# Inter- and intra-rater reliability of clinical tests associated to functional lumbar segmental instability and motor control impairment in patients with LBP : a systematic review

**Reference:**

Denteneer Lenie, Stassijns Gaëtane, de Hertogh Willem, Truijen Steven, van Daele Ulrike.- Inter- and intra-rater reliability of clinical tests associated to functional lumbar segmental instability and motor control impairment in patients with LBP : a systematic review

Archives of physical medicine and rehabilitation - ISSN 0003-9993 - (2016), p. 1-14

Full text (Publishers DOI): http://dx.doi.org/doi:10.1016/j.apmr.2016.07.020


# Inter- and intra-rater reliability of clinical tests associated to functional lumbar segmental instability and motor control impairment in patients with LBP: a systematic review

Lenie Denteneer, Dra, MT, PT, Gaetane Stassijns, PhD, MD, Willem De Hertogh, PhD, MT, PT, Steven Truijen, PhD, MSc, Ulrike Van Daele, PhD, MT, PT

PII: S0003-9993(16)30896-6

DOI: 10.1016/j.apmr.2016.07.020

Reference: YAPMR 56643

To appear in: *ARCHIVES OF PHYSICAL MEDICINE AND REHABILITATION*

Received Date: 28 April 2016

Revised Date: 15 July 2016

Accepted Date: 15 July 2016

Please cite this article as: Denteneer L, Stassijns G, De Hertogh W, Truijen S, Van Daele U, Inter- and intra-rater reliability of clinical tests associated to functional lumbar segmental instability and motor control impairment in patients with LBP: a systematic review, *ARCHIVES OF PHYSICAL MEDICINE AND REHABILITATION* (2016), doi: 10.1016/j.apmr.2016.07.020.

Title: Inter- and intra-rater reliability of clinical tests associated to functional lumbar segmental instability and motor control impairment in patients with LBP: a systematic review.

1. Lenie Denteneer **(1st author)**, Dra, MT, PT
   Faculty of Medicine and Health Sciences, rehabilitation and physiotherapy
   University of Antwerp, Universiteitsplein 1, 2610 Wilrijk
   Tel: 0032494884189
   lenie.denteneer@uantwerpen.be
   **= corresponding author**
2. Gaetane Stassijns **(2nd author)**, PhD, MD
   Faculty of Medicine and Health Sciences, University of Antwerp
   Universiteitsplein 1, 2610 Wilrijk
   Antwerp University Hospital, Physical medicine and rehabilitation
   Wilrijkstraat 10, 2650 Edegem
   gaetane.stassijns@uza.be
3. Willem De Hertogh **(3th author)**, PhD, MT, PT
   Faculty of Medicine and Health Sciences, rehabilitation and physiotherapy
   University of Antwerp, Universiteitsplein 1, 2610 Wilrijk
   willem.dehertogh@uantwerpen.be
4. Steven Truijen **(4th author)**, PhD, MSc
   Faculty of Medicine and Health Sciences, rehabilitation and physiotherapy
   University of Antwerp, Universiteitsplein 1, 2610 Wilrijk
   Steven.truijen@uantwerpen.be
5. Ulrike Van Daele **(last author)**, PhD, MT, PT
   Faculty of Medicine and Health Sciences, rehabilitation and physiotherapy
   University of Antwerp, Universiteitsplein 1, 2610 Wilrijk
   Ulrike.vandaele@uantwerpen.be

The submitted manuscript does not contain information about medical device(s)/drug(s).

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

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

## Abstract

**Objective:** To provide a comprehensive overview of clinical tests associated with functional lumbar segmental instability and motor control impairment in patients with low back pain (LBP) and to investigate their intra and/or inter-rater reliability.

**Data sources:** A systematic computerized search was conducted in four different databases on the 1st of December 2015: Pubmed (1972 - ) , Web of Science (1955 - ), Embase (1947 - ), Medline ( 1946 - ).

**Study selection:** PRISMA guidelines were followed during design, search and reporting stages of this review. The included population are patients with primary LBP. Data was extracted as follows: (1) description and scoring of the clinical tests (2) population characteristics (3) in- and exclusion criteria (4) description of the used procedures (5) results for both intra- and inter-rater reliability and eventually (6) notification on used statistical method. The risk of bias (ROB) of the included articles was assessed with the use of the COSMIN checklist.

**Data synthesis:** A total of 16 records were eligible and 30 clinical tests were identified. All included studies investigated inter-rater reliability and three studies investigated intra-rater reliability. The identified Inter-rater reliability scores ranged from poor to very good (k-0.09-0.89 and ICC 0.72-0.96) and the Intra-rater reliability scores ranged from fair to very good (k 0.51-0.86).

**Conclusions:** Three clinical tests (*aberrant Movement pattern, prone instability test* and the *beighton scale*) could be identified for having an adequate inter-rater reliability. No conclusions could be made for intra-rater reliability. However, further research should focus on better study designs, provide an overall agreement for uniformity and interpretation of clinical tests and should implement research regarding validity.


**Keywords:** low back pain, reliability, motor control impairment, lumbar segmental instability, clinical test

**Abbreviations**

LBP: Low back pain

LSI: Lumbar segmental instability

MCI: motor control impairment

ROB: risk of bias

ICC: intraclass correlation coefficient

GRADE: Grading of Recommendations Assessment, Development and Evaluation

Patients with low back pain (LBP) form a large and heterogeneous population[1]. However, patients who are diagnosed with functional lumbar segmental instability (LSI) associated to motor control impairment (MCI) are considered as a unique subgroup within this LBP population[2]. Throughout the last two decades there has been an increasing awareness of the importance and relevance of the specialized and integrated action of the muscle system in maintaining stability and optimal function of the movement system[3]. A dysfunction of the lumbar local stabilizer system might result in MCI which is associated with a loss of control of joint neutral position and might result in functional LSI[4].

It is important to note that a difference has to be made between radiological LSI and functional LSI. Functional LSI can be present and cause LBP despite the presence of any radiological anomaly. Also, these two categories of instability do not seem to be correlated very well[5].

Because of the significant size of this identified subgroup, clinical tests which assess functional LSI and MCI are widely used among many health care workers. However, a genuine consensus with regard to the best method to be used is lacking up to this point[5].

As new tests are developed, testing for reliability is considered before validity as a test cannot be considered valid if the results turn out to be inconsistent[6]. Therefore, to determine functional LSI and MCI in individual patients with LBP, reliable outcome measures need to be available[7]. Once sufficiently reliable outcome measures are identified they can form a base for further research regarding validity. Eventually, this process can lead to objectively subgrouping patients with LBP which in turn can enable clinicians to establish a tailor-made management strategy[8].

An important advantage of these clinical tests is that they can be performed without any additional (expensive) devices. Consequently, they can be considered as a cost-effective method to evaluate pain and disability in patients with LBP.

The aim of this study is to provide a comprehensive overview of all clinical tests associated to functional LSI and MCI in patients with LBP and to investigate their inter- and/or intra-rater reliability.

## Methods

A systematic literature review was conducted according to the PRISMA guidelines[9].

### *Data sources and search strategy*

A systematic computerized search (Appendix 1) was conducted by one author in 4 different databases on the 1st of December 2015: Pubmed (1972 - ) , Web of Science (1955 - ), Embase (1947 - ), Medline ( 1946 - ). To reduce the search bias, the search strategy was conducted using Medical Subject Headings (MeSH terms). A brief search in ClinicalTrials.gov was conducted to identify any possible ongoing studies. Identified records were uploaded into Endnote (Thomas Reuters), and duplicates were removed. Two independent researchers selected studies for inclusion in a two-step process. First, articles were screened based on title and abstract. If there was no consensus between the two reviewers, the article was included into the second stage without deliberation. The second stage, the screening on full text, was also conducted

independently by both reviewers. In case of disagreement, the reviewers came to a consensus during a deliberation session.

