Manual Therapy xxx (2012) 1–8

Contents lists available at SciVerse ScienceDirect

Manual Therapy

journal homepage: www.elsevier.com/math

Review article

Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies

Hannah Carlsson*, Eva Rasmussen-Barr

Department of Neurobiology, Care Sciences and Society, Division of Physiotherapy, Karolinska Institutet, Alfred Nobels allé 23, 141 83 Huddinge, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 6 May 2012
Received in revised form
23 August 2012
Accepted 30 August 2012

Keywords:
Low-back pain
Reproducibility of results
Physical examination
Rehabilitation
Movement control
Screening

ABSTRACT

Background: Most people experience back pain at some point during their lives. Reports suggest that core stability interventions in subjects with non-specific low-back pain may increase function, thus decreasing pain. Reliable and validated clinical tests are required for implementing adequate rehabilitation and for evaluating such interventions.
Objective: This systematic literature overview seeks to assess the risk of bias and summarise the results of articles assessing the inter-and intra-observer reliability of clinical screening tests for movement control in subjects with non-specific low-back pain.
Method: A search was conducted in electronic search engines up until October 2011. The terms 'low-back pain', 'test', 'movement control', 'motor control' and 'physical examination' were defined and used. An appraisal tool (QAREL) was used to assess the risk of bias. Results of the studies were summarised.
Results: Eight studies were included and assessed. All examined inter-observer reliability and three also examined intra-observer reliability. The grading of the studies varied from five to nine positive items out of eleven possible. Inter-observer reliability ranged between poor and very good agreement. Intra-observer reliability ranged between moderate and very good agreement.
Conclusion: Most of the tests are presented in studies conducted with a high risk of bias. Their clinical implications can therefore not be suggested. Two tests, prone knee bend and one leg stance are assessed across studies with moderate and good reliability respectively and presented in studies conducted with a lower risk of bias. Their utilisation in clinical work may be recommended.

1. Background

Fifty-nine to eighty-four percent of the population exhibit symptoms of back pain at some point during their lives (Walker, 2000). Eighty to ninety percent will recover within six weeks irrespectively of coping strategies or treatment. However, up to 86% of these patients will suffer from recurrence within a year or more (Manchikanti et al., 2009; Stanton et al., 2011).

Up to 90% of low-back pain (LBP) symptoms are classified as non-specific, and the suggested cause is mechanical overuse and dysfunction of the spine and surrounding structures (Panjabi, 1996; Waddell, 2004; Panjabi, 2006). One hypothesis for the persistence and recurrence of non-specific LBP is impaired function of the deep core muscles proposed to be important for stability, robustness and movement control of the lower spine (Hodges and Moseley, 2003; Moseley and Hodges, 2006; Reeves et al., 2007). To date, core-stabilising and motor-control exercises are commonly used in the treatment of LBP. Previous studies indicate that such interventions may increase motor control, thus decreasing pain and improving function (Tsao and Hodges, 2007, 2008; Macedo et al., 2009; Rasmussen-Barr et al., 2009).

Clinimetrical measurements of core stability and movement control are not easy. Compensatory movement is proposed to be an indication that the central nervous system is unable to control the stabilising muscles of the spine optimally (Comerford and Mottram, 2001; Schabrun and Hodges, 2012). One theory, relative flexibility, may result in movements occurring through the least amount of effort in a kinetic chain (Sahrmann, 2002; O'Sullivan, 2005).

Functional and active movement control tests may be used to identify movement dysfunction in LBP. During such a tests, muscles have to be co-activated in an integrated pattern to retain control, i.e. the patient's ability to isometrically contract muscles concurrently, producing movement in another segment. This is referred to as dissociation and is applicable in the clinic to determine compensatory movement (Woolsey et al., 1988; Sahrmann, 2002). It has been

* Corresponding author. Tel.: +46 736988396.
E-mail address: hannah.m.carlsson@gmail.com (H. Carlsson).

1356-689X/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.math.2012.08.004

suggested that these tests should be used for predictive or preventive purposes, or else when evaluating core stability interventions (Murphy et al., 2006; Luomajoki et al., 2007; Roussel et al., 2007; Tidstrand and Horneij, 2009). Renewed interest in movement screening has prompted a need for valid, reliable and objective clinical tests.

Currently, no systematic literature overview of the reliability of screening tests for movement control of the lumbar spine seems to have been done.

2. Objective

This study seeks to summarise results and systematically assess the risk of bias in intra- and inter-observer reliability studies evaluating objective and active screening tests for movement control in subjects with non-specific LBP.

3. Method

3.1. Data sources and search strategy

A systematic computerised search was conducted by the author (HC) through the search provider Pubmed and in the Chinal, Amed, Pedro and Swemed+ databases between January and March 2010. A second search was conducted at the end of October 2011. The following limitations were used for Pubmed: males/females, English/Swedish, young adults 19–24, adults 19–44, middle-aged 45–64, age: 65+ years and publication date 1998–2011. An initial search used the key term 'low-back pain' to cover a wide range of publications, thus indicating the amount of publications in this area. Additional key terms used were: 'test', 'movement control', 'motor control' and 'physical examination'. To reduce the search bias, a final search strategy was conducted using Medical Subject Headings (MeSH) terms from the articles already included. These terms were: '*Disability Evaluation', 'Exercise Test', 'Low-Back Pain/*diagnosis', 'Lumbar Vertebrae', 'Movement/physiology', 'Physical Examination/*methods', 'Observer Variation', 'Postural Balance/physiology' and 'Reproducibility of Results'. References from the articles studied were also included in the search.

