European Spine Journal (2021) 30:957–989
https://doi.org/10.1007/s00586-020-06712-0

REVIEW ARTICLE

# Psychometric properties of chronic low back pain diagnostic classification systems: a systematic review

**Ahmed Omar Abdelnaeem[1,4] · Aliaa Rehan Youssef[1,2] · Nesreen Fawzy Mahmoud[1] · Nadia Abdalazeem Fayaz[1] · Robert Vining[3]**

Received: 23 November 2020 / Revised: 23 November 2020 / Accepted: 27 December 2020 / Published online: 20 January 2021


## Abstract

**Objectives** To identify and critically appraise studies evaluating psychometric properties of functionally oriented diagnostic classification systems for Non-Specific Chronic Low Back Pain (NS-CLBP).

**Methods** This review employed methodology consistent with PRISMA guidelines. Electronic databases and journals: (PubMed, EMBASE, Cochrane, PEDro, CINAHL, Index to chiropractic literature, ProQuest, Physical Therapy, Journal of Physiotherapy, Canadian Physiotherapy and Physiotherapy Theory and Practice) were searched from inception until January 2020. Included studies evaluated the validity and reliability of NS-CLBP diagnostic classification systems in adults. Risk of bias was assessed using a Critical Appraisal Tool.

**Results** Twenty-two studies were eligible: Five investigated inter-rater reliability, and 17 studies analyzed validity of O'Sullivan's classification system (OCS, $n=15$), motor control impairment (MCI) test battery ($n=1$), and Pain Behavior Assessment (PBA, $n=1$). Evidence from multiple low risk of bias studies demonstrates that OCS has moderate to excellent inter-rater reliability (kappa > 0.4). Also, two low risk of bias studies support of OCS-MCI subcategory. Three tests within the MCI test battery show acceptable inter- and intra-rater reliability for clinical use (the "sitting knee extension," the "one leg stance," and the "pelvic tilt" tests). Evidence for the reliability and validity of the PBA is limited to one high bias risk study.

**Conclusions** Multiple low risk of bias studies demonstrate strong inter-rater reliability for OCS classification specifically OCS-MCI subcategory. Future studies with low risk of bias are needed to evaluate reliability and validity of the MCI test battery and the PBA.

**Keywords** Classification · Non-specific chronic low back pain · Pelvic girdle pain · Motor control · Psychometric properties

✉ Ahmed Omar Abdelnaeem
ahmed.omar@pt.cu.edu.eg

Aliaa Rehan Youssef
aliaa.rehan@gmail.com

Nesreen Fawzy Mahmoud
dr_noonfawzy@yahoo.com; dr_nesreenfawzy@cu.edu.eg

Nadia Abdalazeem Fayaz
Nadia.fayaz55@gmail.com

Robert Vining
robert.vining@palmer.edu

[1] Faculty of Physical Therapy, Cairo University, Cairo, Egypt

[2] Faculty of Physical Therapy, Ahram Canadian University, Giza, Egypt

[3] Palmer Center for Chiropractic Research, Palmer College of Chiropractic, Davenport, IA, USA

[4] Cairo, Egypt

## Introduction

Low back pain (LBP) is a highly prevalent problem [1, 2] and a major cause of disability worldwide [1, 3], with a tendency to recur or persist, leading to chronicity [4, 5]. Often, no definitive pathology or underlying mechanism can be identified [6]. To account for uncertain and potentially multiple contributing factors, chronic LBP is commonly labeled as Non-Specific Chronic LBP (NS-CLBP) [1, 7–9]. NS-CLBP is a general diagnosis encompassing a wide range of conditions, symptoms, and clinical features. However, such diagnosis does not distinguish characteristics to guide specific clinical decision-making [10].

There is currently no optimum treatment strategy for NS-CLBP [11]. This is partially due to a lack of standardized, valid, and reliable methods which classify specific characteristics. An effective classification system is expected to

be based on studies reporting selection criteria in clear and standardized terms [12]; requirements that have been proposed as a key research priority [13].

There are many approaches to classify NS-CLBP, such as identifying symptom sources [14–16], functional characteristics, and/or psychosocial risk [17]. Functional evaluation, which generally assesses how people move and perform movement tasks, is typically designed to both inform treatment and be used as a clinical outcome. Functional classification is promising, in part because it identifies characteristics of a condition and offers information about potential mechanisms contributing to chronicity. Active treatments designed to address functional deficits may also positively influence self-efficacy, reduce fear avoidance behaviors, and promote symptom self-management [18, 19].

Despite the potential importance of functional classification as a diagnostic process, to the authors' knowledge, no previous comprehensive evaluation of the reliability and validity of existing classification systems has been performed. Therefore, the aims of this review were to describe and critically appraise the psychometric properties of functionally oriented NS-CLBP diagnostic classification systems.

## Methods

This systematic review was registered on PROSPERO (CRD42015023958) [20] and conducted according to PRISMA guidelines [21].

### Search strategy

A systematic electronic and manual search of the literature published in English, from inception until January 2020, was conducted in the following electronic databases and Journals: PubMed, EMBASE, Cochrane, PEDro, CINAHL, Index to chiropractic literature, ProQuest, Physical Therapy, Journal of Physiotherapy, Canadian Physiotherapy and Physiotherapy Theory and Practice.

The search strategy consisted of keywords and Medical Subject headings (MeSH) related to "Non specific," "mechanical," "low back pain," "simple backache," "lumbar strain," "spinal degeneration," "classification," "clinical test," "clinical examination," "clinical sign," "valid*," and "reliabl*." Searching strategy details are available in Appendix 1. Reference lists of eligible articles were also searched for relevant publications.

### Eligibility criteria

Inclusion and exclusion criteria are summarized in Table 1.

### Data collection and analysis

#### Selection of studies

Studies from electronic databases and manual searches were imported into EndNote X5.0.1 and checked for eligibility by two reviewers (AA and NFM) independently; first by title, abstract, and finally by full text. Discordance was resolved through discussion with co-authors (ARY and RV).

#### Data extraction

Two reviewers (AA and NFM) independently extracted all relevant information into an Excel spreadsheet. All discrepancies were resolved through discussion.

#### Risk of bias assessment

Risk of bias was assessed independently by two authors (AA and NFM) using the Critical Appraisal Tool (CAT) developed by Brink and Louw [22] (Appendix 2). This tool

**Table 1** Studies eligibility criteria

| Inclusion criteria | Exclusion criteria |
|---|---|
| Any study design (observational cross-sectional, cohort and case–control studies) that investigated the reliability and/or validity of NS-CLBP classification systems | Intervention studies, case reports and series, review articles, systematic reviews, conference proceedings, editorials and letters |
| Enrolled adults with NS-CLBP (> 3 months) | Included participants with: |
| Investigated any diagnostic classification system for NS-CLBP (patho-anatomy, motion based, psychosocial classifications, etc.) at an outpatient clinical setting | Potentially confounding comorbidities such as spinal deformity, locomotor system disease or neck pain<br>Radicular symptoms<br>A specific verified spinal diagnosis in the prior 2 months (e.g., fracture, neoplasm and spinal surgery)<br>Any confirmed condition(s) suggesting serious pathology |
| Defined NS-CLBP as chronic pain located between costal margins and gluteal folds | |
| Used any clinical examination approach such as history taking, clinical examination, checklists or questionnaires | |

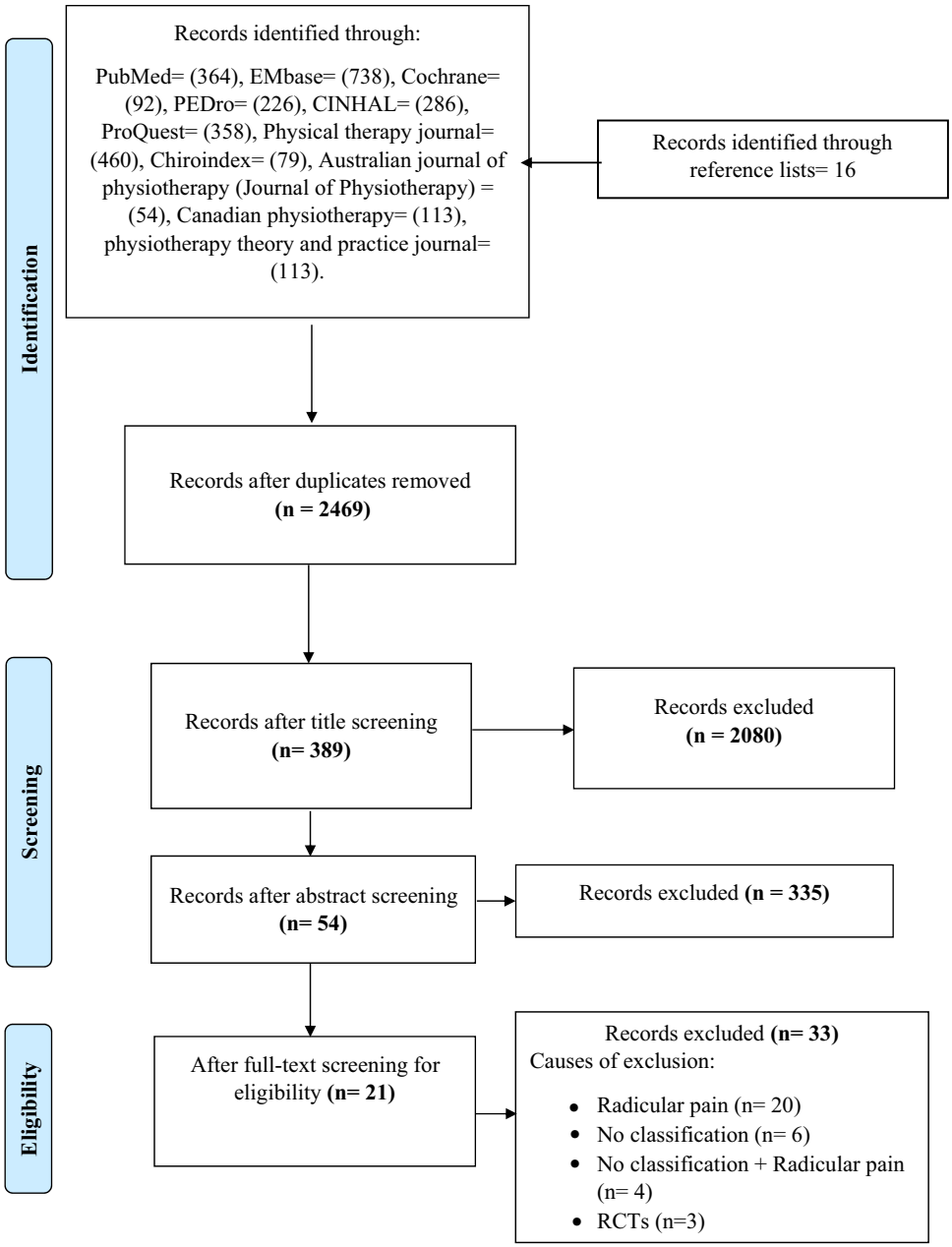



**Fig. 1** PRISMA flow diagram of literature search and studies included

consists of 13 items designed to appraise the quality of the validity and/or reliability studies; 4 items assess reliability studies, 4 items evaluate validation studies, and the remaining 5 items evaluate both validity and reliability studies. Each item is scored as "Yes," "No," or "Not Applicable (N/A)" [22]. Disagreement was resolved through consensus discussion. Studies were considered of low risk of bias if they scored ≥60% [23–25].

## Results

### Selection of studies

Database and manual searching identified a total of 2899 articles. After full article screening, 22 studies published in 21 articles (one article reported 2 studies) were included. The article screening process is outlined in Fig. 1.

### Characteristics of included studies

The methodological approach of each included study is depicted in Tables 2 and 3. Of the 22 included studies, 5

**Table 2** Description of included reliability studies

| References | Reliability type | Classification basis | Participants characteristics | Examiners' qualification | Testing procedure (methods) | Reference test | Classification sub-grouping |
|---|---|---|---|---|---|---|---|
| | Five reliability studies (3 for O'Sullivan classification system and 2 for MCI battery test system) | | | | | | |
| Fersum et al. [26] | Inter-rater | OCS | *N* = 26 NS-CLBP, sex = 11 F and 15 M, mean Age (range) = 32.4 (18–65)Y, pain intensity (Mean) = 6/10, LBP duration (Mean) = 4.9 (Y), ODI score (Mean) = 21.2/100, HSCL (Mean) = 1.53/4, Ørebro (Mean) = 87.5/210 | 4 P.T.s, (mean experience = 12 y, range 7–20 y): Three P.T.s had a master's degree in manual therapy, and one was the system developer. (Classification system training workshops duration ranged from 69 to 140 h, with an average of 106.3 h. Further, a pilot study including 6 participants was conducted) | Assessment of spinal ROM, specific MS. testing, posture, movements and functional tasks<br>Each participant was examined independently twice on 2 days, within a 1-week period | The developer's classification of each participant | Five levels<br>1. Specific or Non-Specific<br>2. Centrally mediated dominant or non-dominant psychosocial<br>3. Peripherally mediated pain<br>4. Pelvic girdle pain with excessive or decreased force closure<br>5. LBP with MCI (a. Flexion, b. flexion with lateral shift, c. active extension, d. passive extension and e. multidirectional dysfunction) or movement impairment (a. Flexion, b. flexion with lateral shift, c. active extension, d. passive extension and e. multidirectional)<br>6. Psychosocial factors contributors (fear avoidance behavior, psychological and social drivers) |