*Study selection*

The included population are patients with primary LBP. The current review will include studies concerning the reliability of clinical tests associated with functional LSI and/or associated with MCI. These clinical test have to be investigated for their reliability in patients with LBP. An overview of in-and exclusion criteria are shown in Text Box 1.

*Data extraction*

Relevant data was extracted as follows: (1) description and scoring of the clinical tests (2) population characteristics (3) in- and exclusion criteria (4) description of the used procedures (5) results for both intra- and inter-rater reliability and eventually (6) notification on used statistical method.

*Risk of bias assessment*

The risk of bias assessment (ROB) of the included articles was independently assessed by two authors with the use of the COSMIN checklist which was developed in 2010 according to a Delphi study by international experts in health related measurement instruments[10]. The COSMIN checklist evaluates 10 possible psychometric property boxes. One of these boxes concern "reliability", for which 14 questions need to be answered[11]. The COSMIN 4-point scale was used to assess the ROB for each question and enables an objective ROB scoring of articles in systematic reviews. The COSMIN checklist provides an option to calculate a final score. This final score is obtained by taking the lowest rating (namely the "worse score counts" algorithm). For each article, the final ROB is then rated as "excellent", "good", "fair", or "poor"[12]. It was decided not to use this final score system since it might induce a bias in interpretation due to a too strict scoring system. Instead, individual scores on the different COSMIN subitems will be assessed and will decide if they form an important bias factor for this review.

The current review will assess four factors to come to final conclusions for each clinical test (ROB, consistency, directness and precision). The "precision" factor assesses the sample sizes for each clinical test, and therefore the COSMIN subitem which investigates the included sample size will be excluded from the ROB.

*Reliability*

Reliability can be defined as the consistency of measurements, or of an individual's performance on a test or of the absence of measurement errors[6]. Results from one subject examined by the same observer (intra-rater or test-retest reliability) or by several observers examining the same subject (inter-rater reliability) should stay consistent[6]. Nominal and ordinal unpaired data are often analyzed with Cohen's or the weighted kappa ($\kappa$) coefficient[13] . $\kappa$ values may vary between -1 and 1.

Agreement can be interpreted as κ < 0.20 = poor, κ: 0.21-0.40 = fair, κ: 0.41-0.60 = moderate, κ: 0.61-0.80 = good, κ:0.81-1.0 = excellent[14]. In some situations, when the prevalence of a given response to a test is either very high or very low, the interpretation of the k statistic does not satisfactorily reflect the true level of agreement[15]. Other statistical tools have been developed to account for this, such as the "prevalence-adjusted bias-adjusted κ" (PABAK), which corrects for this type of bias[16]. Continuous data is often analyzed with intra class correlation coefficients (ICC)[13]. Whatever the type of ICC that is calculated, it is suggested that an ICC close to 1 indicates 'excellent' reliability. An ICC exceeding 0.70 indicates good reliability and an ICC below 0.70 indicates moderate to poor reliability[17].

*Factors for final conclusions*

To come to final conclusions for each of the identified clinical tests, the current review will assess four factors which are all part of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) method[18]:

1) ROB: the outcome of the COSMIN checklist will be systematically used to guide interpretation of results. Included records with "excellent" and "good" scores throughout the different subitems will be given higher value to come to final conclusions than the ones with "fair" or "poor" outcome.

2) Consistency: tests with at least a "good" or "excellent" kappa value and if applicable at least a "good" ICC value, will be taken into account to come to final conclusions. Also, higher value will be given to clinical tests which have been investigated by at least two independent authors.

3) Directness: studies with similar population, similar test type (description); and similar outcome (kappa or ICC levels) are given higher value to come to final conclusions

4) Precision: larger sample sizes or study population (especially if compiled from multiple publications) are part of the last factor that will decide the final conclusions for the current review.

## Results

The database search strategy yielded 673 records. After removal of duplicates, 435 records were screened on title and abstract for eligibility with 55 records going through to the second screening phase. After screening on full text a total of 16 [19-34] records were determined to be eligible (Figure 1). In total, 30 clinical tests were identified representing a total patient sample of 618 patients (Table 1). A detailed description for the interpretation and execution of each clinical test is provided within Supplemental File 1.

*Risk of bias assessment*

Out of the 16 included studies, one study[26] scored "excellent" or "good" on all subitems of the COSMIN checklist and therefore was identified as a low ROB study. A total of nine studies [19,23-25,29,31-34] scored at least at one occasion "fair" and six studies[20-22,27,28,30] were rated at least one time as "poor" within one of the subitems of the COSMIN checklist (Table 2).

*Inter-rater reliability*

Results for inter-rater reliability are shown in table 1. All included articles assessed the inter-rater reliability of the identified clinical tests. A total of 21 clinical tests were investigated by a single author[20-22,25-30,34]. Out of the remaining nine clinical tests who were investigated by multiple authors, five tests were in agreement on inter-reliability (*beighton scale, instability catch, posterior shear test, active straight leg raise* and *sitting knee extension test*) [20,22,24,26,27,30] and four were not (*prone instability test, aberrant movement pattern, reversal lumbopelvic rhythm* and the *passive lumbar extension test*)[19,20,23,24,26,30,31,33].

1) Clinical tests described by multiple authors who were not in agreement for inter-rater reliability: First, *The prone instability test* has been described in seven studies (311 participants) and was mainly described as a good reliable test by Hicks et al[26] (k=0.87), Fritz et al[24] (k=0.69), Rabin et al[30] (k=0.67) and Alyazedi et al[19] (k 0.71). On the other hand, Schneider et al[33] and Fritz et al[23] rated the *prone instability test* as a moderate reliable test (k=0.46-0.54) and Ravenna et al[31] rated is as being a poor reliable test (k 0.04-0.10). Second, The *aberrant movement pattern* was investigated in six different studies (273 participants) and was mainly considered to have a good inter-rater reliability by Rabin et al[30] (k=0.64), Biely et al[20] (k=0.68) and Alyazedi et al[19] (k=0.79). Hicks et al[26] scored the *aberrant movement pattern* as a moderate reliable test (k=0.60). Fritz et al rated this test as a poor inter-rater reliable test (k-0.07-0.18)[23,24] Third, For the *lumbopelvic rhythm test* (94 participants), Biely et al[20] rated a high kappa value (k=0.89) and Hicks et al[26] rated a poor kappa value (k=0.16) and fourth, the *passive lumbar extension test* (70 participants) was rated by Alyazedi et al[19] (k=0.46) as a moderate reliable test and as a good reliable test by Rabin et al[30] (k=0.76).

2) Clinical tests described by multiple authors who were in agreement for inter-rater reliability: First, good inter-rater reliability was found for the *beighton scale* (k=0.79 and ICC=0.72); 112 participants ; 2 studies[24,26] and *sitting knee extension test* (k=0.72 and 0.96); 52 participants ; 2 studies [22,27]. Second, Moderate inter-rater reliability was found for the *active straight leg raise* (k=0.39-0.53); 66 participants ; 2 studies [30,32] and Third, fair reliability was found for the *instability catch* (k=0.25-0.35); 94 participants ; 2 studies [20,26] and the *posterior shear test* (k=0.27-0.35); 112 participants ; 2 studies [24,26] .