3.2. Study selection

Eligible articles were intra- and/or inter-observer reliability studies of objective and active movement control tests published before November 2011, written in either Swedish or English. The tests included were proposed for examining dysfunctional movement control in the trunk and pelvis in subjects with non-specific LBP lasting more than six weeks, as well as healthy subjects. Eligible observers were certified practitioners, i.e. physiotherapists or chiropractors. Only simple measurement devices were used, i.e. stopwatch, metronome, ruler and pressure biofeedback.

An eligibility assessment was conducted by screening relevant titles and abstracts. The two authors (ERB and HC) independently performed the assessment in an un-blinded and standardised manner according to the eligibility criteria set out. If disagreement arose, this was discussed until a consensus was achieved. The study selection procedure is presented in Fig. 1.

3.3. Reliability

Reliability refers to the agreement between several results of one test or the amount of measurement error acceptable when using the test in clinical practice (Atkinson and Nevill, 1998). Results from one subject examined by the same observer (intra-observer or test–retest reliability) or by several observers examining the same subject should stay consistent (inter-observer reliability) (Baumgarter, 1989). 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 (Atkinson and Nevill, 1998). Nominal and ordinal unpaired data are often analysed with Cohen's or the weighted kappa (κ) coefficient (Rankin and Stokes, 1998). κ values may vary between −1 and 1. Agreement can be interpreted as $\kappa < 0.20$ = poor, κ:0.21–0.40 = fair, κ:0.41–0.60 = moderate, κ:0.61–0.80 = good/substantial, κ:0.81–1.0 = very good/excellent (Landis and Koch, 1977). Continuous data is often analysed with intra-class correlation coefficients (ICC) (Rankin and Stokes, 1998).

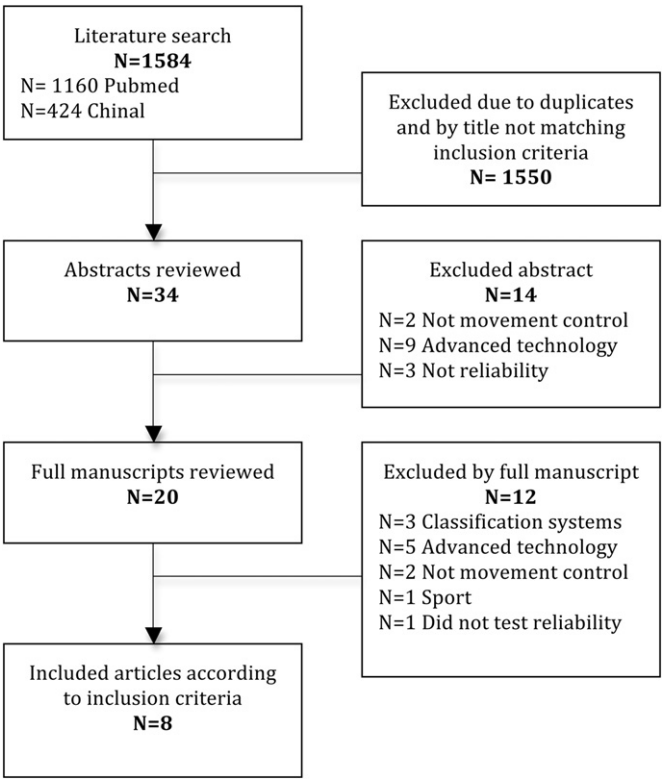

Fig. 1. Flowchart of literature search and studies included.

3.4. Risk of bias assessment

The risk of bias assessment tool used was the Quality Appraisal of Reliability Studies (QAREL) checklist. The QAREL checklist has previously been tested for validity (Lucas et al., 2010) and has been used in a systematic review aimed at reliability of diagnostic tests (Simopoulos et al., 2012). The checklist includes 11 evaluation items graded 'yes', 'no', 'unclear' or 'not applicable', specifying the methodological components assessed separately rather than giving an overall score. Initially, both authors (ERB, HC) graded each study independently and subsequently discussed the results until a completely unanimous grade was allocated to each item.

3.5. Data extraction

One author (HC), extracted information and summarised the results. Information was extracted from each study on (1) characteristics of trial participants (sex, age and pain duration), (2) characteristics of observers (experience, degree and amount of training of protocol), (3) characteristics of settings (private physiotherapy practice, out-patient physiotherapy centre or medical clinic), (4)

Please cite this article in press as: Carlsson H, Rasmussen-Barr E, Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies, Manual Therapy (2012), http://dx.doi.org/10.1016/j.math.2012.08.004

type of intervention (intra-observer reliability, inter-observer reliability and the number and description of clinical screening tests) and (5) data on reliability (κ and ICC values). Information extracted is presented in tables emphasising the risk of bias, the studies design and results. Narrative summaries of the results are also displayed.