**Table 2** (continued)

| References | Reliability type | Classification basis | Participants characteristics | Examiners' qualification | Testing procedure (methods) | Reference test | Classification sub-grouping |
|---|---|---|---|---|---|---|---|
| Dankart et al. [28] | Inter-rater | OCS | *First study*: $N=35$ NS-CLBP (MCI), sex = 18 F and 17 M, mean Age ± SD = 37(± 12.73), mean BMI = 23.172.2 kg/m, mean LBP duration ± SD = 5.6(± 6.0) Y; Revised-ODI = 37 (± 11) %<br>*Second study*: $N=25$ participants from the first study were randomly selected | *First study*: 2 P.T.s; the classification system developer with 18 y experience in LBP and trained P. T. with 12 y experience in LBP<br>*Second study*: 13 (P.T.s and M.Ds). With different levels of system familiarity<br>Moderate familiar: ($n=8$ (1GP, 1 MD, 3 P.T.s)), mean clinical experience = 20 y (range 10–29 y). All examiners attended the CS developers' clinical workshops<br>Very familiar: ($n=5$ P.T.s), mean clinical experience = 9 y (range 7–11 y). All examiners received a postgraduate training under direct supervision of CS developer | *First study*: A full clinical examination was performed. Participants were then subclassified into one of the five OCS patterns<br>Most participants were evaluated by the second examiner within 24 h or 1 week the latest<br>*Second study*: Twenty-five participants from the first study were randomly selected and videotaped performing a series of postures and functional movements and a case report was written | First study results | MCI:<br>A. Flexion<br>B. Flexion /lateral shifting<br>C. Active extension<br>D. Passive extension<br>E. Multidirectional |
| Luomajoki et al. [27] | Inter- and intra-rater | The individual tests of motor control test battery | *NS-CLBP*: $N=27$, sex = (18F and 9 M), Age (Mean) (SD) = 50.8 (6.2)Y, RMDQ (SD) = 8.5 (5.5)<br>*Controls without LBP*: $N=13$, mean Age (SD) = 55.1 (5.1), Sex = (8F and 5 M) | 4 P.T.s (2 MCI specialists with postgraduate degrees in manual therapy and 25 y of working experience, the other 2 P.T.s had 5 years of experience and underwent 3-day intensive MCI training | 10 MCI tests<br>Raters were blinded to participants' diagnosis and peers' evaluation results. Performance was recorded anonymously, and raters watched each video only once. Reviewed after 2 weeks | | 10 MCI tests (3 tests for flexion and extension control and four tests for rotational control): Waiter's bow, pelvic tilt, one leg stance Rt., one leg stance Lt.,<br>sitting knee extension, rocking backwards, rocking<br>forwards, dorsal tilt of pelvis, prone active knee-flexion and crook lying |

**Table 2** (continued)

| References | Reliability type | Classification basis | Participants characteristics | Examiners' qualification | Testing procedure (methods) | Reference test | Classification sub-grouping |
|---|---|---|---|---|---|---|---|
| Enoch et al. [29] | Inter-rater | The individual tests of motor control test battery | *NS-LBP*: $N=25$, sex = (14 F and 11 M), Age (mean (SD) = 47 (12) Y, *No LBP*: $N=15$, sex = 12 F and 3 M), Age(Mean) (SD) = 45 (19), Pain on day of examination (NRS range 0–10) (*n* (% of group)): No pain (*No LBP*) = 15 (100) 1–3 (*NS-LBP*) = 13 (52) > 3 (*NS-LBP*) = 12 (48) | Danish Manual Therapy Society instructors with 20 years of clinical experience, including conducting these tests for LMC. A pilot testing of 10 subjects with LBP was initially performed in order to become familiar with the test procedures and to reduce examiners' bias | Two examiners applied the five tests for MCI on all subjects. The subjects were examined independently in separate rooms in random order on the same day | | 5 MCI tests: one for repositioning (RPS) and four for dynamic stability, including sitting forward lean (SFL), sitting knee extension (SKE), bent knee fall out (BKFO) and leg lowering (LL) |

N, Number; M, Male; F, Female; y, year; Rt., Right; Lt., Left; P.T, physiotherapist; FP, Flexion Pattern; AEP, Active Extension Pattern; PEP, Passive Extension Pattern; BMI, Body Mass Index; NRS, Numeric Rating Scale; HSCL, Hopkins Symptoms Check List; RMDQ, Ronald Morris disability questionnaire; ODI, Oswestry Disability Index; LMC, Lumbar Motor Control; CS, Classification System; ROM, Range of motion; MS, Muscle(s); MD, Medical doctor; MCI, Motor Control Impairment

investigated inter-rater reliability [26–29], with one study reporting both intra- and inter-rater reliability [27] (Table 2).

Validity was assessed in 17 studies; 15 studies evaluated O'Sullivan's Classification System (OCS) [30–44], one assessed the 10-item Motor Control Impairment (MCI) test battery [45], and one investigated the Pain Behavior Assessment (PBA) classification system [46]. All validity studies were cross-sectional design except 2 studies [42, 45] were cross-sectional case–control design (Table 3).

## Demographic data of participants in eligible studies

Sixteen studies enrolled asymptomatic participants as controls [27, 29–38, 40–43, 45]. Sample size ranged from 12 [39] to 200 [46]. Participant mean age ranged from 28.4 [41] to 55.1 years [27]. The mean body mass index (BMI) ranged from 20.8 [42, 43] to 26.9 kg/m$^2$ [45], although four studies did not report BMI [26, 27, 29, 46].

## Classification systems

All eligible studies described three different classification systems (Tables 2, 3):

1. *OCS* ($n=18$ studies) classifies NS-CLBP as predominantly centrally (e.g., central sensitization) or peripherally mediated (e.g., injury, inflammation of peripheral tissues). OCS also includes a psychosocial assessment step and separates pain presumed from lumbar and pelvic origin. Functional testing of lumbar and pelvic girdle pain evaluates presumed motor control impairment by identifying specific postural and movement characteristics [26] (Appendix 3).
2. *MCI Test Battery* ($n=3$ studies) used specific movements/positions to differentiate participants with MCI from normal individuals. This battery consists of 10 individual tests that identify possible flexion, extension and rotational dysfunction. Assessment is dichotomous (impairment or no impairment) with the severity described on 3 levels (none, mild, moderate/severe) [45].
3. *PBA classification* ($n=1$ study) rates: (1) pain perception; (2) overt pain behavior (e.g., guarding movements); (3) effort during physical test performance; and (4) consistency of behavior across different situations of clinical testing. Categories include no pain, low pain or high pain behaviors [46].

## Reliability of different classification systems

OCS and MCI test Battery inter-rater reliability testing was assessed in 5 studies (Table 4).

Inter-rater reliability of using the entire OCS classification system (all steps) was moderate (kappa ($K$) > 0.4) [26].