3) Clinical tests described by a single author: First, The *painful arc* (63 participants), *sagittal deviation* (31 participants), *hip extension* (42 participants), *joint position sense* (25 participants), *sitting forward lean* (25 participants), *bent knee fall out* (25 participants), *waiters bow* (27 participants), *rocking backward* (27 participants), *multifidus* lift (32 participants) and the *pelvic tilt test* (27 participants) were all rated as tests with a good inter-rater reliability (k 0.61-0.81 and ICC 0.90-0.96)

[20,22,25-28]. Second, Moderate inter-rater reliability was indicated for *the prone knee bend extension* (27 participants), *prone knee bend rotation* (27 participants), *rocking forward* (27 participants) and the *lumbar extension load test* (30 participants) with kappa values ranging from 0.47 to 0.58[27,30]. Third, The *crook lying* (27 participants) showed fair inter-rater reliability (k=0.38)[27] and Fourth, The *Gower sign* (63 participants) and the *active hip abduction test* (30 participants) had a poor inter-rater reliability (k-0.09-0.00)[26,30].

4) Clinical tests described by a single author but with inconsistent results for inter-rater reliability: First, For The *single leg stance test*, Loumajoki et al[27] (27 participants) found a moderate inter-rater reliability for the right leg (k=43) and a good inter-rater reliability for the left leg (k=0.65). Second, For the *thoracolumbar dissociation test,* Elgueta et al[21] (20 participants) found a good reliability pre training (k=0.66) and a moderate reliability post training (k=0.51). Third, Sedaghat et al[34] (34 participants) investigated the inter-rater reliability of *manual palpation of the m. transversus abdominis* during different functional movements. The reliability ranged from fair to poor (k=0.07-0.30)[34] and Fourth, for *palpation of the m. multifidus*, Qvistgaard et al[29] (60 participants) calculated a poor inter-rater reliability (k 0.12) for a perfect match and a moderate inter-rater reliability for an acceptable match (k=0.48).

*Intra-rater reliability*

Three studies investigated the intra-rater reliability, addressing the following clinical tests: *the single leg stance* (63 participants), *active straight leg raise* (36 participants), *sitting knee extension* (27 participants), *waiters bow* (27 participants), *rocking backward* (27 participants), *pelvic tilt* (27 participants), *prone knee bend extension* (27 participants), *prone knee bend rotation* (27 participants), *rocking forward* (27 participants), *crook lying* (27 participants) and *thoracolumbar dissociation test* (20 participants). All identified tests were investigated by a single author except for the *single leg stance* test but for this test, the two identified studies were in agreement with each other. Reliability values (k-values) ranged from 0.67 to 0.95, indicating a good to very good intra-rater reliability[21,27,32] except for the *rocking backward test* and the *thoracolumbar dissociation test* who were rated as being fair intra-rater reliable tests (k 0.51-0.78)[21,27].

## Discussion

The current review aimed to provide a complete overview of inter-and/or intra-reliability for clinical tests associated to functional LSI and MCI in patients with LBP. A total of 30 different clinical tests were identified.

*overall completeness and applicability of the evidence for inter-rater reliability*

A total of 21 clinical tests were investigated for inter-rater reliability by a single author[20-22,25-30,34]. Out of the remaining nine clinical tests who were investigated by multiple authors, five were in agreement on inter-rater reliability[20,22,24,26,27,30] and four were not[19,20,23,24,26,30,31,33]. The clinical tests who were not in agreement between different authors will be further discussed:

**First**, the *prone instability test* was identified as a very good inter-rater reliable test (k=0.87) in the lowest ROB study, namely Hicks et al[26]. Two articles[19,24] who rated the *prone instability test* as a good reliable test (k=0.69-0.71) both showed fair scores on the COSMIN subitem "handling missing items". Three articles rated the PIT as a poor or moderate reliable test (k=0.04-0.54)[23,31,33]. Two of these studies showed "fair" COSMIN scores regarding the "time interval" question[23,33] and two studies scored "fair" on the question for "handling missing items"[23,31]. In order to come to final conclusions, the GRADE method was further applied and more value was given to studies without flaws against the time interval. All outcome across the different studies was systematically calculated trough kappa calculations. The description of the *prone instability test* throughout the seven included articles was also very similar and the test results were based upon a large sample of 311 patients who all had LBP. Therefore, in the current review, the *prone instability test* is finally concluded to be a good inter-rater reliable test. This is in slight contrast to the review of Ferrari et al[35] who has previously rated the *prone instability test* as a moderate to good reliable test.

**Second,** the *aberrant movement pattern* was scored by Hicks et al[26] as a moderate reliable test (k=0.60). However, a kappa value of 0.60 can be considered as a borderline good reliable outcome since a kappa value of 0.61 is considered as good reliable and 0.60 is considered as moderate reliable. Three other authors[19,20,30] identified the *aberrant movement pattern* as a good inter-rater reliable test (k=0.64-0.79). However, both Biely and Rabin et al[20,30] failed to include a second measurement moment and were scored "poor" on this COSMIN subitem. Alyazedi et al[19] on the other hand did show overall better COSMIN scores and only lost points on the "handling missing items" question. Two articles rated the *aberrant movement pattern* as a poor inter-rater reliable test, but it is important to note that these articles were both written by the same author (fritz et al[23,24]). One study scored "fair" on "time interval"[23] and the other study scored "fair on the "handling missing items" question[24]. Next to the investigation of ROB and consistency of this clinical test, the two remaining factors of the GRADE method namely "directness" (The *aberrant movement pattern* was consistently described throughout the different included studies, only patients with LBP were evaluated and similar kappa calculations were used in all included studies) and "precision" (a total of 279 patients with LBP were included) helped to come to the final conclusion that the aberrant movement pattern can be considered as a moderate to good reliable clinical test. This result lies in contrast with Ferrari's review[35] who stated the aberrant movement pattern is only insufficient to moderate reliable. The current review enlarged the previous identified results with an extra three records representing an enlarged sample size of 131 patients.

**Third,** for the *passive lumbar extension test*, inconclusive results might be explained because of a different interpretation between authors: a positive test is described by Rabin et al[30] when LBP is elicited and by Alyazedi et al[19] when a heaviness is felt or if the feeling as though the lower back were about to 'come off'. These minor but significant differences might cause a different interpretation of a positive clinical test. In a review conducted by Ferrari et al[35] the *passive lumbar extension test* was concluded as being the most suitable test for detecting functional LSI. This conclusion was however based upon a single

author (Rabin et al[30]). It is important to note that in the current review Rabin et al was identified with a "poor" COSMIN score for the subitem that investigates the presence of a second measurement moment. Also, this study included a low sample size of patients with LBP (n=30). In the current review, an additional article was identified namely Alyazedi et al[19] which included a larger sample size and was identified with overall higher COSMIN scores. However, Based on the four factors of de GRADE method, the current review cannot make a final conclusion regarding the *passive lumbar extension test* because both of the included authors interpreted the outcome different, a presence of ROB in the study described by Rabin et al[30] and the inconclusive results regarding the reliability between both studies.

**Fourth,** the above identified problem for the *passive lumbar extension test* was also noted for the *lumbopelvic rhythm test*. This might explain the major differences for reliability between the two included authors who investigated this test. Finally, because of the very low reliability outcome calculated by Hicks et al[26] and the lower COSMIN scores in the other study (Biely et al[20]), the authors of the current review could not identify the *lumbopelvic rhythm test* as a good reliable test.