3.6. Data analysis

In view of the studies' objectives, heterogeneous populations and the tests and outcomes investigated, descriptive statistics were used to summarise findings across all reliability studies. The risk of bias conclusion was based on the number of positive items (0–11) as well as on the best evidence synthesis. A high risk of bias was considered if $\leq$7 items were positive and a low risk was considered if $\geq$8 items were positive. Seven positive items were considered a moderate risk of bias (Simopoulos et al., 2012). The computer software used for this review included Microsoft Office 2008 (Microsoft Corp., Redmond, Washington), and Endnote (Version X4, Thomson Reuters, New York, NY).

4. Results

Relevant articles matching the inclusion criteria were found in Pubmed and Cinahl. One study (Davis et al., 2011) was retrieved through the MeSH terms, 'Low-back Pain', and 'Observer Variation'. In total, the two authors read 34 abstracts and 20 articles. Finally, eight articles met the study objectives as shown in Fig. 1 and are summarised in Table 1. Both authors assessed the risk of bias.

A variety of screening tests were assessed in the included studies and were conducted with various methods. Six of the nineteen tests were assessed across studies (Table 2) (Van Dillen et al., 1998; Luomajoki et al., 2007; Roussel et al., 2007; Tidstrand and Horneij, 2009; Enoch et al., 2011).

4.1. Risk of bias assessment

The risks of bias of the included studies, according to the QAREL list are presented in Table 2. Most of the tests were assessed with a high risk of bias. The grading of the studies varied from 5 to 9 positive items out of 11 possible. Four studies were considered to be of a high risk of bias (Van Dillen et al., 1998; Murphy et al., 2006; Sedaghat et al., 2007; Enoch et al., 2011). One was considered to be of a moderate risk (Tidstrand and Horneij, 2009) and three were considered to be of low risk of bias (Luomajoki et al., 2007; Roussel et al., 2007; Davis et al., 2011).

4.2. Summary of results

Nineteen screening tests and one test battery for movement control were assessed for reliability in the eight studies. All were examined for inter-observer reliability. Overall, the tests showed κ values between 0.06 and 0.94 and ICC values between 0.30 and 0.98, indicating poor to very good agreement (Table 3).

Three of the included studies (Luomajoki et al., 2007; Roussel et al., 2007; Davis et al., 2011), showing a low risk of bias, presented κ or ICC values ranging between 0.38 and 0.96, indicating fair-to-very good reliability. The hip abduction test (Davis et al., 2011) showed very good reliability (ICC = 0.96) and the crook lying hip abduction/lateral rotation (Luomajoki et al., 2007) test showed fair reliability (κ = 0.38). The one leg stance test, sitting knee extension test, prone knee flexion test, rocking forward test, rocking backward test, pelvic tilt test (Luomajoki et al., 2007) and the active straight leg raising test (Roussel et al., 2007) showed moderate to good reliability (κ = 0.43–0.72).

The four studies that were considered to be of a high risk of bias (Van Dillen et al., 1998; Murphy et al., 2006; Sedaghat et al., 2007; Enoch et al., 2011) presented inter-observer reliability between κ = 0.06–0.76 and ICC = 0.90–0.98, indicating poor-to-very good reliability.

Ten tests were examined for intra-observer reliability (Table 4). Overall, these ten showed κ values between 0.51 and 0.95, varying between moderate and very good agreement and ICC of 0.90–0.96, indicating very good agreement. These tests were included in studies with a low risk of bias (Luomajoki et al., 2007; Roussel et al., 2007; Davis et al., 2011).

Four of the studies (Murphy et al., 2006; Tidstrand and Horneij, 2009; Davis et al., 2011; Enoch et al., 2011) presented an average agreement of ICC/$\kappa \geq$ 0.70 indicating good-to-very-good agreement. Of these, only Davis et al. (2011) was assessed with a low risk of bias (Table 3).

One of the eight studies (Davis et al., 2011) used both quantitative (scale) and dichotomous (positive/negative) outcome variables. Two studies (Sedaghat et al., 2007; Enoch et al., 2011), used only quantitative outcome variables, while one presented results as cm and mm Hg (Enoch et al., 2011). The remaining studies presented dichotomous outcome variables (Table 1).

Only two tests were rated across studies with the same ratings. The prone knee flexion test was assessed for inter-observer reliability and was rated with moderate agreement (Van Dillen et al., 1998; Luomajoki et al., 2007). The one-leg stance test was examined for intra-observer reliability and rated with very good agreement (Luomajoki et al., 2007; Roussel et al., 2007). Both these tests were presented in studies with a low risk of bias.

5. Discussion

The present study is, to our knowledge, the first systematic review assessing risk of bias and summarising the results of intra- and inter-observer reliability of movement-screening tests in subjects with non-specific low-back pain. The included studies presented a variety of movement screening tests and were conducted with various methods. Only a few tests were assessed across studies and the majority of studies included were conducted with a high risk of bias.