**Table 3** Description of included validity studies

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| | *Validity studies (O' Sullivan classification system (OCS): MCI subcategories)* | | | | | | |
| Dankart et al. [38] | Cross section | OCS | *NS-CLBP*: *N* = 33 (20 FP and 13 AEP), sex = (FP = 4 F and 16 M), (AEP = 8 F and 5 M), Age = (FP = 35.7 (11.2)), (AEP = 39.9 (11.3)), BMI = (FP = 24.6 (2.5)) and (AEP = 24.2 (2.8)), VAS(24 h/10) in FP = 4.2 (1.9) and in AEP = 5.7 (2.1), ODI = FP = 36.6%(11.0) and in AEP = 41.2 (14.2), TSK in FP = 40.2 (8.2) and in AEP = 41.3 (8.8). *No LBP*: *N* = 34, sex = (16 F and 18 M), Age = 32.0 (12.2), BMI = 23.3 (2.9) | Two P.T.s subclassified the subjects with NS-CLBP | Laboratory testing consisted of recordings of 3 superficial trunk muscles (TrIO, (sLM) and ICLT) and lumbosacral kinematics (lumbar angle, lower lumbar angle and sacral angle) obtained from a series of postures and movements. Participants underwent the laboratory testing within a week of the clinical examination (performed in a blinded manner) | Mutual agreement between clinicians classification of FP or AEP | No LBP, FP and AEP |
| Dankart et al. [36] | Cross section | OCS | *NS-CLBP*: *N* = 33 (20 FP and 13 AEP), sex = (FP = 4 F and 16 M), (AEP = 8 F and 5 M), Age = (FP = 35.7 (11.2)), (AEP = 39.9 (11.3)), BMI = (FP = 24.6 (2.5)) and (AEP = 24.2 (2.8)), VAS (24 h/10) in FP = 4.2 (1.9) and in AEP = 5.7 (2.1), ODI = FP = 36.6%(11.0) and in AEP = 41.2 (14.2), TSK in FP = 40.2 (8.2) and in AEP = 41.3 (8.8). *No LBP*: *N* = 34, sex = (16 F and 18 M), Age = 32.0 (12.2), BMI = 23.3 (2.9) | Two P.T.s Qualification not reported | comprehensive subjective and physical examinations Participants with a FP or AEP as determined independently by both clinicians were selected. Laboratory testing consisted of recording lumbosacral kinematics (upper lumbar angle, lower lumbar angle and sacral angle) obtained from a series of postures and movements known to aggravate LBP | none | OCS: No LBP, FP and AEP |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| Van Hoof et al. [41] | Cross section | OCS | *NS-CLBP*: *N* = 8, sex = 8 M, Age = 28.3 (8.7), BMI = 22.3 (2.7), Years of cycling = 7.3 (2.5), Average pain (NRS) 4w prior (cycling) = 5.6 (1.2), Average pain (NPRS) 4w prior (ADL) = 3.3 (1.8). *No LBP*: *N* = 9, sex = 9 M, Age = 28.4 (9), BMI = 22.8 (1.9), years of cycling = 8.4 (5.1), | Two P.T.s Qualification not reported | Subjects performed a 2-h outdoor cycling task on a standard flat parcourse. They were guided by a heart rate monitor and instructed to maintain between 60 and 70% of their age-predicted maximum The optimal length of strain gauge was chosen (ranging from 50 to 120 mm) to allow maximal flexion. Lower lumbar kinematics were measured by a remote posture monitoring system, Saddle angle by goniometer and pain level NRS. The pain level was measured at start, every 15 min during cycling and at 30 min, one, two and 24 h after cycling | none | FP VS normal |
| Burnett et al. [35] | Cross section | OCS | *NS-CLBP*: *N* = 9, Sex = (5 F and 4 M), Age = 42.37 (± 9.7), BMI = 22.97 (± 1.7), VAS 1 week prior to examination = 2.3 (± 1.7) *No LBP*: *N* = 9, sex = (5 F and 4 M), Age = 37.67 (± 7.9), BMI = 23.47 (± 2.0) | Two experienced manipulative P.T.s | Prior to data collection, subjects performed MVIC for all trunk muscles (multifidus, ES at (T12 and T9), RA, IO and EO). Subjects were also instructed to ride at 75% of their maximum heart rate and at a cadence between 90 and 100 rpm until the onset of LBP (pain group) or until the general discomfort was too great (non-pain group). EMG and spinal kinematics data were collected at the beginning and then every 5 min throughout the ride. All MVICs were collected for 5 secs, and 3 trials were performed | None | FP VS normal |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| O'Sullivan et al. [30] | Cross section | OCS | *NS-CLBP (FP)*: $N=24$, sex = 24 M, Age = 38.79 (9.24) y, BMI = 26.43 (2.86) kg/m$^2$<br>*No LBP*: $N=21$, sex = 21 M, Age = 38.24 (9.33) y, BMI = 25.05 (3.32) kg/m$^2$ | Number and qualification not reported | A questionnaire designed to obtain subject's activity levels and commonly adopted postures, both at work and home. Subjects were photographed in their natural sitting and maximal slumped sitting postures, natural standing and maximal sway standing postures, and lifting and maximal standing lumbar flexion postures. Measures of lumbar, hip and knee angles were recorded. Trunk muscle endurance was measured with the Biering-Sorensen test<br>Investigators were blinded to group allocation during both testing and data management | None | FP VS normal |
| Hemming et al. [43] | Cross section | OCS | *NS-CLBP*: $N=50$ (23 AEP, 27 FP), sex = (AEP = 19 F and 4 M) (FP = 6 F and 21 M)), Age = (AEP = 43.7 (11.2)) (FP = 41.0 (10.0)), BMI = (AEP = 20.8 (4.9)) (FP = 23.4 (3.5)), Site of Back Pain Right AEP = 8 (34.8%), FP = 5 (18.5%). Left AEP = 2 (8.7%), FP = 3 (11.1%). Central AEP = 13 (56.4%), FP = 19 (70.4%)<br>*No LBP*: $N=28$, sex = (16 F and 12 M), Age = 38.5 (11.2), BMI = 21.5 (4.1) | The lead investigator (RH), a chartered P.T. with 4 years clinical experience, who had received specialist training in the MDCS. The second assessor (LS), a senior P.T./researcher trained in the classification approach and who has published work in this area | All testing was performed at a single visit. Each session took approximately 90–120 min to complete. Nine functional tasks (reach up, sitting-to-standing, standing-to-sitting, step up, step down, box lift, box replace, bending to retrieve and returning from retrieving a pen from the floor) were visually observed and video recorded. A 30-s rest period between each testing condition. Functional tasks were performed randomly | None | Two subclassified MCI (AEP and FP MCI) and a asymptomatic control group |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| O'Sullivan et al. [33] | Cross section | OCS | *NS-CLBP (FP)*: $N=15$, sex = 9 (F and 6 M), Age = 38.8 (± 12), Wt. = 73.9 (± 18.4), height = 171.3 (± 8.0), ODI = 26.1 (± 13.3), Short Form McGill = 15.5 (± 5.6). *No LBP*: $N=15$, sex = 9 (F and 6 M), Age = 38.2 (± 10.9), Wt. = 71.6 (± 11.8), height = 172.3(± 7.1), | Experienced P.T.s | Spinal position sense (SPS) was evaluated using the 3-Space Fastrack Model with participants trying to reproduce a criterion position of neutral lordosis in sitting. Participants were blinded and wore only shorts and undergarments to reduce sensory cues from clothing. They then were positioned into a neutral spinal posture for 5 secs to remember the position. They then relaxed into full lumbar flexion for 5 secs before being asked to reproduce the position. Participants received no feedback on repositioning accuracy during testing | None | Flexion pattern VS normal |
| Sheeran et al. [40] | Cross section | OCS | NS-CLBP: $N=90$ (FP = 51, AEP = 39), sex = (FP = 29 F and 22 M)(AEP = 30 F and 9 M), Age = (FP = 33.0 (10.3)) (AEP = 37.0 (11.4)), BMI = (FP = 25.1 (3.6)) (AEP = 24.9 (3.8)), VAS = (FP = 4.8 (1.3)) (AEP = 4.5 (1.4)), RMDQ = (FP = 7.3 (3.8)) (AEP = 6.2 (3.5)) *No LBP*: $N=35$, sex = (22 F and 13 M), Age = 36.0 (10.3), BMI = 23.3 (2.2) | Spinal Kinematics C7, T12, and S1 spinous processes were identified by the researcher (LS), checked by the P.T. (VS) Qualification not given | SPS (C7, T12, and S1 spinous processes) and trunk EMG (Lt. and Rt. sLM, ICLT, EO, and TrIO) were evaluated during participants' attempts to reproduce a target position of neutral lumbar lordosis and neutral thoracic kyphosis during sitting and standing trials Participants were blindfolded and wore loose clothing to minimize sensory cues. Four repositioning trials in standing and sitting with 5 s of relaxed standing and sitting between each trial were then performed | None | FP, AEP and control |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| O'Sullivan et al. [31] | Cross section | OCS | *NS-CLBP*: *N* = 15, sex = 5F and 10 M, Age = 31.3 (10.3), BMI = 24.3 (3.2), NRS = 3.3 (1.9), ODI (%) = 14.1 (7.8), FABQ = 34.8 (14.4). *No LBP*: *N* = 15, sex = 5F and 10 M, Age = 32.1 (9.2), BMI = 23.8 (2.0) | All participants were examined by 2 investigators<br>Qualification not reported | Prior to testing, all participants with NSCLBP completed measures of average daily pain (NRS, ODI, FABQ and Tampa). Lumbo-pelvic RE was determined using a wireless posture monitor | None | FP and control |
| Sheeran et al. [44] | Cross section | OCS | NS-CLBP: *N* = 87 (FP = 50, PEP = 14 and AEP = 23), sex = (FP = 43%M and 57% F) (AEP = 35%M and 65% F) (PEP = 8%M and 92% F), age = (FP = 33.3 (10.1)) (AEP = 39.7 (12.9)) (PEP = 33.4 (8.3)), BMI = (FP = 25.2 (3.7)) (AEP = 25.0 (3.7)) (PEP = 25.0 (4.5)), VAS = (FP = 4.9 (1.4))(AEP = 4.6 (1.5)) (PEP = 4.6 (1.5)), RMDQ = (FP = 7.3 (3.8)) (AEP = 6.0 (2.8)) (PE = 7.1 (4.7)), No LBP: *N* = 31, sex = (40%M and 60%F), age = (35.2 (9.7)), BMI = (24.8 (2.2)) | Two experienced P.T.s both fully trained in OCS<br>Qualification not reported | Cardiff DST Classifier was employed to identify clinical subgroups of LBP based on repositioning accuracy for participants performing a sitting and standing posture task | The two P.T.s' classification of FP, PEP, AEP and control | FP, PEP, AEP and control |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| Dankart et al. [37] | Cross section | OCS | *NS-CLBP*: *N* = 33 (20 FP and 13 AEP), sex = (FP = 4 F and 16 M), (AEP = 8 F and 5 M), Age = (FP = 35.7 (11.2)), (AEP = 39.9 (11.3)), BMI = (FP = 24.6 (2.5)) and (AEP = 24.2 (2.8)), VAS(24 h/10) in FP = 4.2 (1.9) and in AEP = 5.7 (2.1), ODI = FP = 36.6%(11.0) and in AEP = 41.2 (14.2), TSK in FP = 40.2 (8.2) and in AEP = 41.3 (8.8). *No LBP*: *N* = 34, sex = (16 F and18 M), Age = 32.0 (12.2), BMI = 23.3 (2.9) | 2 P.T.s. (one was the classification system developer) | Participants with NS-CLBP were blindly assessed by 2 P.T.s (active and functional movement tests, articular tests, neurologic examination and tests for spinal motor control). sEMG recordings of 10 superficial trunk muscles were obtained for each subject during "usual" and "slumped" sitting. Three trials of 5 s duration each were conducted, with approximately 1-min rest between each trial. A maximum of 1 week separated examination and laboratory testing | None | No LBP and 2 subgroups of NS-CLBP (FP and AEP) |
| Hemming et al. [42] | Case control | OCS | *NS-CLBP*: *N* = 50 (23 AEP, 27 FP), sex = (AEP = 19 F and 4 M) (FP = 6 F and 21 M)), Age = (AEP = 43.7 (11.2)) (FP = 41.0 (10.0)), BMI = (AEP = 20.8 (4.9)) (FP = 23.4 (3.5)), Site of Back Pain Right AEP = 8 (34.8%), FP = 5 (18.5%). Left AEP = 2 (8.7%), FP = 3 (11.1%). Central AEP = 13 (56.4%), FP = 19 (70.4%). *No LBP*: *N* = 28, sex = (16 F and 12 M), Age = 38.5 (11.2), BMI = 21.5 (4.1) | The lead investigator, a chartered P.T. with 4 y clinical experience, who had received specialist training in the OCS. The second assessor, a senior P.T. trained in the OCS and who has published work in this area | Nine functional tasks were evaluated (reach up, sitting-to-standing, standing-to-sitting, step up, step down, box lift, box replace, bending to retrieve and returning from retrieving a pen from the floor). Each task was repeated three times. sEMG data were collected from spinal extensor (sLM and LT) and abdominal muscles (TrA/IO and EO) bilaterally | None | (AEP and FP MCI) and a asymptomatic control group |
| | *Validity studies (OCS: SIJD subcategories)* | | | | | | |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| O'Sullivan et al. [32] | Cross section | OCS | *SIJD*: *N* = 13, sex = (11 F and 2 M), Age = 32.3 (± 11.2) Y, BMI = 23.8 (± 4.2)KG/M$^2$, Duration of symptoms (months) = 40.8 (± 35.7), Subjects postpartum (*N*) = 5, Subjects posttrauma (*N*) = 13, Subjects with bladder dysfunction (*N*) = 13. *Control*: *N* = 13, sex = (11 F and 2 M), Age = 31.4 (± 11.4)Y, BMI = 22.6 (± 3.5)KG/M$^2$, Subjects postpartum (*N*) = 2 | Number and qualification not reported | Spirometry (for Respiratory pattern) and ultrasonography (for Diaphragmatic excursion and pelvic floor descent) were performed separately with the participant the supine lying position during the following test conditions: at rest, while performing an ASLR, and while performing an ASLR with manual pelvic compression through the ilia. A positive test is denoted by improved ability to raise the leg. Repeat measures of all variables were recorded in each of the 3 test conditions | None | SIJD and normal |
| Hungerford et al. [34] | Cross section | OCS | *SIJD*: *N* = 14, sex = 14 M, Age = 32.7 (range 24–47) Y, height = 176.8 (range 168–184)CM, weight = 77 (range 71–90)KG. *control*: *N* = 14, sex = 14 M, Age = 33.5 (range 22–50), height = 176 (range 168–182) CM, weight = 72.5 (range 61–85)KG | All subjects were assessed by the same experienced P.T to maintain continuity<br>Qualification not reported | EMG data were recorded from 7 trunk and hip muscles (adductor longus, biceps femoris, TFL, gluteus medius, gluteus maximus, LM and IO) on the side of standing on one leg during 5 left and right trials of hip flexion in standing. Force platform data were used to determine initiation of motion during hip flexion | None | SIJD and normal |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| Beales et al. [39] | Cross section | OCS | Chronic unilateral PGP: *N* = 12, sex = 12 F (*N* = 5 out of 12 nulliparous), Age = 39.8 (± 11.2), BMI = 23.2 (± 4.6) KG/M$^2$, Duration of symptoms = 92.6 (± 78.0) months, Qubec (X/100) = 22.9 (± 18.7), McGill (X/45) = 8.4 (± 2.7), VAS for usual pain (x/100) = 43.7 (± 24.3), Tampa (X/68) = 35.1 (± 9.2), continence dysfunction (*N*) = 7, Etiology (Pregnancy related) (*N*) = 4, Etiology (Trauma) (*N*) = 6, Etiology (Insidious) (*N*) = 2 | Qualification not reported | EMG of the anterior abdominal wall, Rt. CW and the scalene, IAP, ITP, RR, PF kinematics and downward leg pressure of the non-lifted leg were compared between an ASLR lifting the leg on the affected side versus the nonaffected side. First, data were collected for 60 s in resting supine. An ASLR trial was then performed for each leg. A cough at the start of each trial, producing movement of the PF on US, was used to synchronize PF data with the other variables. After coughing the leg was lifted for 45 s. A further trial was performed on each leg for repeatability analyses | | PGP |
| | *Validity study (MCI battery test)* | | | | | | |
| Biele et al. [45] | Case control | MCI Battery test | *NS-CLBP*: *N* = 65, sex = 64.6% F, Age = 46.4 (12.7) Y, BMI = 26.9 (5.7) kg/m$^2$, absolute ODI score (out of 50) 10.6 (7.4)<br>*No LBP*: *N* = 66, sex = 53.0% F, Age = 41.3 (13.3)Y, BMI = 25.5 (5.3) kg/m$^2$, absolute (ODI) score (out of 50) = 2.1 (3.47) | Nine raters were trained for 4 h on the execution and evaluation of the MCTs by a P.T who is a clinical expert in this field. Finally, raters were tested for agreement in a pilot study of 4 subjects<br>Qualification not reported | Eleven tests were performed. If a participant needed more than 3 attempts to perform the test correctly, the test was considered as failed. Each single test was rated as fail (−) or pass (+). The total number of failed tests was summed as a summary score. The subjects were tested in individual treatment rooms, in 1 session | None | The two-level (impairment or not) and three-level (none, mild, moderate/severe) categorization of participants with NS-CLBP |
| | *Validity study (Pain Behavior Assessment system (PBA))* | | | | | | |

**Table 3** (continued)

| Authors | Study design | Classification basis | Participants' characteristics | Examiners' number and qualification | Testing procedure (Methods) | Reference base | Classification subgrouping |
|---|---|---|---|---|---|---|---|
| Meyer et al. [46] | Cross section | PBA | NS-CLBP: *N* = 200, sex = 145 M(72.5%), Age = 43.3 (16.5), Duration of pain, months Median (IQR) due to nonparametric distribution = 34.4 (12–100), Fear of losing job *N* (percent) = (Yes = 47 (23.5)) (No = 56 (28)), No job contract = 97 (48.5), WAI = 21.3 (7.4), FABQaa (activity scale,) = 32.1 (9.6), FABQwa (work scale) = 19.2 (4.6) ODI = 43.3 (16.5) | *Experts*<br>Six experts with at least 5 years of work experience, developed the preliminary PBA. The experts were clinicians in the field of work rehabilitation and included a physician specialized in rehabilitation and insurance medicine. *Assessors*<br>P.T.s, who rated the PBA items during FCE testing. They had 2–15 y of experience in FCE testing. In addition, they had a 2-day course on FCE and trained during a 4-h session on application of PBA items | *Step 1 Development of the preliminary PBA*: The PBA was mainly developed by an expert group based on the assessment of inconsistency, as a part of the FCE<br>*Step 2 Pretesting*: Preliminary PBA testing was performed in rehabilitation centers. The feedback of the assessing physiotherapists was evaluated by experts and guided further adaptations. *Step 3 Construct validity*: Eight P.T.s. administered the FCE, including the revised PBA, to 200 participants | None | It consisted from the following 4 subscales, as observed in clinical practice: (a) pain perception; (b) overt pain behavior; (c) effort; and (d) consistency of behavior. A dichotomous scale (0 = no pain behavior present; 1 = pain behavior present) was applied to all items<br>The total score of the PBA is the sum of all items rated. The final results classify the participants to the following categories: no, low or high pain behavior |

AEP, Active Extension Pattern; ASLR, Active Straight Leg Raise; BMI, Body Mass Index; CW, Chest Wall; EMG, Electromyography; EO, External Oblique; F, Female; FABQa, Fear Avoidance Beliefs Questionnaire activity scale; FABQw, Fear Avoidance Beliefs Questionnaire work scale; FCE, Functional Capacity Evaluation; FP, Flexion Pattern; H, Healthy; ICLT, Iliocostalis Lumborum pars Thoracis; IO, obliquus internus abdominis; IAP, Intra-Abdominal Pressure; ITP, Intra-Thoracic Pressure; IQR, Interquartile Range; LT, Longissimus Thoracis; Lt., Left; McGill, McGill Pain Questionnaire; MCTs, Motor Control Tests; MVIC, Maximum Voluntary Isometric Contraction; M, Male; N, Number; NRS, Numeric Rating Scale; ODI, Oswestry Disability Index; P.T, physiotherapist; PEP, Passive Extension Pattern; PGP, Pelvic Girdle Pain; PF, Pelvic Floor; Quebec, Quebec Back Pain Disability Scale; RR, Respiratory Rate; RMDQ, Roland-Morris Disability Questionnaire; Rt., Right; sLM, superficial Lumbar Multifidus, Tampa, Tampa Scale for Kinesiophobia; TrIO, Transverse fibers of Internal Oblique; TrA, Transversus Abdominus; TFL, Tensor Fascia Lata; VAS, Visual Analogue Scale; WAI, Work Ability Index; y, year

For levels 1–4, the mean agreement (%) was excellent (96%; range 75–100%). For the fifth level, *K*- and mean agreement between 4 testers was strong 0.82 (range 0.66–0.90) and 86% (range 73–92%), respectively. The final classification level had a moderate mean *K* of 0.65 (range 0.57–0.74) and excellent agreement of 87% (range 85–92%) [26]. Within the OCS-MCI subcategory, the most reliably identified subgroup is the passive extension pattern (PEP) ($K=0.90$; strong), while the least reliable is the active extension pattern (AEP) ($K=0.66$; moderate) [26].

The OCS-MCI subcategory was also tested in 2 studies reported within a single article [28]. In the first study, the inter-rater agreement was excellent ($K=0.96$ and %-of-agreement = 97%). In the second study, the inter-rater agreement was moderate ($K=0.61$) on average and ranged from 0.47 to 0.80, while the mean agreement was 70%, ranging from 60 to 84%, among 13 examiners who assessed 25 cases including subjective information from participants and video recorded functional tests [28].

### MCI test battery

In one study, four examiners independently rated video recordings of 27 participants with NS-CLBP and 13 controls performing 10 MCI tests. *K* values for inter-rater reliability ranged between minimal and moderate (0.24–0.71). Six out of 10 tests showed substantial $K>0.6$ inter-rater-reliability. The most reliable tests (for both rater pairs) were the "pelvic tilt" for extension dysfunction, "one leg stance" for rotational dysfunction and "sitting knee extension" and "waiter's bow" tests for flexion dysfunction. The poorest reliability was reported for the "abduction in crook lying" test for rotational dysfunction where both rater pairs had low *K*-values ($K=0.44$; 95%, CI 0.18–0.70 and $K=0.32$; 95% CI 0.10–0.54). Intra-tester reliability ranged from 0.51 to 0.96. All tests, except abduction in crook lying, showed substantial reliability ($K>0.6$) [27].

The second study reported the inter-rater reliability between two examiners who independently examined 25 participants with NS-CLBP and 15 asymptomatic controls using the five MCI clinical tests. Intra-class correlation coefficients (ICCs) were excellent (0.90) for repositioning (RPS), 0.96 for sitting forward lean (SFL), 0.96 for sitting knee extension (SKE), 0.94 for bent knee fall out (BKFO) and 0.98 for leg lowering (LL) [29].