Conclusively, good inter-rater reliability outcome was found for a total of 15 clinical tests (*painful arc, gower sign, sagittal deviation, prone instability test, multifidus lift test, sitting knee extension test, beighton scale, hip extension test, joint position sense, sitting forward lean, bent knee fall out, waiters bow, rocking backward, aberrant movement pattern and the pelvic tilt test*). However, it is questionable that high value should be given to clinical tests who were investigated by a single author. Also, the individual articles often included a low study sample where 3/16 articles included a sample size less than thirty patients, 9/16 articles included a sample size less than fifty patients and with a largest study sample of 63 patients. If all the factors of the GRADE method were being taken into account, the following tests could be identified for having an adequate inter-rater reliability: *aberrant movement pattern, beighton scale* and the *prone instability test*.

The *painful arc* test has bas been assessed as a good reliable test in a low ROB design but has been investigated by a single author. For this reason, the *painful arc* has not yet been recommended as a good reliable test in the current review.

<u>*overall completeness and applicability of the evidence for intra-rater reliability*</u>

A total of ten clinical tests were investigated for intra-rater reliability by a single author[21,27,32]. The *Single leg stance test* was the only one which was investigated by two authors[27,32] and both came to the same conclusions. The identified tests with good intra-rater reliability were: *single leg stance, active straight leg raise, sitting knee extension, waiters bow, pelvic tilt, prone knee bend extension, prone knee bend rotation* and the *crook lying test*. The *single leg stance test* was rated by two authors as a good/very good reliable test. However, Luomajoki et al[27] failed to include a second measurement moment and included a low sample size of 27 patients with LBP. Roussel et al[32] on the other hand showed better COSMIN scores, but included 36 patients which can also be seen as a relatively low patient sample. Also, an inconsistency in the description for this test between the two authors was found. For these reasons, the single leg stance could not yet be recommended as a

good reliable test in the current review. Intra-rater reliability was investigated in Carlsson's review[36] and similar results were noted in the current review.

All included studies investigated inter-rater reliability[19-34] and three studies investigated intra-rater reliability[27,32,34]. There seems to be a lack of interest towards the investigation of intra-rater reliability. This is in line with two previous conducted reviews[35,36] who also reported on the reliability of clinical tests for the assessment of motor control impairment or functional LSI. A possible reason for this is that, in the literature, clinical tests are often used in randomized controlled trials, multicenter trials or applied by more than one researcher. Therefore, the importance of high and well investigated inter-rater reliability seems to be more relevant if clinical tests are used for scientific research. In the clinical practice however, patients are often followed by the same health care worker and intra-rater reliability gains importance. Conclusively, additional studies addressing the intra-rater reliability of this are needed and recommended.

*Quality of the evidence*

To assess the ROB, the current study used the COSMIN checklist[10]. Out of the 16 included studies, one study was identified as having overall good COSMIN scores and therefore as having a low ROB[26]. The main reasons for higher ROB assessments in the remaining 15 studies are first, the lack of description of how missing data was handled (11/15); second, the lack of an implementation of two measurement moments (5/15) and third, issues regarding time interval between the two measurements (4/15) (table 2). The authors acknowledge the importance of these subitems. Firstly, if outcome within a reliability study is based upon a single measurement, it is questionable that these results are applicable in clinical practice. Secondly, the time interval between the administrations must be appropriate. Time interval should be long enough to prevent recall bias and short enough to ensure that patients have not been changed on the construct to be measured. Thirdly, the lack of description how missing data was handled is an important flaw within a study design however it is not necessarily a fatal flaw in a study[12]. Therefore, the authors of the current study decided to give more weight to flaws against the lack of a second measurement moment and time interval errors.

It appears that the COSMIN assesses both internal and in part the external validity of a study as it includes adequate sample size as one of the ROB criteria. The GRADE method which assesses the "precision" factor (or imprecision to be more exact) is designed to assess if a sample size is adequate. This precision factor is usually a separate construct that relates to the external validity of studies and can be seen separate from systematic ROB (which normally assess internal validity) or random error. For this reason, the adequateness of the sample size was not used to downgrade a study for ROB, but rather to down or upgrade the cumulated data from multiple trials.

*Agreement and disagreement with other studies or reviews*

The results of this review contribute to the literature since it gives new insights and adds to the information provided by two similar reviews: Ferrari et al[35] searched the literature for only four predefined clinical tests and therefore had a different approach than the current review which aimed to give a complete overview of all described clinical test for the assessment of functional LSI associated to MCI. The second review by Carlsson et al[36] seems to be comparable to the current study but identified eight eligible studies where this paper included 16 eligible papers. Carslsson et al discuss the limitations of their inclusion criteria and search strategy and state that they therefore might have missed other relevant articles. By applying a more comprehensive strategy in the current review we believe the chance of missing relevant articles is less.

To our knowledge, this is a unique review on inter- and intra-rater reliability that contains additional information compared to previous reviews within the topic of clinical tests for the assessment of functional LSI associated to MCI.

*Study limitations and potential bias in the review process*

Apart from the brief search in the ClincalTrials.gov database, no additional search was performed for grey literature. The current review also solely included studies written in English. These two factors might have caused a potential publication bias. Even though we were able to identify an extra eight studies in the current review, a total of 21 clinical tests remain researched by a single author which makes it difficult to come to final conclusions. Another problem for formulating a strong conclusion or advice is that only one article could be rated as having overall “good” or “excellent” scores for all the subitems within the COSMIN checklist. Also, some tests seem to be slightly different interpreted and described within different studies which makes it difficult to compare study results. The current study included patients with diagnosed primary LBP. This sample remains a heterogeneous sample and therefore the general conclusions should be interpreted with caution. The current study tried to make a statement regarding the association for each test to LSI or MCI (supplemental online only file 1). It is important to note that this association is solely based on the information that was provided within the included studies. It gives an insight on the possible interpretation for each clinical test, but it does not say anything about validity. In order to truly investigate the association to LSI and MCI, future research should focus on the validity and responsiveness of the identified studies in the current review.

Conclusions

Inter-rater reliability is well investigated and based upon the previously described GRADE method three clinical tests could be concluded to have an adequate inter-rater reliability (*prone instability test, aberrant movement pattern, beighton scale*). The intra-rater reliability of the identified clinical tests in the current review is poorly investigated and no final conclusions could be made. Further investigation should focus on better study designs to improve the overall ROB assessment and

more research on the intra-rater reliability should be done. In the future, authors should also use identical protocols for the description of clinical tests in order to be able to generalize results and compare them in-between different studies. The assessment of reliability is only a first step in the recommendation process for the use of clinical tests. In future research, the identified clinical tests in the current review should be further investigated for validity in high quality studies. Only when the reliability, validity and responsiveness of a clinical test has been thoroughly investigated, a final conclusion regarding the clinical and scientific use of the identified tests can be made.