All the included studies (Van Dillen et al., 1998; Murphy et al., 2006; Luomajoki et al., 2007; Roussel et al., 2007; Sedaghat et al., 2007; Tidstrand and Horneij, 2009; Davis et al., 2011; Enoch et al., 2011) examined inter-observer reliability and three also examined intra-observer reliability (Luomajoki et al., 2007; Roussel et al., 2007; Davis et al., 2011). Inter-observer reliability ranged between poor and very good agreement and intra-observer reliability ranged between moderate and very good agreement.

The tests examined for inter-observer and intra-observer reliability across studies did not show similar agreement except for prone-knee flexion (Van Dillen et al., 1998; Luomajoki et al., 2007) and one-leg stance (Luomajoki et al., 2007; Roussel et al., 2007) which gave moderate and very-good agreement respectively. One reason may be that the studies examined one factor for each test, such as pelvic tilt or lateral shift of the trunk, with a dichotomous outcome. In most of the tests, several factors are included in a positive or negative outcome such as deviating spine, pelvis deviation, compensatory movement from limbs and changes in starting position (Tidstrand and Horneij, 2009). Investigating a single factor in a test instead of many possible factors may simplify the decision for the observer, thus resulting in more realistic κ values. For example, Enoch and colleagues (Enoch et al., 2011) investigated one factor for each investigated test, such as one measure of spinal deviation, hence resulting in very good agreement

Please cite this article in press as: Carlsson H, Rasmussen-Barr E, Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies, Manual Therapy (2012), http://dx.doi.org/10.1016/j.math.2012.08.004

Table 1
Description and results of studies included.

Study	Objective	Subjects	Raters	Method/tests	Outcome variables	Results
Van Dillen et al., 1998	Describe results of inter-observer reliability of physical examination items used to categorize people with LBP. Items used to examine movement control and pain provocation.	$N = 138$. 95 subjects with LBP, of whom 54 women. Forty-three without LBP, of whom 26 women. Mean age (SD) LBP 44.07 (13.29), healthy 39.38 (13.05). Eighteen subjects LBP > 7 days, 71 subjects LBP > 7 weeks. Recruited from an outpatient practice, clinical facilities, advertisements, friends and family.	Five orthopaedic physiotherapists with 5–35 years' work experience. Each studied a training manual and took a written exam. Training continued with video and individual 45-min discussion with first author.	Forward bending and return. Sitting knee-extension. Supine hip abduction and hip rotation. Prone knee-flex and hip rotation Quadruped arm lifting and rocking backwards. Examined symptom provocation, alignment and movement. Subjects rated pain from back pre-tests and rated again during test; same, decrease, increase. Subjects examined by two raters and one observer = three test pairs. Raters blinded.	Dichotomous outcome variables yes/no and ordinal data. Kappa (κ) + percentage agreement for dichotomous outcome. Weighted kappa (κ) and weighted percentage agreement (3 categories) for ordinal data. Three levels of agreement: 1.0: maximum agreement 0.5:partial agreement 0.0: maximum disagreement.	Movement control: $\kappa = 0.26–0.65$. Average $\kappa = 0.51$ Symptom behaviour: $\kappa = 0.89.1.00$. Average $\kappa = 0.96$
Murphy et al., 2006	Examine inter-observer reliability of hip-extension test to identify dynamic instability of e.g. lower back in LBP.	$N = 42$, of whom 31 women. Age 19–60. LBP > 7 weeks. Recruited from spine centre.	Two chiropractors, one with 13 years of work experience and one with less than 1 year Training session 1 h.	Hip extension prone, left and right hip, no more than three repetitions/side. Raters examined subject at the same time. Outcome blinded between raters.	Dichotomous outcome variables positive/negative. Kappa coefficient (κ) for inter-observer reliability. SD and range. No normally distributed data.	$\kappa = 0.72–0.76$ Average $\kappa = 0.74$
Luomajoki et al., 2007	Determine inter- and intra-observer reliability of 10 MCD tests of lumbar spine and whether experience affects reliability.	$N = 40$, of whom 26 women. 27 = LBP, 13 = healthy. Mean age (SD) 52.1 (5.5). Recruited from a private physiotherapy practice	Four physiotherapists. Two raters were clinical specialists in the field with over 25 years' experience and post-graduate degrees in manual therapy. Two raters had five years of work experience. Raters with five years of work experience participated in a three-day course in MCD pre-study.	Waiter's bow, sitting knee-extension, rocking backwards, rocking forwards, dorsal tilt of pelvis, prone active knee-flexion, one-leg stance, crook lying. Raters blinded to subjects' diagnosis and each others' test results. Oral and visual to subjects. Subjects filmed and raters analysed each video once. One experienced and one inexperienced rater analysed each film again after two weeks to test intra-observer reliability.	Dichotomous outcomes variables positive/negative. Kappa coefficient (κ) for inter- and intra-observer reliability Percentage agreement. CI (95%)	Inter-observer: $\kappa = 0.38–0.72$. Average $\kappa = 0.59$ Intra observer: $\kappa = 0.51–0.95$. Average $\kappa = 0.78$
(Roussel et al., 2007)	Examine test–retest reliability and internal consistency of two clinical tests. Evaluated inter-observer reliability of breathing pattern during ASLR.	$N = 36$ of whom 21 women LBP > 3 months. Age 21–62. Recruited from a private physiotherapy practice	Two physiotherapists, one with completed master's degree and one with four years' work experience. Both trained to perform tests pre-study two hours × 2. Examined 10 subjects for pilot.	Trendelenburg, ASLR and breathing pattern during ASLR. Subjects examined by rater 1, rested 10 min, filled out a form and then examined by rater 2. Raters blinded to subject's health history and each other's ratings. Test order randomised.	Dichotomous outcome variables positive/negative. Weighted kappa (κ) for inter-observer reliability. Spearman's correlations coefficient for correlation between tests. Sign. Level: 0.05	Inter-observer: $\kappa = 0.38–0.47$. Average $\kappa = 0.43$ Intra-observer: $\kappa = 0.70–0.83$. Average $\kappa = 0.75$
Sedaghat et al., 2007	Examine inter-observer reproducibility of Wisbey-Roth grading system.	$N = 30$, of whom 15 women. Current (68%) or previous history of LBP. Age (SD) 42.7 (13.6).	Five physiotherapists; four women and one man. Work experience between 4 and 20 years. All had experience in examining and treating	Wisbey-Roth: grading system score of 0–5 depending on subjects' ability to activate TrA and LM through palpation and pelvic floor in functional	Dichotomous outcome variables positive/ negative and outcome for 0–5. SEM for agreement.	$\kappa = 0.06–0.30$. Average: $\kappa = 0.18$ ICC across observers = 0.30 ANOVA: $p = 0.64–0.95$

