients with chronic low back pain. Physiother Res Int 2001;6:170–8.
[46] Teixeira da Cunha-Filho I, Lima FC, Guimaraes FR, et al. Use of physical performance tests in a group of Brazilian Portuguese-speaking individuals with low back pain. Physiother Theory Pract 2010;26:49–55.
[47] Viikari-Juntura E, Takala EP, Riihimäki H, et al. Standardized physical examination protocol for low back disorders: feasibility of use and validity of symptoms and signs. J Clin Epidemiol 1998;51:245–55.
[48] Waddell G, Somerville D, Henderson I, et al. Objective clinical evaluation of physical impairment in chronic low back pain. Spine 1992;17:617–28.
[49] Kahraman T, Ozcan Kahraman B, Salik Sengul Y, et al. 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 2016;39:165–70.
[50] Smith RL. Therapists ability to identify safe maximum lifting in low back pain patients during functional capacity evaluation. J Orthop Sports Phys Ther 1994;19:277–81.
[51] Denteneer L, Stassijns G, De Hertogh W, et al. Inter- and intra-rater reliability of clinical tests associated to functional lumbar segmental instability and motor control impairment in patients with LBP: a systematic review. Arch Phys Med Rehabil 2017;98:151–64, e6.
[52] Bassett DR Jr. Validity and reliability issues in objective monitoring of physical activity. Res Q Exerc Sport 2000;71:S30–6.
[53] Leenders NY, Sherman WM, Nagaraja HN. Energy expenditure estimated by accelerometry and doubly labeled water: do they agree? Med Sci Sports Exerc 2006;38:2165–72.
[54] Plasqui G, Westerterp KR. Physical activity assessment with accelerometers: an evaluation against doubly labeled water. Obesity (Silver Spring) 2007;15:2371–9.
[55] van Hedel HJ, Wirz M, Dietz V. Assessing walking ability in subjects with spinal cord injury: validity and reliability of 3 walking tests. Arch Phys Med Rehabil 2005;86:190–6.
[56] Novy DM, Simmonds MJ, Lee CE. Physical performance tasks: what are the underlying constructs? Arch Phys Med Rehabil 2002;83:44–7.
[57] Filho IT, Simmonds MJ, Protas EJ, et al. Back pain, physical function, and estimates of aerobic capacity: what are the relationships among methods and measures? Am J Phys Med Rehabil 2002;81:913–20.
[58] Lee CE, Simmonds MJ, Novy DM, et al. Self-reports and clinician-measured physical function among patients with low back pain: a comparison. Arch Phys Med Rehabil 2001;82:227–31.