1. Denteneer L, Van Daele U, De Hertogh W, et al. Identification of Preliminary Prognostic Indicators for Back Rehabilitation in Patients With Nonspecific Chronic Low Back Pain: A Retrospective Cohort Study. *Spine* 2016;41:522-9.
2. Hicks GE, Fritz JM, Delitto A, et al. Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. *Archives of physical medicine and rehabilitation* 2005;86:1753-62.
3. Comerford MJ, Mottram SL. Movement and stability dysfunction--contemporary developments. *Manual therapy* 2001;6:15-26.
4. Comerford MJ, Mottram SL. Functional stability re-training: principles and strategies for managing mechanical dysfunction. *Manual therapy* 2001;6:3-14.
5. Demoulin C, Distree V, Tomasella M, et al. Lumbar functional instability: a critical appraisal of the literature. *Annales de readaptation et de medecine physique : revue scientifique de la Societe francaise de reeducation fonctionnelle de readaptation et de medecine physique* 2007;50:677-84, 69-76.
6. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26:217-38.
7. O'Sullivan P. Diagnosis and classification of chronic low back pain disorders: maladaptive movement and motor control impairments as underlying mechanism. *Manual therapy* 2005;10:242-55.
8. Airaksinen O, Brox JI, Cedraschi C, et al. Chapter 4. European guidelines for the management of chronic nonspecific low back pain. *European spine journal : official publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society* 2006;15 Suppl 2:S192-300.
9. Moher D, Liberati A, Tetzlaff J, et al. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA StatementThe PRISMA Statement. *Annals of Internal Medicine* 2009;151:264-9.
10. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 2010;19:539-49.
11. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010;10:22.
12. Terwee C, Mokkink L, Knol D, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Quality of Life Research* 2012;21:651-7.
13. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil* 1998;12:187-99.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.

15. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257-68.
16. Chen G, Faris P, Hemmelgarn B, et al. Measuring agreement of administrative data with chart data using prevalence unadjusted and adjusted kappa. *BMC Med Res Methodol* 2009;9:5.
17. Fleiss JL. Reliability of Measurement. *The Design and Analysis of Clinical Experiments*: John Wiley & Sons, Inc., 1999:1-32.
18. Gopalakrishna G, Mustafa RA, Davenport C, et al. Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable. *Journal of clinical epidemiology* 2014;67:760-8.
19. Alyazedi FM, Lohman EB, Swen RW, et al. The inter-rater reliability of clinical tests that best predict the subclassification of lumbar segmental instability: Structural, functional and combined instability. *Journal of Manual and Manipulative Therapy* 2015;23:197-204.
20. Biely SA, Silfies SP, Smith SS, et al. Clinical Observation of Standing Trunk Movements: What Do the Aberrant Movement Patterns Tell Us? *Journal of Orthopaedic & Sports Physical Therapy* 2014;44:262-72.
21. Elgueta-Cancino E, Schabrun S, Danneels L, et al. A clinical test of lumbopelvic control: Development and reliability of a clinical test of dissociation of lumbopelvic and thoracolumbar motion. *Manual Therapy* 2014;19:418-24.
22. Enoch F, Kjaer P, Elkjaer A, et al. Inter-examiner reproducibility of tests for lumbar motor control. *BMC musculoskeletal disorders* 2011;12:114.
23. Fritz JM, Brennan GP, Clifford SN, et al. An examination of the reliability of a classification algorithm for subgrouping patients with low back pain. *Spine* 2006;31:77-82.
24. Fritz JM, Piva SR, Childs JD. Accuracy of the clinical examination to predict radiographic instability of the lumbar spine. *European Spine Journal* 2005;14:743-50.
25. Hebert JJ, Koppenhaver SL, Teyhen DS, et al. The evaluation of lumbar multifidus muscle function via palpation: Reliability and validity of a new clinical test. *Spine Journal* 2015;15:1196-202.
26. Hicks GE, Fritz JM, Delitto A, et al. Interrater Reliability of Clinical Examination Measures for Identification of Lumbar Segmental Instability. *Archives of Physical Medicine and Rehabilitation* 2003;84:1858-64.
27. Luomajoki H, Kool J, de Bruin ED, et al. Reliability of movement control tests in the lumbar spine. *BMC musculoskeletal disorders* 2007;8:90.
28. Murphy DR, Byfield D, McCarthy P, et al. Interexaminer reliability of the hip extension test for suspected impaired motor control of the lumbar spine. *Journal of manipulative and physiological therapeutics* 2006;29:374-7.
29. Qvistgaard E, Rasmussen J, Laetgaard J, et al. Intra-observer and inter-observer agreement of the manual examination of the lumbar spine in chronic low-back pain. *European Spine Journal* 2007;16:277-82.
30. Rabin A, Shashua A, Pizem K, et al. The Interrater Reliability of Physical Examination Tests That May Predict the Outcome or Suggest the Need for Lumbar Stabilization Exercises. *Journal of Orthopaedic & Sports Physical Therapy* 2013;43:83-90.
31. Ravenna MM, Hoffman SL, Van Dillen LR. Low Interrater Reliability of Examiners Performing the Prone Instability Test: A Clinical Test for Lumbar Shear Instability. *Archives of physical medicine and rehabilitation* 2011;92:913-9.
32. Roussel NA, Nijs J, Truijen S, et al. Low Back Pain: Clinimetric Properties of the Trendelenburg Test, Active Straight Leg Raise Test, and Breathing Pattern During Active Straight Leg Raising. *Journal of Manipulative and Physiological Therapeutics* 2007;30:270-8.
33. Schneider M, Erhard R, Brach J, et al. Spinal Palpation for Lumbar Segmental Mobility and Pain Provocation: An Interexaminer Reliability Study. *Journal of Manipulative and Physiological Therapeutics* 2008;31:465-73.
34. Sedaghat N, Latimer J, Maher C, et al. The reproducibility of a clinical grading system of motor control in patients with low back pain. *Journal of manipulative and physiological therapeutics* 2007;30:501-8.

428 35. Ferrari S, Manni T, Bonetti F, et al. A literature review of clinical tests for lumbar instability in
429 low back pain: validity and applicability in clinical practice. *Chiropractic & manual therapies*
430 2015;23:14.
431 36. Carlsson H, Rasmussen-Barr E. Clinical screening tests for assessing movement control in non-
432 specific low-back pain. A systematic review of intra- and inter-observer reliability studies. *Manual*
433 *therapy* 2013;18:103-10.

434 Figure 1: PRISMA 2009 Flow Diagram

Appendix 1: Search strategy

| Search details: | | | | | |
|---|---|---|---|---|---|
| LBP | AND | Reliability | AND | Clinical test | OR |
| | | | | Physical examination | OR |
| | | | | Muscle test | OR |
| | | | | Joint instability | OR |
| | | | | Instability | OR |
| | | | | Trunk instability | OR |
| | | | | Core stability | OR |
| | | | | Motor control | OR |
| Limits: none | | | | | |