		Recruited from a private physiotherapy practice.	LBP. Three training sessions pre-study with Wisbey-Roth test protocol.	positions. Positions in crook-lying, sitting, standing during arm flexion, standing repeated arm flexion, left trunk rotation and flexion. Raters blinded to subjects' diagnosis and each others' results. Raters examined subjects individually, subjects rested 5 min between examinations.	Weighted kappa (κ) and ICC (2,1 model) for reliability 0–5 and unweighted kappa (κ) for dichotomous outcome ANOVA for order of testing or rater. Significance level 0.05. CI (95%).	
Tidstrand and Horneij, 2009	Examine inter-observer reliability of three functional tests of muscular coordination of the lumbar spine in LBP	$N = 19$, of whom nine women. LBP = 13. Arm/shoulder pain = 6. Age 42 ± 12. Recruited from private physiotherapy centre.	Two experienced raters trained in orthopaedic manual therapy and McKenzie. Both with over five years of experience treating patients with lumbar instability. Raters coordinated evaluation of rating on 10 subjects pre-study (not included in study)	Single-limb stance, sitting on a bobath ball and unilateral pelvic lift. Examined every subject at the same time, evaluated individually. Each subject completed 6 tests, 3 left and 3 right hand side. LBP subjects rated pain (VAS) before and after tests. Exclusion over 70 mm (VAS)	Rated dichotomously positive/negative. Cohen's kappa (κ) coefficient for inter observer reliability and percentage agreement.	$\kappa = 0.54–94$ Average $\kappa = 0.77$
Davis et al., 2011	Determine intra- and inter observer reliability of the active hip abduction test	$N = 64$. Asymptomatic. Age 22–69. Recruited from university population	16 (9 male) holding physical therapy licences and active engagement in clinical practice (0–15 years). Mailed: 14 min tutorial DVD and scoring criteria,	Examine the active hip abduction test. Filmed above and in frontal plane. Five videos examined from each scoring category = 20 video clips. Video clips, sheet detailing, demographic survey and scoring sheets mailed. Individual scoring.	Dichotomous (positive/negative) and ordinal data 0–3 points. ICC for reliability	Inter-observer: ICC = 0.59–0.97. Average ICC = 0.81 Intra-observer: ICC = 0.53–0.96. Average ICC = 0.74
Enoch et al., 2011	Inter-observer reproducibility of tests for lumbar motor control in mixed population with and without LBP	$N = 40$ of whom 26 women. 15 asymptomatic. Age 20–82. Recruited from different private physiotherapy practices.	Two with 20 years of clinical experience. Teachers in Danish Manual Therapy Society. Tested first on 10 LBP subjects.	Joint position sense (cm), sitting forward lean (cm), sitting knee-extension (cm), bent-knee fallout (cm), leg lowering (mm Hg). Examined independently in random order, provided necessary instructions. Half started with examiner A.	Continuous data ICC for reliability type 2.1 and Bland and Altman's limits of agreement	ICC = 0.90–0.98. Average ICC = 0.95

N = number, ICC = intra-class correlation coefficient, LBP = low back pain, VAS = visual analogue scale, MCD = movement control dysfunction, SD = standard deviation, CI = confidence interval, ASLR = active straight-leg raise, TrA = transversus abdominis, LM = lumbar multifidus, SEM = standard error of measurement.