## Validity of different classification systems

All three diagnostic systems underwent some aspect of validation testing (Table 5):

1. *OCS* Fifteen studies assessed OCS validity (12 for OCS-MCI subcategories and 3 for sacroiliac joint dysfunction (SIJD)).

   *OCS-MCI construct validity* was reported by measuring *lumbosacral kinematics* and *trunk muscle activation* in 33 participants with NS-CLBP (20 Flexion Pattern (FP) and 13 AEP) and 34 asymptomatic controls. The biomechanical model used lower lumbar kinematics in sitting and forward bending and two trunk muscle activation variables (lack of flexion relaxation of the superficial lumbar multifidus in slump sitting and end range of forward bending). The model correctly classified 96.4% of cases and distinguished between individuals with No LBP, AEP and FP [38].

   *Discriminant validity of MCI subcategories through spinal kinematics testing*:

   *Sitting postures* were tested to distinguish participants with NS-CLBP with AEP (lordotic lumbar posture) and FP (kyphotic lumbar posture) from asymptomatic controls ($P<0.001$). Participants with NS-CLBP had less ability to consciously alter posture when asked to slump from usual sitting ($P<0.001$) [36]. Similar findings were reported during cycling as participants with NS-CLBP (FP) exhibited greater lumbar region flexion compared to asymptomatic controls ($p=0.018$) and reported remarkable pain increase over 2 hours of cycling ($p<0.001$) [41]. Further, cyclists with NS-CLBP (FP) showed increased, although non-significant, lumbar flexion and rotation tendency compared to controls ($P>0.05$) [35]. *In another study*, participants with NS-CLBP (FP) sat with less hip flexion ($P=0{:}05$), suggesting a relative posterior pelvic tilt. During "usual" sitting, the FP group positioned the lumbar spine significantly closer to end range lumbar flexion compared to asymptomatic controls [30].

   *Functional tasks*: spinal kinematics during functional tasks were not different between the AEP and asymptomatic controls in a single study [43]. However, the AEP group distinctively adopted more upper lumbar and lower thoracic (T6—L3) extension compared to the FP group which adopted more flexion during these activities ($p<0.05$). The FP group also exhibited greater thoracolumbar kyphosis than asymptomatic controls [43].

   *Spinal position sense (SPS)*: Lumbar repositioning accuracy was assessed in 3 studies [31, 33, 40]. Participants with NS-CLBP, compared to asymptomatic controls, developed substantially greater magnitude of Absolute Error (AE) [31, 40] and Variable Error (VE) [40]. The FP group underestimated lumbar target positions [31, 33, 40], while the AEP group overestimated lumbar and underestimated thoracic target positions compared to FP [40]. The Cardiff Dempster–Shafer Theory (DST) Classifier method, based on objective

**Table 4** Overview of the results of reliability studies

| Classification system (study reference) | Subgroup and test variables | Percentage | Kappa Mean (range) | MDC | ICC | (95% CI) | 95% LOA |
|---|---|---|---|---|---|---|---|
| Fersum et al. [26] | First level NS-CLBP Vs specific LBP | 98% | | | | | |
| | Second level NS-CLBP (peripheral pain source Vs central pain source) | 99% | | | | | |
| | Third level PGP | 100% | | | | | |
| | Third level LBP | 99% | | | | | |
| | Fourth level for PGP (increased VS decreased force closure) | 100% | | | | | |
| | Fourth level for LBP (MCI) | 99% | | | | | |
| | Fourth level for LBP (MI) | 75% | | | | | |
| | Fifth level for the primary direction of provocation | 86% (73–92%) | 0.82 (0.66–0.90) | | | | |
| | The final level for detecting psychosocial influence | 87% (85–92%) | 0.65 (0.57–0.74) | | | | |
| Dankarts et al. [28] Study 1 (between expert clinicians) | NS-CLBP (MCI) based on a comprehensive subjective and physical examination | 97% | 0.96 | | | | |
| Study 2 (between examiners and experts) | NS-CLBP (MCI) based on subjective and video | 70% (60–84%) | 0.61 (0.47–0.80) | | | | |
| | Flexion shifting pattern | 82% | | | | | |
| | Passive extension pattern | 77% | | | | | |
| | Flexion pattern | 68% | | | | | |
| | Multidirectional pattern | 68% | | | | | |
| | Active extension pattern | 62% | | | | | |
| Luomajoki et al. [27] | *Tests of movement control dysfunctions*: Sitting knee extension. (flexion dysfunction) | | Inter = 0.72\Intra = 0.95 | | | | |
| | Rocking extension | | Inter = 0.68\Intra = 0.51 | | | | |
| | Pelvic tilt. (extension dysfunction) | | Inter = 0.65\Intra = 0.80 | | | | |
| | One leg stance Lt. (rotation dysfunction) | | Inter = 0.65\Intra = 0.84 | | | | |
| | Waiters bow | | Inter = 0.62\Intra = 0.88 | | | | |
| | Prone knee bend rotation | | Inter = 0.58\Intra = 0.78 | | | | |
| | Rocking flexion | | Inter = 0.57\Intra = 0.72 | | | | |
| | Prone knee bend extension | | Inter = 0.47\Intra = 0.70 | | | | |
| | One leg stance Rt | | Inter = 0.43\Intra = 0.67 | | | | |
| | Crook lying | | Inter = 0.38\Intra = 0.86 | | | | |
| Enoch et al. [29] | *Five motor control tests*: repositioning (RPS) | | | 0.24 | 0.90 | (0.81; 0.94) | (−0.22; 0.26) |
| | Sitting forward lean (SFL) | | | 0.35 | 0.96 | (0.92; 0.98) | (−0.41; 0.29) |
| | Sitting knee extension (SKE) | | | 0.19 | 0.95 | (0.90; 0.97) | (−0.20; 0.18) |
| | Bent knee fall out (BKFO) | | | 0.37 | 0.94 | (0.88; 0.97) | (−0.40; 0.34) |
| | Leg lowering (LL) | | | 2.90 | 0.98 | (0.96; 0.99) | (−3.08; 2.74) |

measures of repositioning sense during sitting and standing, discriminated the No LBP from NS-CLBP (pooled and in subsets) with an accuracy ranging between 93.83 and 98.15%. Further, the DST classifier method distinguished different NS-CLBP subgroups with an accuracy of 96.8%, 87.7% and 70.27% for FP from PEP, FP from

AEP and AEP from PEP subtypes, respectively. Finally, ranking analysis showed that lumbar AE in sitting could distinguish participants with NS-CLBP from No LBP and FP from No LBP, while lumbar constant error in standing consistently discriminated LBP extension subsets (AEP and PEP) from No LBP [44].

*Discriminant validity of MCI subcategories through trunk muscle activity testing*:

Surface electromyography (sEMG) recorded from five trunk muscles during unsupported "usual" and "slumped" sitting postures could not distinguish trunk muscle activity between asymptomatic controls and a pooled NS-CLBP group. However, compared to controls, participants classified with AEP presented with significantly higher co-contraction of lumbar multifidus, ilio-costalis lumborum pars thoracis and transverse fibers of internal oblique muscles ($p < 0.05$) [37]. Burnett et al. [35] reported less co-contraction of the lower lumbar multifidus in the FP group compared to controls during a cycling task [35]. Sheeran et al. [40] reported the NS-CLBP (FP and AEP combined) group produced significantly higher abdominal activity ($p < 0.01$) compared to controls during usual sitting and standing postures. Hemming et al. [42] also reported significantly greater muscle activation in right-sided superficial lumbar multifidus muscles during the functional tasks of step up, reach up and box replace ($p < 0.05$). External oblique muscle contraction during box lift differed significantly between participants with AEP and asymptomatic controls ($p = 0.016$). Significant differences between participants with FP and asymptomatic controls were also reported for left-sided transversus abdominis/internal oblique and superficial lumbar multifidus activity during stand-to-sit tasks ($p = 0.009$) [42].

*SIJD subcategory*: Two studies reported decreased diaphragmatic excursion, altered respiratory patterns and depression of the pelvic floor (PF) in participants with NS-CLBP during the ASLR test compared to controls [32, 39]. When the examiner added manual pelvic compression to the ASLR test, there were no differences between the two groups. Manual pelvic compression during the ASLR theoretically improves load transfer by enhancing passive stability of the SIJs and MC patterns/force closure [32]. Another study reported delayed activation of obliquus internus abdominis (OI), multifidus and gluteus maximus muscles in patients with SIJD compared to controls. Delayed OI and multifidus activation occurred in both the symptomatic and the asymptomatic sides in the SIJD group. Biceps femoris activation occurred earlier in SIJP group [34].

2. *MCI Test Battery*: One study assessed the clinical validity of the MCI test battery for classifying participants with NS-CLBP [45]. For both the two-class (impairment or not) and the three-class (none, mild/moderate and severe) categorization, the ideal number of MCI tests was 10. The overall discrimination potential for two-class categorization was good (Area Under the Curve (AUC) > 0.8, sensitivity = 0.75, specificity = 0.82, Youden = 0.57, LR + = 3.40, LR- = 0.20, effect size = 1.45), with an optimal cutoff of three tests. To classify MCI, at least four failed items are needed. The overall discrimination potential for the three-class categorization was fair (volume under the surface > 0.5, sensitivity = 0.48, sensitivity = 0.50, specificity = 0.82, Youden = 0.40, effect size = 1.56), with an optimal cutoff of three and six tests. At least four failed MCI tests are needed to classify mild/moderate MCI and six or more failed tests classify severe cases [45].
3. *PBA*: The internal consistency (reliability) of PBA showed good person separation index (0.83). Construct validity evaluated by Rasch analysis resulted in 41 items. PBA convergent validity was supported by a significant correlation with other questionnaires [46].

### Risk of bias assessment

Risk of bias assessment showed an excellent inter-assessor agreement (92.2% and $K = 0.84$) [47]. Nine studies [30–32, 34, 39–41, 45] did not clarify evaluators' characteristics (*Item 2*). Reference standard tests (*Item 3*) were reported only in two studies [38, 44] and were performed independently (*Item 9*). All inter-rater reliability studies used raters blinded to each other's findings [26–29] (*Item 4*). All studies reported clear descriptions of measurement procedures (*Item 10*). Appropriate reliability and validity statistical methods (*item 13*) were employed in 21 studies, while one study employed the Kolmogorov–Smirnov test, a less than ideal test used to confirm normality distribution in a small sample study ($n < 50$) [41]. Overall, all five reliability studies [26–29] and two OCS validity studies [38, 44] were rated as low risk of bias. The remaining validity studies were rated as high risk [30–37, 39–43, 45, 46] (Table 6).

## Discussion

This systematic review identified and critically appraised studies reporting reliability and validity of functionally oriented NS-CLBP diagnostic classification systems; specifically, the OCS, MCI test battery, and PBA systems. Of the 3 systems evaluated through studies included in this review, the OCS is the most reliable and valid. All included reliability studies were consistently rated as high quality. However, validity in this context is limited to the capacity to

**Table 5** Overview of results of validity studies

| Reference | Statistical tests | Validity | Summary of results |
|---|---|---|---|
| Dankarts et al. [38] (Construct validity) | Stepwise linear discriminant analysis (LDA) to generate statistical model<br>Model's validity (cross-validation procedure and a holdout validation)<br>Univariate analysis | *Model, Cross-Validation, and Hold-Out Group Validation Accuracy (%) (range)*:<br>Statistical classification model accuracy = 96.4%; valid cases (*n*) = (no LBP = 29, AEP = 10, FP = 14)<br>Model validation<br>Cross-validation accuracy = 94.6%; valid cases (*n*) = (No LBP = 29, AEP = 10, FP = 14)<br>Hold-out validation accuracy = 74.5% (53.9–93.8); valid cases(*n*) = 33% of all cases<br>*Differences Between Subgroups Based on Univariate Analysis*:<br>Forward bending LLx Q4 (°) $F = 12.2$; $p < 0.001$†; post hoc AEP < FP = no LBP<br>Forward bending SA Q2 (°) $F = 9.5$; $p < 0.001$†; post hoc FP > AEP = no LBP<br>Return forward bending SA Q2 (°) $F = 4.3$; $p < 0.018$†; post hoc No LBP = AEP, No LBP < FP, AEP = FP<br>Usual sitting Lx (°) $F = 29.2$; $p < 0.001$†; post hoc AEP < no LBP < FP<br>Usual sitting SA (°) $F = 20.4$; $p < 0.001$†; post hoc AEP = no LBP < FP<br>Forward bending fully bent sLM F = 36.3; $p < 0.001$†; post hoc AEP > FP > no LBP<br>Sitting FRR LM $F = 3.7$; $p < 0.031$†; post hoc No LBP > AEP = FP | The statistical model used 5 kinematic and 2 EMG variables. The model correctly classified 96.4% of cases |
| Dankarts et al. [36] | *T test* used to compare No LBP and pooled NS-CLBP<br>*One-way ANOVA* used to compare No LBP VS subclassified NS-CLBP | *No LBP Versus Pooled NS-CLBP*<br>*Usual sitting*<br>Sacral tilt ($t = -1.95$, $P = 0.06$)<br>Lower lumbar angle ($t = —0.31$, $P = 0.75$)<br>Upper lumbar angle ($t = —0.02$, $P = 0.99$)<br>*Slumped sitting*<br>Lower lumbar angle ($t = 2.78$; $P < 0.007$) in participants with NS-CLBP sitting with more lordosis<br>Participants with NS-CLBP showed less ability to change their lumbo-pelvic posture when moving from *usual to slumped sitting*<br>Sacral tilt ($t = 4.82$; $P < 0.001$)<br>Lower lumbar angle ($t = 4.29$; $P < 0.001$)<br>*No LBP Versus Subclassified NS-CLBP*<br>*Usual sitting*<br>Sacral tilt (F2,63 = 20.44; $P < 0.001$) for AEP subgroup and No LBP subjects, when compared with FP subgroup (who showed a kyphotic posture)<br>Lower lumbar angle (F2,63 = 19.76; $P < 0.001$) for AEP compared with No LBP and FP<br>Upper lumbar angle (F2,63 = 9.86; $P < 0.001$). AEP sat with more lumbar lordosis than No LBP, and the FP sat with a more kyphotic lumbar spine<br>*Slumped sitting*<br>AEP had more anterior sacral tilt (F2,63 = 9.05; $P < 0.001$)<br>Lower lumbar angle larger in AEP than No LBP subjects and FP groups (F2,63 = 16.31; $P < 0.001$)<br>Upper lumbar angle (F2,63 = 1.03; $P = 0.36$) No differences were noted between the three groups<br>Less change between *usual and slouched sitting* for AEP and FP compared with No LBP<br>Sacral tilt (F2,63 = 12.97; $P < 0.001$)<br>Lower lumbar angle (F2,63 = 9.42; $P < 0.001$)<br>For the upper lumbar angle (F2,63 = 12.52; $P < 0.001$), the AEP and No LBP showed greater change compared to FP | Differences in usual sitting posture were only revealed when participants with NS-CLBP were subclassified (AEP and FP based on OCS) |
| Van hoof et al. [41] | Independent *t* test<br>ANOVA (with post-hoc Bonferroni) | Participants with NS-CLBP (FP) were significantly more flexed at the lower lumbar spine during cycling compared to asymptomatic controls ($p = 0.018$) and reported a significant increase in pain over the 2 h of cycling ($p < 0.001$)<br>One-way repeated measures ANOVA revealed a significant main effect for group ($p = 0.035$, $F = 5.546$) which remained just significant when adding saddle angle as a covariate ($p = 0.05$, $F = 4.747$) | FP subgroup demonstrates maladaptive motor control pattern (greater lower lumbar flexion) during cycling, likely related to a significant increase in pain |