Table 1: Reliability of identified clinical tests

| | | | Inter rater reliability | | Intra rater reliability | |
|---|---|---|---|---|---|---|
| Identified clinical tests | Author + year | Specification of test | K or ICC value | Conclusion | K or ICC value | Conclusion |
| **Aberrant movements*** | Hicks 2003[26] | | k = 0.60 (95%CI 0.47-0.73), PA = 84% | Moderate | | |
| | Fritz 2005[24] | | k = -0.07 (95%CI -0.45-0.31), PA = 87% | Poor | | |
| | Fritz 2006[23] | | k = 0.18 (95%CI -0.07-0.43), PA = 59% | Poor | | |
| | Alyazedi 2015[19] | | k = 0.79 (95%CI 0.39-1.19), PA = 98% | Good | | |
| | Rabin 2013[30] | | k = 0.64 (95%CI 0.32-0.90), PA = 83% | Good | | |
| | Biely 2014[20] | | k = 0.68 (95%CI 0.34-1.00), PA = 87% | Good | | |
| **Prone instability test*** | Hicks 2003[26] | | k = 0.87 (95%CI 0.80-0.94), PA = 91% | Very good | | |
| | Fritz 2005[24] | | k = 0.69 (95%CI 0.59-0.79), PA = 85% | Good | | |
| | Fritz 2006[23] | | k = 0.52 (95%CI 0.29 - 0.75), PA = 78% | Moderate | | |
| | Schneider 2008[33] | PIT test 1 (feet on the floor) | k = 0.54 (95%CI 0.27-0.81), PABAK = 0.58, PA = 79% | Moderate | | |
| | | PIT test 2 (feet off the floor) | k = 0.46 (95%CI 0.15-0.77), PABAK = 0.58, PA = 79% | Moderate | | |
| | Ravenna 2011[31] | PIT acknowledged | k = 0.10 (95%CI -0.27-0.47), PA = 63%, PABAK = 0.27(95%CI -0.08-0.61), bias index 0.03, prevalence index = 0.43 | Poor | | |
| | | PIT ignored | k = 0.04 (95%CI -0.34-0.42), PA = 73%, PABAK = 0.47 (95%CI 0.15-0.78), bias index 0.00, prevalence index 0.67 | Poor | | |
| | Alyazedi 2015[19] | | k = 0.71 (95%CI 0.45-0.98), PA = 90% | Good | | |
| | Rabin 2013[30] | | k = 0.67 (95%CI 0.29-1.00), PA = 90% | Good | | |
| **Beighton scale*** | Hicks 2003[26] | | k = 0.79 (95%CI 0.68-0.87) | Good | | |
| | Fritz 2005[24] | | ICC = 0.72 (95%CI 0.50-0.85) | Good | | |
| Painful arc | Hicks 2003[26] | Painful arc in flexion | k = 0.69 (95%CI 0;54-0.84), PA = 92% | Good | | |
| | | Painful arc on return | k = 0.61 (95%CI 0.44-0.78), PA = 90% | Good | | |
| Instability catch | Hicks 2003[26] | | k = 0.25 (95%CI -0.10-0.60), PA = 92% | Fair | | |
| | Biely 2014[20] | | k = 0.35 (95%CI 0.00-0.71), PA = 65% | Fair | | |
| Gower sign | Hicks 2003[26] | | k = 0.00 (95%CI -1.09-1.09), PA = 98% | Poor | | |
| Reversal of lumbopelvic rhythm during | Hicks 2003[26] | | k = 0.16 (95%CI-0.15-0.46), PA = 87% | Poor | | |
| | Biely 2014[20] | | k = 0.89 (95%CI 0.69-1.0), PA = 96% | Very good | | |
| Posterior shear test | Hicks 2003[26] | | k = 0.35 (95%CI 0.20-0.51), PA = 74% | Fair | | |
| | Fritz 2005[24] | | k = 0.27 (95%CI 0.14-0.41), PA = 64% | Fair | | |
| Palpation transversus abdominis | Sedaghat 2007[34] | crook lying/sitting | k = 0.07 (95%CI -0.26-0.40) | Poor | | |
| | | closed chain lower limb movement | k = 0.06 (95%CI -0.14-0.26) | Poor | | |
| | | trunk movement | k = 0.23 (9%CI 0.11-0.35) | Fair | | |
| | | fast upper limb movement | k = 0.30 (95%CI 0.18-0.42) | Fair | | |
| | | functional upper limb movement not able to perform | k = 0.30 (95%CI 0.12-0.48) | Fair | | |
| | | functional upper limb movement able to perform | k = 0.15 (95%CI -0.10-0.40) | Poor | | |
| | | for all palpations tests | k weighted = 0.29 | Fair | | |
| Palpation multifidus | Qvistgaard 2007[29] | perfect match | k = 0.12 | Poor | | |
| | | acceptable match | k = 0.48 | Moderate | | |
| Multifidis lift test | Hebert 2015[25] | L4-L5 no weight | k = 0.75 (95%CI 0.52-0.97), p<.001, PA = 86% | Good | | |
| | | L4-L5 with weight | k = 0.79 (95%CI 0.57-1.00), p<.001, PA = 90% | Good | | |
| | | L5-S1 no weight | k = 0.81 (95%CI 0.62-1.00), p<.001, PA = 91% | Very good | | |
| | | L5-S1 with weight | k = 0.80 (95%CI 0.59-1.00),p<.001, PA = 90% | Good | | |
| Single leg stance | Roussel 2007[32] | Left side | | | k =0.83, p<.001 | Very good |
| | | Right side | | | k =0.75, p<.001 | Good |
| | Luomajoki 2007[27] | Right side | k = 0.43 | Moderate | k = 0.67 | Good |
| | | Left side | k = 0.65 | Good | k = 0.84 | Very good |
| Passive lumbar extension test | Alyazedi 2015[19] | | k = 0.46 (95%CI 0.20-0.72), PA = 73% | Moderate | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Rabin 2013[30] | | k = 0.76 (95%CI 0.46-1.00), PA = 88% | Good | | |
| Active straight leg raise | Roussel 2007[32] | Breathing pattern left side | k weighted = 0.47, p<.01 | Moderate | k weighted = 0.70, p<.001 | Good |
| | | Breathing pattern right side | k weighted = 0.39, p.02 | Fair | k weighted = 0.71, p<.001 | Good |
| | Rabin 2013[30] | | k = 0.53 (95%CI 0.20-0.84), PA = 76% | Moderate | | |
| Sagittal deviation | Biely 2014[20] | | k = 0.68 (95%CI 0.34-1.00), PA = 87% | Good | | |
| Sitting knee extension test | Luomajoki 2007[27] | | k = 0.72 | Good | k = 0.95 | Very good |
| | Enoch 2011[22] | | ICC = 0.96 (95%CI 0.90-0.97) | Good | | |
| Hip extension test | Murphy 2006[28] | left leg | k = 0.72 | Good | | |
| | | right leg | k = 0.76 | Good | | |
| Joint position sense | Enoch 2011[22] | | ICC = 0.90 (95%CI 0.81-0.94) | Good | | |
| Sitting forward lean test | Enoch 2011[22] | | ICC = 0.96 (95%CI 0.92-0.98) | Good | | |
| Bent knee fall out test | Enoch 2011[22] | | ICC = 0.94 (95%CI 0.88-0.97 | Good | | |
| Waiters bow | Luomajoki 2007[27] | | k = 0.62 | Good | k = 0.88 | Very good |
| Rocking backward | Luomajoki 2007[27] | | k = 0.68 | Good | k = 0.51 | Fair |
| Pelvic tilt | Luomajoki 2007[27] | | k = 0.65 | Good | k = 0.80 | Good |
| Prone knee bend extension | Luomajoki 2007 [27] | | k = 0.47 | Moderate | k = 0.70 | Good |
| Prone knee bend rotation | Luomajoki 2007[27] | | k = 0.58 | Moderate | k = 0.78 | Good |
| Rocking forward | Luomajoki 2007[27] | | k = 0.57 | Moderate | k = 0.72 | Good |
| Lumbar extension load test | Rabin 2013[30] | | k = 0.47 (95%CI 0.14-0.78), PA = 74% | Moderate | | |
| Crook lying | Luomajoki 2007[27] | | k = 0.38 | Fair | k = 0.86 | Very good |
| Active hip abduction | Rabin 2013[30] | | k = -0.09 (95%CI -0.35-0.27), PA = 60% | Poor | | |
| Thoracolumbar dissociation | Elgueta 2014[21] | pre training | k weighted = 0.66 (95%CI 0.33-0.84) | Good | k weighted = 0.56 (95%CI0.17-0.86) | Fair |
| | | post training | k weighted = 0.51 (95%CI-0.03-0.79) | Moderate | k weighted = 0.78 (95%CI0.58-0.93) | Good |