Table 2
Overview of risk of bias assessment utilized with Quality Appraisal of Reliability Studies (QAREL) checklist (Lucas et al., 2010).

Parameter	Van Dillen et al., 1998	Murphy et al., 2006	Luomajoki et al., 2007	Roussel et al., 2007	Sedaghat et al., 2007	Tidstrand and Horneij, 2009	Davis et al., 2011	Enoch et al., 2011
1.	Y	Y	Y	Y	Y	Y	Y	Y
2.	Y	Y	Y	Y	Y	Y	Y	Y
3.	Y	Y	Y	Y	Y	Y	Y	Y
4.	N/A	N/A	U	U	N/A	N/A	Y	N/A
5.	N	Y	Y	Y	Y	Y	Y	U
6.	U	U	Y	Y	Y	U	Y	U
7.	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
8.	Y	N/A	U	Y	U	Y	N	Y
9.	N/A	N/A	Y	U	N/A	N/A	Y	N/A
10.	U	Y	Y	Y	U	Y	Y	Y
11.	Y	Y	Y	Y	Y	Y	Y	Y
TOTAL	**5**	**6**	**8**	**8**	**6**	**7**	**9**	**6**
Risk of bias	**High**	**High**	**Low**	**Low**	**High**	**Moderate**	**Low**	**High**

Y = yes, N=no, N/A = not applicable, U = unclear. 1. Was the test evaluated in a sample of subjects who were representative of those to whom the authors intended the results to be applied? 2. Was the test performed by raters who were representative of those to whom the authors intended the results to be applied? 3.Were raters blinded to the findings of other raters during the study? 4.Were raters blinded to their own prior findings of the test under evaluation? 5. Were raters blinded to the subjects' disease status or the results of the accepted reference standard for the target disorder (or variable) being evaluated? 6. Were raters blinded to clinical information that was not intended to form part of the study design or testing procedure? 7. Were raters blinded to additional cues that are not part of the test? 8. Was the order of examination varied? 9. Was the stability (or theoretical stability) of the variable being measured taken into account when determining the suitability of the time interval among repeated measures? 10. Was the test applied correctly and interpreted appropriately? 11. Were appropriate statistical measures of agreement used?

presenting quantitative outcome. It may be discussed whether it is more appropriate in clinical tests to use quantitative outcome variables or dichotomous ones. The hip abduction test showed almost identical results both when dichotomised and when assessed with a scale (Davis et al., 2011). Quantifying test results may enable the observer to obtain more information, which may be more useful for diagnosis and evaluation in the clinic.

In addition to quantitative outcome variables, precise measurements seems more reliable than tactile or visual methods (Van Dillen et al., 1998). Sedaghat et al. (2007) used a 0–5 point scale with a tactile approach, resulting in poor agreement. Here, reliability was assessed for the complete protocol compared to the remaining studies including trials assessed by observation only and for every test. Therefore, it may be discussed whether the latter

Table 3
Results of inter-observer reliability.

Test	Author	Reliability	Percentage agreement (%)	Conclusion
One leg stance	Luomajoki et al., 2007	κ = R:0.43 L:0.65	P1:R/L 88	Moderate–good
	Tidstrand and Horneij, 2009	κ = R:1.00 L:0.88	P2:R97.5 L92.5 R:100 L:95	Very good
Forward bending/	Van Dillen et al., 1998	κ = F:1.00 E:1.00 RF:0.51,	F:100 E:100 RF:76,	Moderate–very good
Waiter's bow	Luomajoki et al., 2007	HER:0.48 LER:0.54 κ = 0.62	HER:91 LER:92 P1:85.7 P2:75	Good
Sitting knee extension	Van Dillen et al., 1998	κ = 0.58	86	Moderate
	Luomajoki et al., 2007	κ = 0.72	P1:90.4 P2:95	Good
	Enoch et al., 2011	ICC = 0.95		Very good
Prone knee flexion	Van Dillen et al., 1998	κ = Ext:0.76 Rot:0.43	Ext/rot:90	Moderate–good
	Luomajoki et al., 2007	κ = Ext:0.47 Rot:0.58	P1:ext 97.6/rot 90.5 P2:ext/rot 87.5	Moderate
Crook lying hip abduction/	Van Dillen et al., 1998	κ = 0.60	88	Moderate
lateral rotation	Luomajoki et al., 2007	κ = 0.38	P1:78.6 P2:65	Fair
	Enoch et al., 2011	κ = 0.94		Very good
Rocking back	Van Dillen et al., 1998	κ = RF:0.78 Rot:0.51	RF:95 Rot:82	Moderate–good
	Luomajoki et al., 2007	κ = 0.57	P1:88 P2:90	Moderate
Rocking forward	Luomajoki et al., 2007	κ = 0.68	P1:92.8 P2:92.5	Good
Quadruped arm lift	Van Dillen et al., 1998	κ = 0.21	55	Fair
Prone hip extension	Murphy et al., 2006	L:0.72 R:0.76		Good
Sidelying hip abduction	Davis et al., 2011	ICC 0–3 point scale:0.97 Pos/neg 0.96		Very good
Active straight leg raise (ASLR) with assessed breathing pattern	Roussel et al., 2007	κ = R:0.38 L:0.47		Fair-to-moderate
Trunk lateral flexion	Van Dillen et al., 1998	κ = 0.26	65	Fair
Prone hip rotation	Van Dillen et al., 1998	κ = RF 0.56 Rot:0.52	RF:83 Rot:74	Moderate
Pelvic tilt	Luomajoki et al., 2007	κ = 0,65	P1:80 P2:92.5	Good
Sitting forward lean	Enoch et al., 2011	ICC = 0.96		Very good
Unilateral pelvic lift	Tidstrand and Horneij, 2009	κ = R:0.61 L:0.47	R:79 L:94	Moderate
Sitting on a bobath ball	Tidstrand and Horneij, 2009	κ = R:0.79 L:0.88	R:89 L:95	Very good
Joint position sense	Enoch et al., 2011	ICC = 0.90		Very good
Leg lowering	Enoch et al., 2011	ICC = 0.98		Very good
Wisbey-Roth grading system	Sedaghat et al., 2007	κ = −0.01–0.56 ICC = 0.30		Poor–fair