**Table 5** (continued)

| Reference | Statistical tests | Validity | Summary of results |
|---|---|---|---|
| Burnett et al. [35] | Independent *t* tests | Increased spinal flexion in the lower thoracic region at the start and finish of the ride ($d=0.96$ and 0.80, respectively) and increased range of axial rotation in the lower lumbar spine for the pain group at the start of the ride ($d=0.89$) were evident<br>The pain group exhibited greater levels of activation of the Rt. erector spinae (T9) ($d=0.83$), Lt. LM (1.24), Rt. rectus abdominus ($d=0.81$) and reduced levels of activation of the Lt. IO ($d=0.81$) at the end of the ride<br>Furthermore, asymmetrical activation of the lower portion of (LM) were observed in the pain group both at the beginning ($d=0.81$) and end ($d=0.99$) of the ride | The findings suggest altered postural motor control and kinematics of the lower lumbar spine are associated with the development of LBP in cyclists |
| O sullivan et al. [30] | Independent *t* tests<br>Spearman's correlations | FP subjects had significantly reduced back muscle endurance ($P<0:01$)<br>FP sat with less hip flexion ($P=0:05$), suggesting increased posterior pelvic tilt in sitting<br>FP subjects postured their spines significantly closer to their end of range lumbar flexion in "usual" sitting than the asymptomatic controls ($P<0.05$) | There is a relationship between reduced lumbar muscle endurance, presumably from habitually posturing the lumbar spine close to end range flexion in sitting |
| Hemming et al. [43] | One-way ANOVAs with post hoc Bonferroni<br>ANOVAs were repeated for the kinematic data with gender as a covariate | *Kinematics (tasks)*<br>*Reach up*:<br>Upper lumbar: $F=4.824$; $p=0.011$; post hoc AEP vs. FP (0.015)<br>*Step down*:<br>Total lumbar: $F=3.248$; $p=0.044$; post hoc AEP vs. FP (0.039)<br>Lower thoracic: $F=4.353$; $p=0.016$; post hoc AEP vs. FP (0.016)<br>Upper lumbar: $F=6.902$; $p=0.002$; post hoc AEP vs. FP (0.002)<br>*Step up*:<br>Lower thoracic: $F=3.967$; $p=0.023$; post hoc AEP vs. FP (0.030)<br>Upper lumbar: $F=7.432$; $p=0.001$; post hoc AEP vs. FP (0.001), FP vs. control (0.025)<br>*Box replace*:<br>Lower thoracic: $F=5.231$; $p=0.007$; post hoc AEP vs. FP (0.008)<br>Upper lumbar: $F=7.844$; $p=0.001$; post hoc AEP vs. FP (0.001), FP vs. control (0.036)<br>*Box lift*:<br>Lower thoracic: $F=5.144$; $p=0.008$; post hoc AEP vs. FP (0.008)<br>Upper lumbar: $F=6.849$; $p=0.002$; post hoc AEP vs. FP (0.001)<br>*Stand to sit*:<br>Total lumbar: $F=4.574$; $p=0.013$; post hoc AEP vs. FP (0.010)<br>Lower thoracic: $F=5.997$; $p=0.004$; post hoc AEP vs. FP (0.006), FP vs. control (0.025)<br>Upper lumbar: $F=10.530$; $p<0.001$; post hoc AEP vs. FP ($<0.001$)<br>*Sit to stand*:<br>Total lumbar: $F=3.334$; $p=0.041$; post hoc AEP vs. FP (0.038)<br>Lower thoracic: $F=6.638$; $p=0.002$; post hoc AEP vs. FP (0.004), FP vs. control (0.018)<br>Upper lumbar: $F=9.050$; $p<0.001$; post hoc AEP vs. FP ($<0.001$)<br>*Pick up pen (bend down)*:<br>Lower thoracic: $F=5.027$; $p=0.009$; post hoc AEP vs. FP (0.033), FP vs. control (0.018)<br>Upper lumbar: $F=5.830$; $p=0.005$; post hoc AEP vs. FP (0.005), FP vs. control (0.044)<br>*Pick up pen (return)*:<br>Lower thoracic: $F=5.478$; $p=0.006$; post hoc AEP vs. FP (0.019), FP vs. control (0.016)<br>Upper lumbar: $F=4.978$; $p=0.009$; post hoc AEP vs. FP (0.007) | Regional spinal curvatures appear to differ in MCI subgroups, during functional tasks. The FP group demonstrated more kyphotic thoraco-lumbar kyphotic postures. No significant differences between AEP and asymptomatic groups were observed, suggesting that these groups adopt similar functional movement strategies |

**Table 5** (continued)

| Reference | Statistical tests | Validity | Summary of results |
|---|---|---|---|
| O Sullivan et al. [33] | Independent *t* tests | For the lumbar spine, the average Repositioning Error (RE) was 1.7 ± 0.8 cm for the LBP (FP) group and 1.1 ± 0.6 for the comparison group<br>Lumbar RE was significantly greater in the LBP group than in the comparison group ($t$ [28] = 2.48; $P = 0.02$)<br>There was also a significant difference between the groups at each individual sensor. RE in both groups was greatest at T12 and decreased continuously to S2 | Individuals with a clinical diagnosis of lumbar segmental instability demonstrate an inability to reposition the lumbar spine accurately into a neutral spinal posture while seated |
| Sheeran et al. [40] | Independent *t* tests<br>One-way ANOVA with post hoc Bonferroni | *Spinal Position Sense*:<br>The NS-CLBP (*combined and subclassified*) group produced significantly greater absolute error (AE) and variable error (VE) than the asymptomatic controls in the sitting and standing thoracic and lumbar spine regions<br>No difference was observed between NS-CLBP (*combined*) and asymptomatic controls in thoracic and lumbar constant error (CE) during sitting and standing<br>Only when the NSCLBP was *subclassified* were differences in CE apparent in the thoracic spine (sitting: $P = 0.001$) and lumbar spine (sitting: $P = 0.003$; standing: $P = 0.041$)<br>In sitting, the FP underestimated the lumbar target and overestimated the thoracic target compared with AEP and asymptomatic groups ($P < 0.01$). Conversely, AEP overestimated the lumbar target and underestimated the thoracic target compared with FP ($P < 0.016$), but not reaching statistical significance compared with the asymptomatic group ($P > 0.016$)<br>In standing, the only significant difference was in the lumbar spine where AEP overestimated the target position compared with the asymptomatic group ($P < 0.016$). There was no difference between subgroups in the thoracic spine during standing ($P > 0.016$)<br>*Trunk Muscle Activity*<br>The NS-CLBP (*combined*) produced significantly higher TrIO and EO activity and comparable LM and ICLT activity during sitting and standing compared to asymptomatic control<br>*After subclassification*, differences were apparent in LM during standing ($P < 0.017$), where FP produced higher activity than the asymptomatic controls ($P < 0.016$), although a statistical significance was not reached between FP and AEP subgroups ($P > 0.016$)<br>No subgroup differences were shown in TrIO and EO (sitting and standing), ICLT (sitting and standing) and LM (sitting) | Subgroups of participants with NS-CLBP had similar neutral spinal position deficits in error magnitude and variability, but subclassification revealed clear differences in the direction of the deficit. Trunk muscle activation was shown to be largely nondiscriminatory between subgroups, except for sLM |
| O Sullivan et al. [31] | Independent *t*-tests<br>A Mann–Whitney U-test<br>Pearson's correlation coefficient (r) and Spearman's correlation coefficient (rs) | Comparing RE between groups:<br>AE ($t = 3.4$, $p = 0.002$) and CE ($t = -2.957$, $p = 0.006$) were both significantly greater in the NSCLBP group<br>VE ($Z = -1.39$, $p = 0.165$) was not significantly different between the groups<br>Correlations with RE:<br>A moderate, statistically significant, positive correlation was found between functional disability and AE ($r = 0.601$, $p = 0.018$)<br>Similarly, a moderate, statistically significant, inverse correlation was found between fear avoidance and CE ($r = -0.577$, $p = 0.002$) | Increased lumbo-pelvic RE in a NS-CLBP group (undershooting the target position). Overall, RE was only weakly to moderately correlated with measures of pain, disability and fear |

**Table 5** (continued)

| Reference | Statistical tests | Validity | | Summary of results |
|---|---|---|---|---|
| Sheeran et al. [44] | Chi square (group difference)<br>ANOVA for BMI<br>The DST is a mathematical method | In discriminating NS-CLBP from no LBP, the classifier accuracy was 96.61%<br>From no LBP, subsets of FP, AEP and PEP achieved 93.83, 98.15% and 97.62% accuracy, respectively<br>Classification accuracies of 96.8%, 87.7% and 70.27% were found when discriminating FP from PEP, FP from AEP and AEP from PEP subsets, respectively<br>Sitting lumbar error magnitude best discriminated NS-CLBP from No LBP (92.4% accuracy) and the FP subset from no LBP (90.1% accuracy)<br>Standing lumbar error best discriminated AEP and PEP from No LBP (94.4% and 95.2% accuracy, respectively) | | Using repositioning accuracy, the Cardiff DST Classifier distinguishes between subsets of LBP |
| Dankarts et al. [37] | Independent *t* tests to compare between the No LBP and pooled NS-CLBP<br>one-way ANCOVA with post hoc comparisons (Bonferroni)<br>Paired *t* tests | *No LBP Versus Pooled NS-CLBP*<br>Usual sitting: No differences between No LBP and NS-CLBP<br>Slumped sLM $t=-3.39$; $p=0.001$<br>Slumped ICLT $t=-2.82$; $p=0.006$<br>FRR in sitting<br>sLM $t=4.6$; $P<0.001$<br>ICLT $t=2.7$; $P<0.001$ | *No LBP Versus Subclassified NS-CLBP*<br>Usual sitting sLM $F=5.5$; $p=0.006$* post hoc (FP < NO < AEP)<br>Usual sitting ICLT $F=8.9$; $P<0.001$* post hoc (FP < NO < AEP)<br>Slumped sLM $F=6.6$; $P=0.003$* post hoc (NO < FP < AEP)<br>Slumped ICLT $F=9.9$; $P<0.001$* post hoc (NO < FP < AEP)<br>Usual TrIO $F=3.2$; $P=0.04$* post hoc (FP < NO < AEP)<br>Slumped TrIO $F=3.4$; $P=0.04$* post hoc (FP < NO < AEP)<br>Difference Between Usual and Slumped Sitting in sLM Activity<br>FP and AEP showed a nonsignificant difference in sEMG of the sLM when moved from usual sitting-to-slumped sitting (+4%; $t=-1.6$; $P=0.11$) (at the level of their LBP) and (37% vs. 36%; $t=0.42$; $P=0.685$)respectively<br>The No LBP group showed a clear difference (23% vs. 14%; $t=4.4$; $P<0.001$) suggesting a relaxation response | No differences were found during usual sitting when the participants with NS-CLBP were pooled. Analysis based on subgrouping the participants revealed significant differences in muscle activation patterns |
| Hemming et al. [42] | Mann–Whitney *U* tests were conducted to establish pairwise differences between groups | *Kruskal–Wallis ($p<0.05$) and post hoc Mann–Whitney ($p<0.0167$) of sEMG of muscles during functional tasks*:<br>*Step up*: Rt. sLM: $p=0.046$; post hoc AEP vs. control (0.015)<br>*Reach up*: Rt. sLM: $p=0.039$; post hoc AEP vs. control (0.013)<br>*Stand to sit*: Lt. TrA/IO: $p=0.02$; post hoc FP vs. control (0.009), Lt. sLM: $p=0.02$; post hoc FP vs. control (0.009)<br>*Box replace*: Rt. sLM: $p=0.026$; post hoc AEP vs. control (0.007)<br>*Box lift*: Rt. EO: $p=0.019$; post hoc AEP vs. control (0.008) | | sEMG findings demonstrated participants with NS-CLBP exhibited increased muscle activity compared to controls |