*PIT: prone instability test, k: kappa, ICC: intra correlation coefficient, PA: percentage agreement, PABAK: prevalence and Bias Adjusted Kappa, 95%CI: 95% confidence interval, p: significance level*
**recommended clinical test for inter-rater reliability*

Table 2: Overview of the risk of bias assessment with the COSMIN checklist

| | Alyazedi 2015 | Biely 2014 | Elgueta 2014 | Enoch 2011 | Fritz 2005 | Fritz 2006 | Hebert 2015 | Hicks 2003 | Luomajoki 2007 | McCarthy 2007 | Murphy 2006 | Qvistgaard 2007 | Rabin 2013 | Ravenna 2011 | Roussel 2007 | Schneider 2008 | Sedaghat 2007 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentage of missing items described? | G | E | G | G | G | G | E | G | G | E | G | G | E | G | G | G | G |
| Description of handling missing items? | F | F | F | F | F | G | E | G | F | G | F | F | E | F | G | F | F |
| At least two measurements? | E | P | P | E | E | E | E | E | P | E | P | E | P | E | E | E | E |
| Administrations independent? | E | N/A | N/A | E | E | E | E | E | N/A | E | N/A | G | N/A | E | E | E | G |
| Time interval stated? | E | N/A | N/A | E | E | F | F | E | N/A | E | N/A | F | N/A | E | E | F | E |
| Were patients stable in-between measurements? | E | N/A | N/A | G | G | E | G | E | N/A | G | N/A | G | N/A | G | G | G | G |
| Time interval appropriate? | E | N/A | N/A | E | E | F | F | E | N/A | E | N/A | F | N/A | E | E | E | E |
| Conditions similar for both measurements? | E | N/A | N/A | E | G | G | G | E | N/A | G | N/A | G | N/A | G | E | G | G |
| Any important flaws in design? | E | E | E | E | E | E | E | E | E | E | E | E | E | E | E | E | E |
| ICC for continuous scores? | N/A | N/A | N/A | E | E | E | N/A | E | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | E |
| Kappa for dichotomous/ordinal/nominal scores? | E | E | E | N/A | E | E | E | E | E | E | E | E | E | E | E | E | E |
| Weighted kappa for ordinal scores? | N/A | N/A | E | N/A | N/A | E | N/A | E | N/A | E | N/A | N/A | N/A | N/A | N/A | N/A | E |
| Weighting scheme described for ordinal scores? | N/A | N/A | G | N/A | N/A | G | N/A | E | N/A | E | N/A | N/A | N/A | N/A | N/A | N/A | E |

*E: excellent; G: Good; F: Fair; P: Poor; N/A: not applicable*

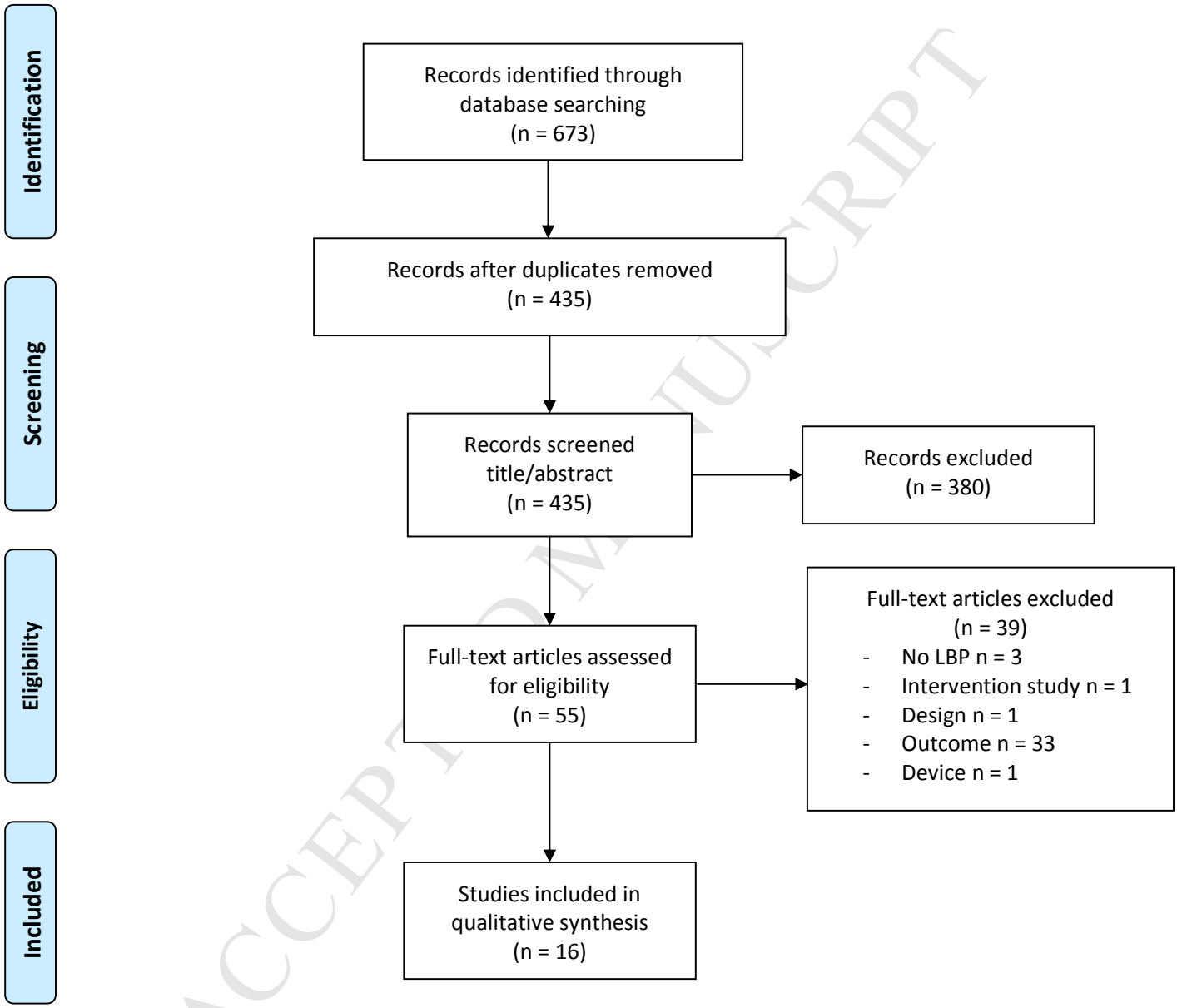
Table 3: Included articles and risk of bias assessment