Percentage agreement for Davis et al. (2011) is not included due to this was presented for every scoring point and not for complete test. R = right, L = left, P1 = pair 1, P2 = pair 2, RF = relative flexibility, Rot = rotation, HER = hip extension with return LER = lumbar extension with return.

Please cite this article in press as: Carlsson H, Rasmussen-Barr E, Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies, Manual Therapy (2012), http://dx.doi.org/10.1016/j.math.2012.08.004

Table 4
Results of intra-observer reliability.

Test	Author	Reliability	Percentage agreement (%)	Conclusion
Sitting knee extension	Luomajoki et al., 2007	$\kappa = 0.95$	100	Very good
Crook lying hip abduction/lateral rotation	Luomajoki et al., 2007	$\kappa = 0.86$	97.5	Very good
Prone knee flexion	Luomajoki et al., 2007	κ = Ext:0.70 Rot:0.78	Ext:92.1 rot:O1 92.5 O2 100	Good
Rocking forward	Luomajoki et al., 2007	$\kappa = 0.51$	O1:95 O2:100	Moderate
Rocking back	Luomajoki et al., 2007	$\kappa = 0.72$	97.5	Good
Trunk flexion (Waiter's bow)	Luomajoki et al., 2007	$\kappa = 0.88$	O1:97.5 O2:100	Very good
Pelvic tilt	Luomajoki et al., 2007	$\kappa = 0.80$	95	Good
One leg stance/Trendendelenburg	Luomajoki et al., 2007	κ = R:0.67 L:0.84	O1:R92.5 L87.5	Good–very good
	Roussel et al., 2007	κ = R:0.75 L:0.83	O2:R/L100	Good–very good
Side lying hip abduction	Davis et al., 2011	ICC = 0.53–0.93		Moderate–very good
Active straight leg raise (ASLR) with assessed breathing pattern	Roussel et al., 2007	κ = R:0.70 L:0.71		Good

Ext = extension, Rot = rotation, L = left, R = right, O1 = observer 1, O2 = observer 2.

study (Sedaghat et al., 2007) should have been included in the present review as it differs the most from the others.

All the studies presented adequate information on observer demographics, reducing the risk of bias, as reliability results seem to depend on the observers. The sitting knee-extension, crook lying/bent-knee fallout, forward bending/waiters' bow, one-leg stance, rocking backwards and prone knee-flexion tests were analysed across studies (Van Dillen et al., 1998; Luomajoki et al., 2007; Enoch et al., 2011) (Table 3) but showed diverse agreement. Two of the studies (Tidstrand and Horneij, 2009; Enoch et al., 2011) only used experienced observers and presented results with good-to-very-good agreement. The rest showed a wide range of clinical experience between observers, resulting in a wide range of test reliability. A reason for good test reliability may be the observers' level of experience, suggesting that experienced observers demonstrate better reliability than inexperienced (Sahrmann, 2002; Dankaerts et al., 2006; Luomajoki et al., 2007).

Apart from experience, observer training may play an important role for the reliability of a test. The studies examined single tests and complex test batteries, differing numbers of observers and a wide range of observer training. In general, a single test (Murphy et al., 2006; Davis et al., 2011) showed a higher level of agreement than did more complex test batteries (Van Dillen et al., 1998; Roussel et al., 2007; Sedaghat et al., 2007). Varying training quality between the studies may have contributed to the diverse results. However, advanced training did not seem to influence the results.

As κ calculations depend on even outcomes, diverging subject categories may have influenced the diverse reliability results of included tests. Even outcomes can be difficult to attain but the probability increases with observers are blinded to two equally large groups, predicted for even outcomes (Banerjee et al., 1999). Studies (Luomajoki et al., 2007; Tidstrand and Horneij, 2009; Enoch et al., 2011) including both healthy and LBP subject groups showed slightly higher agreement (moderate-to-very-good) in general compared to those (Murphy et al., 2006; Roussel et al., 2007; Sedaghat et al., 2007) with only one group of subjects (poor-to-very-good).