**Table 5** (continued)

| Reference | Statistical tests | Validity | Summary of results |
|---|---|---|---|
| O'Sullivan et al. [32] | Two groups (SIJD group and control group) for three conditions (resting supine position, ASLR and ASLR with compression) (ANOVA) | *Respiratory function*<br>Minute ventilation (the total volume of gas entering the lung per minute; it is equal to the tidal volume multiplied by the respiratory rate.)<br>Between the *SIJD* and *control* groups was significantly different ($F$ (1,24) = 5.49; $P$ = 0.028)<br>Between the three *testing conditions* was significantly different in *the SIJD* ($F$ (1.28,30.63) = 6.43; $P$ = 0.011)<br>In *SIJD* between the resting supine and ASLR conditions ($F$ (1,24) = 5.17; $P$ = 0.032)<br>In *SIJD* between ASLR and ASLR with compression also was identified ($F$ (1.24) = 4.42; $P$ = 0.046)<br>Respiratory rate<br>Between the two groups ($F$ (1,24) = 10.42; $P$ = 0.004) (increased the rate in SIJD)<br>Between the three testing conditions ($F$ (1.25,29.95) = 5.85; $P$ = 0.016) (increased during ASLR)<br>*Diaphragmatic excursion*<br>Only a significant difference in the magnitude of diaphragmatic excursion did exist between the *three conditions* ($F$ (2,48) = 22.25; $P$ = 0.001) (in *SIJD*)<br>Between the *resting supine* and *ASLR* ($F$ (1,24) = 60.93; $P$ = 0.001) (decreased diaphragmatic excursion during ASLR in the participants with SIJD<br>Between *ASLR* and *ASLR with compression* it also was significant ($F$ (1,24) = 34.85; $P$ = 0.001) (diaphragmatic excursion increased, returning to a level comparable with that of the comparison group<br>Between the *resting supine* and *ASLR with compression* ($F$ (1,24) = 9.62; $P$ = 0.005) (demonstrating that it did not return to the resting level)<br>*Pelvic floor descent*<br>The distinguishing feature of this interaction was the magnitude of PF descent during ASLR in the SIJP group:<br>Between the two groups ($F$ (1,24) = 22.95; $P$ = 0.001)<br>Between the two conditions (ASLR and ASLR with compression) ($F$ (1,24) = 26.82; $P$ = 0.001) | The key feature of this interaction was an increase in minute ventilation in the group with SIJD during ASLR without manual pelvic compression<br>In the participants with SIJP, it was observed that minute ventilation decreased to a level like that in the comparison group during performing ASLR with pelvic compression |
| Hungerford et al. [34] | Two tailed paired Student's $t$ tests were performed for all variables between the Lt. and Rt. sides in the control group and between the symptomatic side and the asymptomatic side in the SIJD group | *Comparison of EMG onsets between the SIJD group and control (the symptomatic side)*:<br>OI, multifidus and gluteus maximus were significantly delayed on the *symptomatic* side (all $P \leq 0.01$)<br>The onset of biceps femoris occurred significantly earlier ($P < 0.03$) on the *symptomatic* side<br>*Comparison of EMG onsets between the SIJD group and control subjects (the asymptomatic side)*:<br>For OI and multifidus activation showed a significant difference:<br>OI was ($P \leq 0.01$) and multifidus was ($P = 0.05$) (delayed in the *SIJD* group in comparison with control)<br>*Onset of EMG activity in the SIJD group (the symptomatic side VS. the asymptomatic side)*:<br>For OI, multifidus and gluteus maximus, there was a significant difference:<br>Onset of OI and multifidus activation on the symptomatic side was significantly delayed ($P \leq 0.01$), as gluteus maximus was ($P \leq 0.05$) | The onset of OI and multifidus contraction occurred before initiation of weight transfer in the control subjects. The onset of OI, multifidus and gluteus maximus contraction was delayed on the symptomatic side in SIJD compared with control subjects, and the onset of biceps femoris EMG activity was earlier. In addition, EMG onsets were different between the symptomatic and asymptomatic sides in SIJD subjects |

**Table 5** (continued)

| Reference | Statistical tests | Validity | Summary of results |
|---|---|---|---|
| Beales et al. [39] | Repeated measure analysis of variance and post hoc Student *t* tests (patterning and bracing)<br>Paired Student *t* tests | *Obliquus Internus Abdominis (Patterning)*:<br>The IO on the affected side showed greater activation lifting the leg on the affected side ($P=0.0254$)<br>*Obliquus Internus Abdominis (Bracing)*:<br>During an ASLR on the affected side, there was symmetrical activation of the IOs ($P=0.235$) consistent with a bracing pattern, but asymmetrical tonic activation during a nonaffected side ASLR ($P=0.034$)<br>*Obliquus Externus Abdominis (Patterning)*:<br>There was no difference in EO activation lifting either the leg on the affected or nonaffected side (affected side EO: $P=0.150$; nonaffected side EO: $P=0.456$)<br>*Obliquus Externus Abdominis (Bracing)*:<br>Activation of EO was symmetrical during ASLR on the affected side (muscle: $P=0.087$) but asymmetrical during ASLR on the nonaffected side (muscle: $P=0.002$)<br>*Rectus Abdominis Patterning and Bracing*:<br>Side by respiration was significant for the affected RA (affected RA: side by respiration, $P=0.033$)<br>*Intra-Abdominal Pressure and Intrathoracic Pressure*:<br>Respiratory fluctuation of IAP and ITP did not vary lifting either leg (IAP: $P=0.185$, ITP$=0.571$). The baseline shift in IAP was greater during an ASLR on the affected side ($P=0.044$) but did not change for ITP ($P=0.892$)<br>*PF Movement*:<br>There was greater PF downward movement in response to an ASLR on the affected side ($P=0.012$)<br>*Contralateral Leg Downward Pressure*:<br>Downward leg pressure with the non-lifted leg did not differ during either ASLR ($P=0.326$) | Performing an ASLR on the symptomatic side resulted in a pattern of bracing the abdominal wall. This was associated with increased intraabdominal pressure and depression of the pelvic floor when compared with an ASLR on the nonaffected side |
| Biele et al. [45] | The receiver operating characteristic (ROC) statistics including the optimal cutoff by means of the Youden index, AUC, sensitivity and specificity | *The diagnostic criteria for the two-level classification system (LBP/no LBP) was*:<br>Items ($n$) = 10, AUC = 0.85, sensitivity = 0.75, specificity = 0.82, Youden = 0.57, LR + = 3.40, LR- = 0.20, effect size = 1.45, cutoff = (> 3)<br>*The diagnostic criteria for the three-level classification system (no MC, mild/moderate MC and severe MC) was*:<br>Items ($n$) = 10, VUS = 0.52, sensitivity (chronic status) = 0.48, sensitivity (acute status) = 0.50, specificity = 0.82, Youden = 0.40, effect size = 1.56, cutoff (≤, ≥) = 3, 6 | The ideal number of test items was 10<br>The overall discrimination potential for the two-level categorization system of the test is good (AUC = 0.85) with an optimal cutoff of three failed tests<br>The overall discrimination potential for the three-level categorization system of the test is fair (VUS > 0.8). The optimal cutoff of the 10-item solution is 3 and 6, respectively |
| Meyer et al. [46] | Rasch analysis (construct validity)<br>The person separation index (PSI) (Internal consistency)<br>Spearman's correlation coefficients (Convergent Validity) | *Construct validity of PBA*:<br>Rasch analysis removed 11 items due to misfit and redundancy, resulting in a final PBA of 41 items. Item mean fit residual was −0.33 (SD 1.06) and total item Chi square 100.39 (df = 82, $p=0.08$)<br>*Internal consistency of PBA*:<br>The PSI value was 0.83. This value should be ≥ 0.7 for group comparisons and ≥ 0.85 for individual use<br>Differential item functioning (DIF) analysis for age and gender revealed no bias<br>*Convergent validity of PBA*:<br>The Spearman's correlation coefficients rho between the PBA and the other questionnaires were all significant with values of − 0.59 for the WAI, 0.39 for the FABQa, 0.37 for the FABQw and 0.70 for the ODI | The PBA is a valid method to assess pain behavior in participants with NS-CLBP |

AEP, Active Extension pattern; ASLR, Active Straight Leg Raising test; Cardiff DST classifier, Cardiff Dempster–Shafer Theory Classifier; EO, external oblique; FABQa, Fear Avoidance Beliefs Questionnaire activity scale; FABQw, Fear Avoidance Beliefs Questionnaire work scale; FP, Flexion Pattern; FRR, Flexion Relaxation Ratio; ICLT, iliocostalis lumborum pars thoracic; Lt., Left; LDA, Linear Discriminant Analysis; LM, Lumbar Multifidus; No LBP, no low back pain; NS-CLBP, nonspecific chronic low back pain; ODI, Oswestry Disability Index; PEP, Passive Extension Pattern; PF, Pelvic Floor; RE, Repositioning Error; RA, rectus abdominis; Rt., Right; SA, Sacral Angle; sLM, superficial lumbar multifidus; TrIO, transverse fibers internal oblique; WAI, Work Ability Index; N.B. The alpha level for statistical significance was set at $p<0.05$

systematically identify different muscle activation patterns and spinal kinematic changes from each other and controls (construct and discriminant validity) as demonstrated in two high-quality studies. The remaining reviewed studies that assessed some aspects of OCS validation had high risk of bias. Limited evidence supports acceptable inter- and intra-rater reliability for clinical use of the following MCI test battery tests: "sitting knee extension" (to identify flexion dysfunction), “one leg stance” (for rotational dysfunction) and “pelvic tilt” (for extension dysfunction). Evidence supporting validity of the PBA is inconclusive.

### The OCS system

The OCS is currently the most studied, functionally based classification system with inter-rater reliability among various stages ranging from moderate to excellent [26]. For the OCS-MCI subcategory, three low risk of bias studies reported strong reliability (FP, AEP, PEP, flexion/lateral shift pattern and multidirectional pattern) [26, 28], with PEP as the most reliable subgroup, and AEP as the least reliable [26]. This review identified two low risk of bias studies demonstrating construct [38] and discriminant validity [44] of OCS-MCI subcategories based on determining and explaining aberrant muscle activity and spinal kinematic changes in participants with NS-CLBP. These studies generally adhered to guidelines for developing and validating classification systems [12, 48–56].

### MCI test battery

Based on 2 low risk of bias reliability studies [27, 29], 3 individual tests included in the MCI test battery show good to excellent reliability to classify people with NS-CLBP with or without MCI. Two previous systematic reviews concluded similarly [57, 58]. Current evidence suggests clinical use of the "sitting knee extension" test to identify flexion dysfunction, the “one leg stance” test for rotational dysfunction and the “pelvic tilt” test for extension dysfunction are suitable for clinical use based on good–excellent values both for intra- and inter-rater reliability [27, 29]. Because validity of the 10-test battery is based on a single study [45] with a high risk of bias, recommending routine clinical use is premature.

### PBA

The PBA consistently recognizes and classifies pain behavior into three categories (none, low, and high). However, evidence is limited to a single study with a high risk of bias [46]. People with no or low levels of pain behavior are likely to benefit from a physically oriented rehabilitation program with little emphasis on psychological and behavioral approaches. Conversely, those with high pain behavior may benefit from programs that emphasize psychological and behavioral factors. This reasoning incorporates a biopsychosocial approach similar to stratified care informed by the 9-item STarT Back questionnaire [59]. Unlike the STarT back questionnaire, the PBA also includes observing movement tasks and behaviors.

### Risk of bias assessment of individual studies

All reliability studies for OCS and MCI test battery were rated as high quality. The main risk of bias in reliability studies was the lack of randomizing test order.

Only two OCS validation studies [38, 44] were rated as high-quality largely because they employed expert opinion as a reference standard. However, no currently available objective tests classify function [54]. When no such tests are available, expert opinion, though limited, represents the best available reference standard [60–62]. The main risk of bias for the MCI test battery and PBA was the absence of a reference standard. The choice of statistical methods was considered appropriate for all studies; although one study used [41] the Kolmogorov–Smirnov test, which is not recommended for testing normality [63, 64].

### Implications for clinical practice

The findings of this review suggest that clinicians can use OCS to reliably classify functional characteristics of patients with NS-CLBP. Upper lumbar and lower thoracic spine kinematic studies offer mechanistic evidence supporting the rationale for assessing MCI. However, evidence supporting the validity of the 10-item MCI test battery is inconclusive because it is available from only 1 high risk of bias study. Because the effectiveness of therapies informed by functional classification is generally unknown, it is unclear if such diagnosis can be used to both inform effective care and/or as an objective measure of condition severity or response to care.

### Implication for future research

Standardized assessment protocols for determining MCI require well-defined procedures, operational definitions and quantifiable values. Standardizing these will facilitate more clinically useful findings and the ability to pool data from clinical trials [29]. Included classification systems used sagittal plane MCI assessment. Future studies should consider frontal and transverse planes to more comprehensively assess complex movement strategies. Further studies with lower risk of bias are needed to confirm the clinical usefulness of PBA classification. Finally, RCTs validating the clinical effectiveness of treatments based on functional assessment are needed.

**Table 6** Quality assessment of the included studies with the Clinical Appraisal Tool (CAT)

| Study | Item 1[a] V[n]+R[o] | Item 2[b] V+R | Item 3[c] V | Item 4[d] R | Item 5[e] R | Item 6[f] R | Item 7[g] V | Item 8[h] R | Item 9[i] V | Item 10[j] V+R | Item 11[k] V | Item 12[l] V+R | Item13[m] V+R | Risk of bias (% "Yes") * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Reliability studies* | | | | | | | | | | | | | | |
| Fersum et al. [26] | Yes | Yes | Yes | Yes | N/A | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Low (88%) |
| Dankart et al. [28] study 1 | Yes | Yes | N/A | Yes | N/A | Yes | N/A | Yes | N/A | Yes | N/A | Yes | Yes | Low (100%) |
| Dankart et al. [28] study 2 | Yes | Yes | Yes | Yes | N/A | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Low (100%) |
| Luomajoki et al. [27] | No | Yes | N/A | Yes | Yes | No | N/A | Yes | N/A | Yes | N/A | Yes | Yes | Low (78%) |
| Enoch et al. [29] | Yes | Yes | N/A | Yes | N/A | Yes | N/A | Yes | N/A | Yes | N/A | Yes | Yes | Low (100%) |
| Validity studies | | | | | | | | | | | | | | |
| Dankarts et al. [38] | Yes | Yes | Yes | N/A | N/A | N/A | Yes | N/A | Yes | Yes | Yes | Yes | Yes | Low (100%) |
| Dankarts et al. [36] | Yes | Yes | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (56%) |
| Van Hoof et al. [41] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | No | High (33%) |
| Burnett et al. [35] | Yes | Yes | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (56%) |
| O'Sullivan et al. [30] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (44%) |
| Hemming et al. [43] | Yes | Yes | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (56%) |
| O'Sullivan et al. [33] | Yes | Yes | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (56%) |
| Sheeran et al. [40] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (44%) |
| O'Sullivan et al. [31] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (44%) |
| Sheeran et al. [44] | Yes | Yes | Yes | N/A | N/A | N/A | No | N/A | Yes | Yes | Yes | Yes | Yes | Low (89%) |
| Dankarts et al. [37] | Yes | Yes | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (56%) |
| Hemming et al. [42] | Yes | Yes | No | N/A[p] | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (56%) |

**Table 6** (continued)

| Study | Item 1[a] V[n]+R[o] | Item 2[b] V+R | Item 3[c] V | Item 4[d] R | Item 5[e] R | Item 6[f] R | Item 7[g] V | Item 8[h] R | Item 9[i] V | Item 10[j] V+R | Item 11[k] V | Item 12[l] V+R | Item13[m] V+R | Risk of bias (% "Yes") * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| O'Sullivan et al. [32] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (44%) |
| Hungerford et al. [34] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (44%) |
| Beales et al. [39] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (44%) |
| Biele et al. [45] | Yes | No | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (44%) |
| Meyer et al. [46] | Yes | Yes | No | N/A | N/A | N/A | No | N/A | No | Yes | No | Yes | Yes | High (56%) |

Items: (1[a]) Detailed information of the sample of subjects; (2[b]) Qualification or competence of raters; (3[c]) Explanation of the reference standard; (4[d]) Raters blindness in inter-rater reliability; (5[e]) Raters blindness in intra-rater reliability; (6[f]) Variation of the order of examination; (7[g]) Time period between the reference standard and the index test; (8[h]) Stability of variable; (9[i]) Reference standard independent of the index test; (10[j]) Description of the index test; (11[k]) Description of the reference standard; (12[l]) Explanation of the withdrawals; (13[m]) Statistical methods; *Percentage calculated using as total number of items those that are applicable to the study; R[n]=Reliability; V[o]=Validity; N/A[p]=Not applicable; Items 4, 5, 6 and 8 are only applicable to reliability studies

## Review strengths and limitations

This review employed methodology consistent with PRISMA guidelines. However, as with all systematic reviews, articles may have been missed in the database searches. A meta-analysis was not feasible, due to heterogeneity in the methodological design and the statistical analyses employed. Only articles in the English language were included. Another limitation is the inclusion low-quality studies (small sample sizes and no reference standards).