| | Author | Reported clinical tests | Demographic values | In- and exclusion criteria |
|---|---|---|---|---|
| 1 | Alyazedi 2015[19] | aberrant movements<br>prone instability test<br>passive lumbar extension test | n=40 (25% women)<br>mean age 35 years (SD 12.22) | Inclusion: a new episode of LBP, experienced a similar episode of LBP before with the first episode occurring at least 3 months before the date of recruitment, currently LBP for at least 3 months<br>Exclusion: previous spine surgery, history of traumatic fracture, scoliosis greater than 20°, pregnancy, inability to actively flex and extend the spine adequately, medical red flags |
| 2 | Biely 2014[20] | aberrant movements<br>instability catch<br>reversal of lumbopelvic rhythm<br>sagittal deviation | n=31 (48% women)<br>mean age 44.4 years (SD 12.3)<br>mean ODI 28 (SD 14.1) | Inclusion: current episode of LBP that started within last 7 weeks and rated as 4 or greater on 11 point likert scale<br>Exclusion: age under 25 or above 65, clinical signs of systemic illness, definitive neurologic signs (weakness or numbness), previous spinal surgery, diagnosed osteoporosis, stenosis or inflammatory joint disease that interfered with upright stance, pregnancy |
| 3 | Elgueta 2014[21] | thoracolumbar dissociation | n=20 LBP (70% women)<br>mean age 31 years (SD 9) | Inclusion: LBP<br>Exclusion: suspected or confirmed final pathology of non-musculoskeletal origin, pregnancy, nerve root compression, spine surgery |
| 4 | Enoch 2011[22] | sitting knee extension<br>joint position sense<br>sitting forward lean test<br>bent knee fall out | n=25 LBP (56% women)<br>mean age 47 years (SD 12) | Inclusion: People with nonspecific LBP, age 18-85 years.<br>Exclusion: neurological disorders, rheumatologic disorders, acute pain in hip and legs, diabetes, cancer, inability to speak Danish |
| 5 | Fritz 2005[24] | aberrant movements<br>prone instability test<br>posterior shear test<br>beighton scale | n=49 (57% women)<br>mean age 39.2 years (SD 11.3)<br>mean ODI 20.4 (SD 13.3) | Inclusion: patient referred for flexion-extension Rx based on suspicion of instability, LBP without radiation into lower extremities, less than 60 years of age<br>Exclusion: contraindication for Rx, previous lumbar fusion surgery, inability to actively flex and extend spine |
| 6 | Fritz 2006[23] | aberrant movements<br>prone instability test | n=60 (48% women)<br>mean age 36.6 years (SD 10.5)<br>mean ODI 42.3% (SD 11.0) | Inclusion: 18-65 years, referred to physical therapy with primary complaint of LBP less than 90 days with or without referral into lower extremity, ODI higher than 25%<br>Exclusion: lateral shift, acute kyphotic deformity, symptoms could not be reproduced with lumbar ROM or palpation or when signs of nerve root compression were present, pregnancy, prior spine surgery |
| 7 | Hebert 2015[25] | multifidus lift test | n=32 (44% women)<br>mean age 31.38 years (SD 12.70)<br>mean ODI 30.31 (SD 11.00) | Inclusion: current LBP, 18-60 years, minimum ODI 20 points<br>Exclusion: history of lumbar spine surgery, signs or symptoms of lumbar radiculopathy, medial red flags, osteoporosis, recently treated for LBP with spinal manipulation or trunk muscle stabilization exercise |
| 8 | Hicks 2003[26] | aberrant movements<br>painful arc<br>instability catch<br>gower sign<br>reversal of lumbopelvic rhythm<br>prone instability test<br>posterior shear test<br>beighton scale | n=63 (60% women)<br>mean age 36 years (SD 10.3)<br>mean ODI 17.8 (SD 11.3) | Inclusion: LBP without radiation below the knee<br>Exclusion: pregnancy, acute fracture, tumor, infection. Previous lumbar fusion surgery |
| 9 | Luomajoki 2007[27] | single leg stance<br>sitting knee extension<br>waiters bow<br>rocking backward<br>rocking forward<br>pelvic tilt<br>prone knee bend extension<br>prone knee bend rotation<br>crook lying | n=27 LBP (67% women)<br>mean age 50.8 years (SD 6.2)<br>mean RMDQ 8.5 (SD 5.5) | Inclusion: Nonspecific LBP, patients who would be able to perform the tests very well (to prevent bias of results through too many incorrect test results)<br>Exclusion: serious pathologies (non-healed fractures, anomalies, tumors), acute trauma, acute LBP |
| 10 | Murphy 2006[28] | hip extension test | n=42 (74% women)<br>mean age 37.8 years (SD 14.4)<br>mean RMDQ 5.8 (SD 4.3) | Inclusion: > 7 weeks LBP, pain between buttock and T12 with or without leg pain, ability to perform hip extension in prone position without pain<br>Exclusion: visceral pathology, spinal infection, spinal fracture, pain when performing hip extension |
| 11 | Qvistgaard 2007[29] | palpation multifidus | n=60 (no further information about demographic information ) | Inclusion: LBP more than one month, 18-60 years<br>Exclusion: clinical signs of an acute disc herniation, inflammatory disease, ongoing insurance claim, significant medical disease, intellectual or language problems |

| | | | | |
|---|---|---|---|---|
| 12 | Rabin 2013[30] | aberrant movements<br>prone instability test<br>passive lumbar extension test<br>active straight leg raise<br>lumbar extension load test<br>active hip abduction | n=30 (50% women)<br>mean age 33.5 years (SD 8.0)<br>mean ODI 34.9 (SD 10.9) | Inclusion: age 18-60 years, main complaint of LBP and/or related leg symptoms<br>Exclusion: pregnancy, non-mechanical origin of symptoms, LBP due to fracture, osteoporosis, use of cortico's, rheumatoid arthritis, sings for nerve root compression |
| 13 | Ravenna 2011[31] | prone instability test | n=30 (43.3% women)<br>mean age 36.1 years (SD 11.8)<br>mean MODI 23.9 (SD 10.0) | Inclusion: recurrent LBP, age 18-60 years, current symptoms, but no acute flare-up<br>Exclusion: no BMI>30kg/m2, disc herniation, radiating symptoms below the knee, history of spinal surgery or fracture, spinal deformity, inflammatory condition, neurologic disease, pregnancy, primary hip problems |
| 14 | Roussel 2007[32] | single leg stance<br>active straight leg raise | n=36 (58% women)<br>mean age 37.4 years (SD 11.6) | Inclusion: age 18-65 years, chronic nonspecific LBP<br>Exclusion: specific LBP, spinal fracture history, severe degenerative change, severe scoliosis, osteoporosis, obesity, radicular signs, malignancies, metabolic or rheumatologic diseases |
| 15 | Schneider 2008[33] | prone instability test | n=39 (no further information about demographic information) | Inclusion: 18-65 years, low back pain history, ability to tolerate prone lying<br>Exclusion: history of prior lumbar surgery, stenosis, scoliosis greater than 20°, unstable spondylolisthesis, positive nerve root tension signs or radiculopathy, any red flags |
| 16 | Sedaghat 2007[34] | palpation transversus abdominis | n=34 (38% women)<br>mean age 42.7 years (SD 13.6) | Inclusion: nonspecific LBP within 6 months before testing<br>Exclusion: irritable low back pain condition that would become exacerbated with repeated testing |

*LBP: low back pain, ODI: oswestry disability index, RMDQ: roland morris disability questionnaire, MODI: modified oswestry disability index, ROM: range of motion, BMI: body mass index*

# PRISMA 2009 Flow Diagram






*From:* Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(6): e1000097. doi:10.1371/journal.pmed1000097

**For more information, visit www.prisma-statement.org.**

Text Box 1: Inclusion and exclusion criteria

| Inclusion criteria |
| --- |
| 1. Intra and/or inter reliability had to be conducted on a population with LBP<br>2. Clinical tests who identify lumbar segmental instability associated to motor control impairment<br>3. Only simple measurement devices such as a ruler, goniometer and laser pointer are allowed<br>4. adults (no children, adolescents or elderly)<br>5. Articles written in English |
| **Exclusion criteria** |
| 1. Patients with other pathologies than LBP<br>2. Patients with LBP not primarily originating from the lumbar spine (e.g. Sacroiliac / pelvic pathology, pregnancy related LBP and malignity)<br>3. High technological medical devices (e.g. MRI, 3D motion analysis system,…)<br>4. Reviews, practice guidelines, pilot studies, case reports, commentaries, editorials, letters, study protocols and books |