## Conclusions

Evidence from multiple studies with low risk of bias demonstrates OCS as a reliable classification method. Strong inter-rater reliability also exists for using 3 tests of the 10-item MCI test battery. Evidence for the reliability and validity of the PBA is limited to one study with high risk of bias. While clinicians are encouraged to categorize the functional capacity of patients with NS-LBP using reliable methods, research evidence is not yet available to answer questions about the effectiveness of care informed by such classification.

# Appendix 1

## Search strategies of the searched databases and journals

| Database/journal | Last citation no | Keywords |
|---|---|---|
| PubMed | 364 | (((((Non specific OR non-specific OR nonspecific OR mechanical))) AND ((low back pain OR simple backache OR lumbar strain OR spinal degeneration))) AND ((clinical test OR clinical examination OR clinical sign))) AND ((valid* OR reliabl*)) simple search |
| EMbase | 738 | (clinical) AND (test* OR exam* OR sign*) AND (non-specific OR nonspecific OR 'non specific' OR mechanical OR simple) AND (low back pain OR back pain OR LBP) AND (reliab* OR valid*) in English only and limited to human plus searching in EMbase only |
| Cochrane | 92 | (Non specific or non-specific or nonspecific or mechanical) and (low back pain or simple backache or lumbar strain or spinal degeneration) and (clinical test or clinical examination or clinical sign) and (valid* or reliabl*) in search manager choose in Trials, Methods Studies, Technology Assessments and Economic Evaluations (Word variations have been searched) |
| PEDro | 226 | Non specific low back pain (abstract and title) in advanced search (method clinical trials) |
| CINHAL | 286 | (Non specific OR non-specific OR nonspecific OR mechanical) AND (low back pain OR simple backache OR lumbar strain OR spinal degeneration) AND (clinical test OR clinical examination OR clinical sign) AND (valid* OR reliabl*) in advanced search |
| ProQuest | 358 | (("non specific" OR "non-specific" OR "nonspecific" OR "mechanical back pain") AND ("back pain" OR "lumbar strain" OR "simple backache") AND ("clinical test" OR "clinical examination" OR "clinical sign") AND ("valid*" OR "reliab*")) AND la.exact("ENG") |
| Physical therapy journal | 460 | "non specific" "non-specific" "nonspecific" "mechanical back pain" "back pain" "lumbar strain" "simple backache" "clinical test" "clinical examination" "clinical sign" "valid*" "reliab*" |
| Chiroindex | 79 | "non specific" "non-specific" "nonspecific" "mechanical back pain" "back pain" "lumbar strain" "simple backache" "clinical test" "clinical examination" "clinical sign" "valid*" "reliab*" |
| Australian journal of physiotherapy | 54 | Non specific in *Title/Abs/Keywords* OR nonspecific in*Title/Abs/Keywords* OR non-specific in *Title/Abs/Keywords* AND Low Back Pain in*Title/Abs/Keywords* OR Mechanical low back pain in *Title/Abs/Keywords* OR simple backache in *Title/Abs/Keywords* |
| Canadian physiotherapy In advanced search | 113 | Non specific OR non-specific OR nonspecific OR mechanical AND low back pain AND clinical tests OR clinical examination OR clinical sign AND valid* OR reliabl* |
| physiotherapy theory and practice journal | 113 | Non specific OR non-specific OR nonspecific OR mechanical AND low back pain AND clinical test OR clinical examination OR clinical sign AND valid* OR reliabl* |

## Appendix 2

### Systematic review critical appraisal tool (Reproduced from Brink and Louw (2011))

Item 1: If human subjects were used, did the authors give a detailed description of the sample of subjects used to perform the (index) test on?

Why the criterion should be evaluated: The validity and reliability of a test will be affected by the sample characteristics or composition, and therefore, the study has to report on the sample characteristics because the validity and reliability scores will then only be applicable to that particular population. A study does not contribute to validity and reliability testing if the subjects were not recruited appropriately

This item can be scored yes if:

1 the sample characteristics (e.g., height, weight, age, diagnosis and symptom status) were described or the manner of recruiting subjects was stated or if selection criteria were applied

If none of the above have been described or if insufficient information was provided, select "no." If inhuman or inanimate objects were used, select N/A

Item 2: Did the authors clarify the qualification, or competence of the rater(s) who performed the (index) test?

Why the criterion should be evaluated: The amount of experience of the rater(s), performing the (index) test, will influence the validity and reliability scores and needs to be explained

This item can be scored yes if:

1 the rater(s) characteristics (e.g., qualification, specialization and amount of experience using the instrument under investigation) have been described

If the above have not been described or insufficient information was provided, select "no"

Item 3: Was the reference standard explained?

Why the criterion should be evaluated: The index test scores need to be compared to the scores obtained from the reference standard in order to test validity, and therefore, the reference standard needs to be explained appropriately

This item can be scored yes if:

1 the reference standard is likely to produce correct measurements;

2 the reference standard is the best method available; and

3 details (name of the instrument, references to the accuracy of the instrument) of the reference standard are reported

If none of the above is applicable to the reference standard's description, then select "no"

Item 4: If inter-rater reliability was tested, were raters blinded to the findings of other raters?

Why the criterion should be evaluated: When raters have access to the findings of other raters, it compromises the quality of the reliability testing procedure by inflating the agreement among the raters, and therefore, blinding needs to be performed

This item can be scored yes if:

1 it is stated that the raters were blinded to each other's findings or if a description that implies that the raters were blinded was reported

If no information is provided, then select "no." If intra-rater reliability was examined, then select "N/A"

Item 5: If intra-rater reliability was tested, were raters blinded to their own prior findings of the test under evaluation?

Why the criterion should be evaluated: If raters have knowledge of their prior own findings, it will influence the findings of their repeated measurements and could inflate the rater agreement, and therefore, appropriate measures, depending on the characteristics or the study design of the research study, need to be applied to ensure blinding

This item can be scored yes if:

1 rater(s) has/have examined the same subjects on more than one occasion, it should be stated whether the rater(s) was/were blinded to the subjects they have examined previously

If insufficient information is provided, then select "no." If inter-rater reliability was examined, then select "N/A"

Item 6: Was the order of examination varied?

Why the criterion should be evaluated: If the order is varied, in which the raters examine the subjects when inter-rater reliability is tested, it reduces the risk of systematic bias. If the order is varied in which subjects are examined by one rater when intra-rater reliability is tested, it reduces the risk of the rater recalling the previous test scores and reduces bias

This item can be scored yes if:

1 the order in which subjects were tested varied between raters if inter-rater reliability was tested;

2 the order of subjects was varied when intra-rater reliability was tested

If insufficient information is provided, then select "no." If varied order of examination is unnecessary or impractical (e.g., rater(s) digitizing or reading X-rays) then select "N/A"

Item 7: If human subjects were used, was the time period between the reference standard and the index test short enough to be reasonably sure that the target condition did not change between the two tests?

Why the criterion should be evaluated: The index test and the reference standard should be performed at the same time; however, this is not always possible. It becomes important to know whether it is possible that the test variable did not change between the two tests, otherwise it will affect the index test's validity performance

This item can be scored yes if:

1 result from the index test and the reference standard were collected on the same subjects at the same time;

2 a delay between measurements occurs, it is important that the target condition should not change between measurements

If the time period between performing the index test and the reference standard was sufficiently long that the target condition may have changed between the two tests or if insufficient information is provided, then select "no." If inhuman or inanimate objects were used, then select N/A

### Item 8: Was the stability (or theoretical stability) of the variable being measured considered when determining the suitability of the time interval between repeated measures?

Why the criterion should be evaluated: For reliability, the test variable should not change between repeated measures, otherwise it will decrease the amount of agreement obtained between and within the rater(s)

This item can be scored yes if:

1 the stability of the variable is known or reported, and reviewers then decide on an appropriate time interval between repeated measures (stability of a test variable can only be determined if there is a reference standard);

2 there is no reference standard, then the reviewers should agree upon the theoretical stability of the variable and decide on an appropriate time interval between repeated measures

If insufficient information is provided, then select "no"

### Item 9: Was the reference standard independent of the index test?

Why the criterion should be evaluated: If the reference standard and the index test are not independently performed, then the index test cannot replace the reference standard on its own

This item can be scored yes if:

1 it is clear from the study that the index test did not form part of the reference standard

If it appears that the index test formed part of the reference standard, then select "no"

### Item 10: Was the execution of the (index) test described in enough detail to permit replication of the test?

Why the criterion should be evaluated: Variations in the execution of the reference standard and the (index) test might affect the agreement between the two tests and it is also important to be able to replicate the same study procedure in another setting when needed

This item can be scored yes if:

1 the study reported a clear description of the measurement procedure (e.g., the positioning of the instrument or rater and execution sequence of events);

2 citations of methodology were supplied

The extent to which details is expected to be reported depends on the ability of different procedures to influence the results and on the type of instrument or test under evaluation

If insufficient information is provided, then select "no"

### Item 11: Was the execution of the reference standard described in enough detail to permit its replication?

Why the criterion should be evaluated: For the same reason as item 10

This item can be scored yes if:

1 the study reported a clear description of the measurement procedure (e.g., the positioning of the instrument or rater and execution sequence of events);

2 citations were supplied

If insufficient information is provided, then select "no"

### Item 12: Were withdrawals from the study explained?

Why the criterion should be evaluated: The sample composition will influence the validity and reliability performance of the (index) test; therefore, it is important to know whether any withdrawals from the sample might have changed the composition of the sample

This item can be scored yes if:

1 it is clear what happened to all subjects who entered the study;

2 subjects who entered but did not complete the study are considered

If it appears that subjects who entered but did not complete the study were not accounted for or if insufficient information is provided, then select "no." If inhuman or inanimate objects were used, then select N/A

### Item 13: Were the statistical methods appropriate for the purpose of the study?

Why the criterion should be evaluated: The aim of validity and reliability studies is to report on an estimate of validity and reliability for the particular test and appropriate statistical methods need to be implemented in order to produce this estimate

This item can be scored yes if:

1 the analysis is appropriate in terms of the type of data (e.g., categorical, continuous and dichotomous);

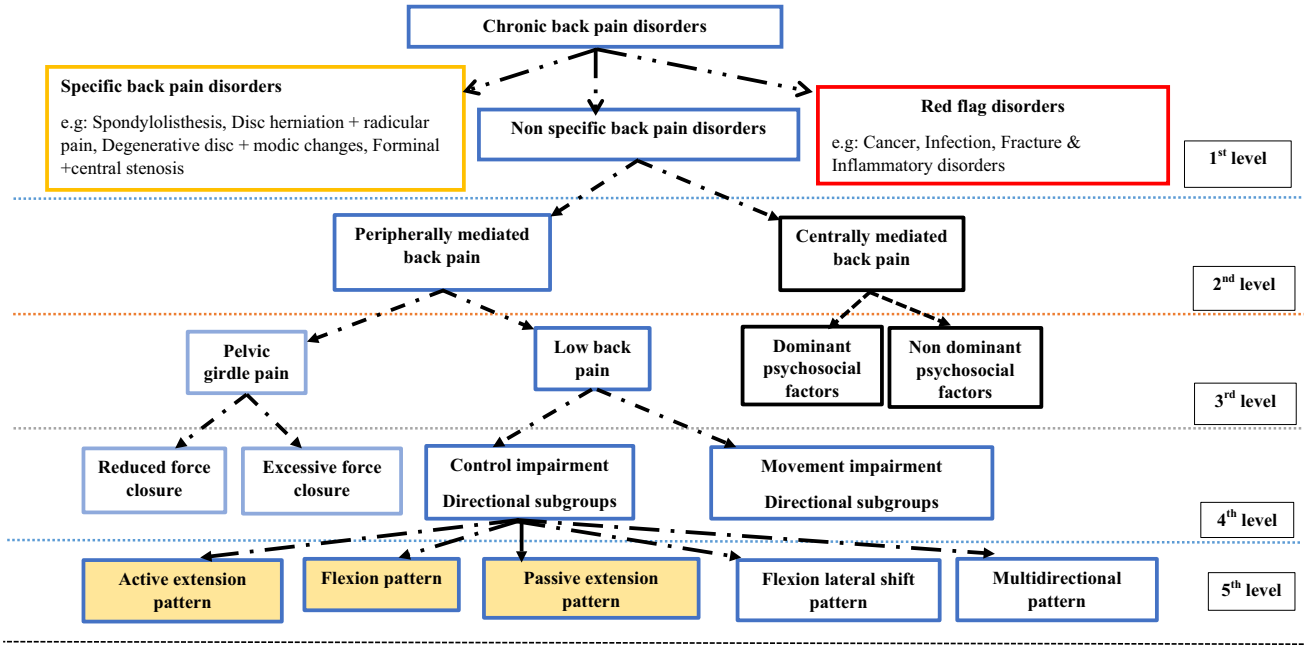
2 statistical analysis for validity studies incorporates, for example means, differences between measurements, 95% confidence interval and ANOVA; and

3 statistical analysis for reliability studies incorporates, for example, interclass correlation coefficient and 95% confidence interval

If the analysis is not appropriate or if insufficient information was provided, then select "no"

# Appendix 3

## Classification processes of OCS




## Compliance with ethical standards

**Conflict of interest** The authors have no conflicts of interest to declare that are relevant to the content of this article.

## References


1. Hartvigsen J, Hancock MJ, Kongsted A et al (2018) What low back pain is and why we need to pay attention. Lancet 391:2356–2367. https://doi.org/10.1016/S0140-6736(18)30480-X
2. Woolf AD, Pfleger B (2003) Burden of major musculoskeletal conditions. Bull World Health Organ 81:646–656
3. Buchbinder R, van Tulder M, Öberg B et al (2018) Low back pain: a call for action. Lancet 391:2384–2388. https://doi.org/10.1016/S0140-6736(18)30488-4
4. da C Menezes Costa L, Maher CG, Hancock MJ, et al (2012) The prognosis of acute and persistent low-back pain: a meta-analysis. CMAJ 184:E613–E624. https://doi.org/10.1503/cmaj.111271
5. Burton AK, McClune TD, Clarke RD, Main CJ (2004) Long-term follow-up of patients with low back pain attending for manipulative care: outcomes and predictors. Man Ther 9:30–35. https://doi.org/10.1016/s1356-689x(03)00052-3
6. Wáng YXJ, Wu A-M, Ruiz Santiago F, Nogueira-Barbosa MH (2018) Informed appropriate imaging for low back pain management: a narrative review. J Orthop Transl 15:21–34. https://doi.org/10.1016/j.jot.2018.07.009
7. Hancock MJ, Maher CG, Latimer J et al (2007) Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. Eur Spine J 16:1539–1550. https://doi.org/10.1007/s00586-007-0391-1
8. Maher C, Underwood M, Buchbinder R (2017) Non-specific low back pain. Lancet 389:736–747. https://doi.org/10.1016/S0140-6736(16)30970-9
9. Balagué F, Mannion AF, Pellisé F, Cedraschi C (2012) Non-specific low back pain. Lancet 379:482–491. https://doi.org/10.1016/S0140-6736(11)60610-7
10. Vining RD, Minkalis AL, Shannon ZK, Twist EJ (2019) Development of an evidence-based practical diagnostic checklist and corresponding clinical exam for low back pain. J Manipulative Physiol Ther 42:665–676. https://doi.org/10.1016/j.jmpt.2019.08.003
11. Patel S, Psychol C, Friede T et al (2012) Systematic review of randomized controlled trials of clinical prediction rules for physical therapy in low back pain. Spine. https://doi.org/10.1097/BRS.0b013e31827b158f
12. Amundsen PA, Evans DW, Rajendran D et al (2018) Inclusion and exclusion criteria used in non-specific low back pain trials: a review of randomised controlled trials published between 2006 and 2012. BMC Musculoskelet Disord 19:113. https://doi.org/10.1186/s12891-018-2034-6
13. Foster NE, Hill JC, Hay EM (2011) Subgrouping patients with low back pain in primary care: are we getting any better at it? Man Ther 16:3–8. https://doi.org/10.1016/j.math.2010.05.013
14. Petersen T, Laslett M, Thorsen H et al (2003) Diagnostic classification of non-specific low back pain. A new system integrating patho-anatomic and clinical categories. Physiother Theory Pract 19:213–237. https://doi.org/10.1080/09593980390246760
15. Vining R, Potocki E, Seidman M, Morgenthal P (2013) An evidence-based diagnostic classification system for low back pain. J Can Chiropr Assoc 57:189–204

16. Spitzer WO, LeBlanc FE, Dupuis M, Abenhaim L, Belanger AY, Bloch R, Bombardier C, Cruess RL, Drouin G, Duval-Hesler N, Laflamme J, Lamoureux G, Nachemson A, Page JJ, Rossignol M, Salmi LR, Salois-Arsenault S, Suissa SW-DS (1987) Scientific approach to the assessment and management of activity-related spinal disorders. A monograph for clinicians. Report of the Quebec Task Force on Spinal Disorders. Spine 12:S1-59
17. Alrwaily M, Timko M, Schneider M et al (2016) Treatment-based classification system for low back pain: revision and update. Phys Ther 96:1057–1066. https://doi.org/10.2522/ptj.20150345
18. Cosio D, Lin E (2018) Role of active versus passive complementary and integrative health approaches in pain management. Glob Adv Heal Med 7:216495611876849. https://doi.org/10.1177/2164956118768492
19. Alhowimel A, AlOtaibi M, Radford K, Coulson N (2018) Psychosocial factors associated with change in pain and disability outcomes in chronic low back pain patients treated by physiotherapist: a systematic review. SAGE Open Med. https://doi.org/10.1177/2050312118757387
20. Booth A, Clarke M, Dooley G et al (2012) The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev 1:2. https://doi.org/10.1186/2046-4053-1-2
21. Moher D, Liberati A, Tetzlaff J, Altman DG (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6:e1000097. https://doi.org/10.1371/journal.pmed.1000097
22. Brink Y, Louw QA (2012) Clinical instruments: reliability and validity critical appraisal. J Eval Clin Pract 18:1126–1132. https://doi.org/10.1111/j.1365-2753.2011.01707.x
23. May S, Littlewood C, Bishop A (2006) Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. Aust J Physiother 52:91–102. https://doi.org/10.1016/S0004-9514(06)70044-7
24. May S, Chance-Larsen K, Littlewood C et al (2010) Reliability of physical examination tests used in the assessment of patients with shoulder problems: a systematic review. Physiotherapy 96:179–190
25. Barrett E, McCreesh K, Lewis J (2014) Reliability and validity of non-radiographic methods of thoracic kyphosis measurement: a systematic review. Man Ther 19:10–17. https://doi.org/10.1016/j.math.2013.09.003
26. Vibe Fersum K, O'Sullivan PB, Kvale A, Skouen JS (2009) Inter-examiner reliability of a classification system for patients with non-specific low back pain. Man Ther 14:555–561. https://doi.org/10.1016/j.math.2008.08.003
27. Luomajoki H, Kool J (2007) Reliability of movement control tests in the lumbar spine. BMC Musculoskelet Disord 8:90. https://doi.org/10.1186/1471-2474-8-90
28. Dankaerts W, O'Sullivan PB, Straker LM et al (2006) The inter-examiner reliability of a classification method for non-specific chronic low back pain patients with motor control impairment. Man Ther 11:28–39. https://doi.org/10.1016/j.math.2005.02.001
29. Enoch F, Kjaer P, Elkjaer A et al (2011) Inter-examiner reproducibility of tests for lumbar motor control. BMC Musculoskelet Disord 12:114. https://doi.org/10.1186/1471-2474-12-114
30. O'Sullivan PB, Mitchell T, Bulich P et al (2006) The relationship beween posture and back muscle endurance in industrial workers with flexion-related low back pain. Man Ther 11:264–271. https://doi.org/10.1016/j.math.2005.04.004
31. O'Sullivan K, Verschueren S, Van Hoof W et al (2013) Lumbar repositioning error in sitting: healthy controls versus people with sitting-related non-specific chronic low back pain (flexion pattern). Man Ther 18:526–532. https://doi.org/10.1016/j.math.2013.05.005
32. O'Sullivan PB, Beales DJ, Beetham JA et al (2002) Altered motor control strategies in subjects with sacroiliac joint pain during the active straight-leg-raise test. Spine 27:E1-8. https://doi.org/10.1097/00007632-200201010-00015
33. O'Sullivan PB, Burnett A, Floyd AN et al (2003) Lumbar repositioning deficit in a specific low back pain population. Spine 28:1074–1079. https://doi.org/10.1097/01.BRS.0000061990.56113.6F
34. Hungerford B, Gilleard W, Hodges P (2003) Evidence of altered lumbopelvic muscle recruitment in the presence of sacroiliac joint pain. Spine 28:1593–1600. https://doi.org/10.1097/00007632-200307150-00022
35. Burnett A, Cornelius M, Dankaerts W, O'Sullivan P (2004) Spinal kinematics and trunk muscle activity in cyclists: a comparison between healthy controls and non-specific chronic low back pain subjects—a pilot investigation. Man Ther 9:211–219. https://doi.org/10.1016/j.math.2004.06.002
36. Dankaerts W, O'Sullivan P, Burnett A, Straker L (2006) Differences in sitting postures are associated with nonspecific chronic low back pain disorders when patients are subclassified. Spine 31:698–704. https://doi.org/10.1097/01.brs.0000202532.76925.d2
37. Dankaerts W, O'Sullivan P, Burnett A, Straker L (2006) Altered patterns of superficial trunk muscle activation during sitting in nonspecific chronic low back pain patients: importance of subclassification. Spine 31:2017–2023. https://doi.org/10.1097/01.brs.0000228728.11076.82
38. Dankaerts W, O'Sullivan P, Burnett A et al (2009) Discriminating healthy controls and two clinical subgroups of nonspecific chronic low back pain patients using trunk muscle activation and lumbosacral kinematics of postures and movements: a statistical classification model. Spine 34:1610–1618. https://doi.org/10.1097/BRS.0b013e3181aa6175
39. Beales DJ, Ther MM, O'Sullivan PB, Briffa NK (2009) Motor control patterns during an active straight leg raise in chronic pelvic girdle pain subjects. Spine 34:861–870. https://doi.org/10.1097/BRS.0b013e318198d212
40. Sheeran L, Sparkes V, Caterson B et al (2012) Spinal position sense and trunk muscle activity during sitting and standing in nonspecific chronic low back pain: classification analysis. Spine 37:E486–E495. https://doi.org/10.1097/BRS.0b013e31823b00ce
41. Van Hoof W, Volkaerts K, O'Sullivan K et al (2012) Comparing lower lumbar kinematics in cyclists with low back pain (flexion pattern) versus asymptomatic controls—field study using a wireless posture monitoring system. Man Ther 17:312–317. https://doi.org/10.1016/j.math.2012.02.012
42. Hemming R, Sheeran L, van deursen R, Sparkes V, (2019) Investigating differences in trunk muscle activity in non-specific chronic low back pain subgroups and no-low back pain controls during functional tasks: a case-control study. BMC Musculoskelet Disord 20:459. https://doi.org/10.1186/s12891-019-2843-2
43. Hemming R, Sheeran L, van Deursen R, Sparkes V (2017) Non-specific chronic low back pain: differences in spinal kinematics in subgroups during functional tasks. Eur Spine J. https://doi.org/10.1007/s00586-017-5217-1
44. Sheeran L, Sparkes V, Whatling G et al (2019) Identifying non-specific low back pain clinical subgroups from sitting and standing repositioning posture tasks using a novel cardiff Dempster–Shafer theory classifier. Clin Biomech. https://doi.org/10.1016/j.clinbiomech.2019.10.004
45. Biele C, Moller D, von Piekartz H et al (2019) Validity of increasing the number of motor control tests within a test battery for discrimination of low back pain conditions in people attending a physiotherapy clinic: a case–control study. BMJ Open 9:e032340. https://doi.org/10.1136/bmjopen-2019-032340
46. Meyer K, Klipstein A, Oesch P et al (2016) Development and validation of a pain behavior assessment in patients with chronic low back pain. J Occup Rehabil 26:103–113. https://doi.org/10.1007/s10926-015-9593-2

47. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data data for categorical of observer agreement. Biometrics 33:159–174
48. Ford J (2003) A systematic review on methodology of classification system research for low back pain. In: Musculoskeletal physiotherapy Australia 13th biennial conference, Sydney, Australia, 2003
49. Anderson JA (1977) Problems of classification of low-back pain. Rheumatol Rehabil 16:34–36. https://doi.org/10.1093/rheumatology/16.1.34
50. Deyo RA, Haselkorn J, Hoffman R, Kent DL (1994) Designing studies of diagnostic tests for low back pain or radiculopathy. Spine 19:2057S-2065S. https://doi.org/10.1097/00007632-199409151-00007
51. Fairbank JCT, Pynsent PB (1992) Syndromes of back pain and their classification. In: The Lumbar spine and back pain. Edinburgh: Churchill Livingstone
52. Petersen T, Thorsen H, Manniche C, Ekdahl C (1999) Classification of non-specific low back pain: a review of the literature on classifications systems relevant to physiotherapy. Phys Ther Rev 4:265–281. https://doi.org/10.1179/108331999786821690
53. Ford J, Story I, O'Sullivan P, McMeeken J (2007) Classification systems for low back pain: a review of the methodology for development and validation. Phys Ther Rev 12(33–42):10p
54. Woolf CJ, Bennett GJ, Doherty M et al (1998) Towards a mechanism-based classification of pain. Pain 77:227–229
55. McCarthy CJ, Arnall FA, Strimpakos N et al (2004) The biopsychosocial classification of non-specific low back pain: a systematic review. Phys Ther Rev 9:17–30. https://doi.org/10.1179/108331904225003955
56. Fairbank J, Gwilym S, France J, Daffner S (2011) The role of classification of chronic low back pain. Spine 1:36. https://doi.org/10.1097/BRS.0b013e31822ef72c
57. Salvioli S, Pozzi A, Testa M (2019) Movement control impairment and low back pain: state of the art of diagnostic framing. Medicina (Kaunas). https://doi.org/10.3390/medicina55090548
58. Carlsson H, Rasmussen-Barr E (2013) Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra-and inter-observer reliability studies. Man Ther 18:103–110. https://doi.org/10.1016/j.math.2012.08.004
59. Murphy SE, Blake C, Power CK, Fullen BM (2016) Comparison of a stratified group intervention (STarT back) with usual group care in patients with low back pain: a nonrandomized controlled trial. Spine 41:645–652. https://doi.org/10.1097/BRS.0000000000001305
60. Mjøsund HL, Boyle E, Kjaer P et al (2017) Clinically acceptable agreement between the ViMove wireless motion sensor system and the Vicon motion capture system when measuring lumbar region inclination motion in the sagittal and coronal planes. BMC Musculoskelet Disord 18:124. https://doi.org/10.1186/s12891-017-1489-1
61. Gracovetsky S, Newman N, Pawlowsky M et al (1995) A database for estimating normal spinal motion derived from non-invasive measurements. Spine 20:1036–1046. https://doi.org/10.1097/00007632-199505000-00010
62. Mannion AF, Knecht K, Balaban G et al (2004) A new skin-surface device for measuring the curvature and global and segmental ranges of motion of the spine: reliability of measurements and comparison with data reviewed from the literature. Eur Spine J 13:122–136. https://doi.org/10.1007/s00586-003-0618-8
63. Öztuna D, Elhan AH, Tüccar E (2006) Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. Turkish J Med Sci 36:171–176
64. Thode HC (2002) Statistics: textbooks and monographs 164 Testing for normality. CRC Press, New York, NY