The QAREL checklist (Lucas et al., 2010) was used to grade the risk of bias in the studies included. This tool was designed to rate studies investigating reliability in diagnostic tests. Several of the studies showed a rather high risk of bias according to the tool, therefore not presenting trustworthy reliability of the tests in spite of presenting high levels of reliability. This implies that a higher κ value is not always equivalent with a reliable and trustworthy test.

Our results indicate that studies with the lowest risk of bias, according to the QAREL list, presented data on both intra- and inter-observer reliability explaining test protocols and procedures satisfactorily (Luomajoki et al., 2007; Roussel et al., 2007; Davis et al., 2011). Overall, they presented information on blinding procedures well, such as blinding towards other observers, to disease/symptoms and to clinical information. However, these three studies, except for Davis and colleagues (Davis et al., 2011), did not present the highest average reliability, thus resulting in a rather low level of inter-observer reliability but high intra-observer reliability.

This review was conducted according to the PRISMA statement for reporting systematic reviews (Liberati et al., 2009). However, search terms and the inclusion criterions narrowed the search result, possibly missing relevant articles. Further, only studies in English or Swedish were included also resulting in a potential bias. A more open search strategy may have identified more studies, giving a wider range of methods. The present search terms were however considered relevant according to the objectives. The small number of studies included, in conjunction with their heterogeneity, precluded a quantitative analysis such as a meta-analysis.

Movement screening tests are to date commonly used in the clinic. Therefore, more studies with similar methods are needed to be able to generalise the reliability of the test results. If more movement screening tests demonstrate satisfactory agreement, such results may add to their feasibility in clinical practice.

5.1. Conclusion

The results of this literature review indicate that most of the screening tests included, designed for assessing impaired motor control in subjects with LBP, are presented in studies conducted with a high risk of bias. The clinical implication of these tests may at this stage not be suggested. However, two of the tests, prone knee bend and one leg stance are assessed across studies with moderate and good reliability respectively and subsequently, presented in studies conducted a lower risk of bias. Therefore, their utilisation in clinical work may be recommended. However, future research is important to evaluate clinical screening tests in more thorough methodological studies enabling the utilisation of trustworthy screening tests in the clinic.

Acknowledgements

The authors wish to thank Tim Crosfield for English proof reading and Björn Äng for valuable advice. This study was financially supported by Minnesfonden with a Master student grant.

References

Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998;26(4):217–38.
Banerjee M, Capozzoli M, Mcsweeney L, Sinha D. Beyond kappa: a review of interrater agreement measures. Can J Stat 1999;27(1):3–23.

Please cite this article in press as: Carlsson H, Rasmussen-Barr E, Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies, Manual Therapy (2012), http://dx.doi.org/10.1016/j.math.2012.08.004

Baumgarter TA. Norm-referenced measurement: reliability. In: J.S.M., M.W.T., editors. Measurement concepts in physical education and exercise science. Champaign: Illinois; 1989.
Comerford MJ, Mottram SL. Movement and stability dysfunction – contemporary developments. Man Ther 2001;6(1):15–26.
Dankaerts W, O'sullivan PB, Straker LM, Burnett AF, Skouen JS. The inter-examiner reliability of a classification method for non-specific chronic low back pain patients with motor control impairment. Man Ther 2006;11(1):28–39.
Davis AM, Bridge P, Miller J, Nelson-Wong E. Interrater and intrarater reliability of the active hip abduction test. J Orthop Sports Phys Ther 2011;41(12):953–60.
Enoch F, Kjaer P, Elkjaer A, Remvig L, Juul-Kristensen B. Inter-examiner reproducibility of tests for lumbar motor control. BMC Musculoskelet Disord 2011:12114.
Hodges PW, Moseley GL. Pain and motor control of the lumbopelvic region: effect and possible mechanisms. J Electromyogr Kinesiol 2003;13(4):361–70.
Landis RJ, Koch GG. The measurement of observer agreement for categorial data. Biometrics 1977;33(1):159–74.
Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. Br Med J 2009:339b2700.
Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol 2010;63(8):854–61.
Luomajoki H, Kool J, De Bruin ED, Airaksinen O. Reliability of movement control tests in the lumbar spine. BMC Musculoskelet Disord 2007;12(8):90–100.
Macedo LG, Maher CG, Latimer J. Motor control exercise for persistent, nonspecific low back pain: a systematic review. Phys Ther 2009;89(1):9–25.
Manchikanti L, Singh VDS, Cohen SP, Hirsch JA. Comprehensive review of epidemiology, scope, and impact of spinal pain. Pain Physician 2009;12(4):E35–70.
Moseley GL, Hodges PW. Reduced variability of postural strategy prevents normalization of motor changes induced by back pain: a risk factor for chronic trouble? Behav Neurosci 2006;120(2):474–6.
Murphy DR, Byfield D, Mccarthy P, Humphreys K, Gregory AA, Rochon R. Inter-examiner reliability of the hip extension test for suspected impaired motor control of the lumbar spine. J Manipulative Physiol Ther 2006;29(5):374–7.
O'sullivan P. Diagnosis and classification of chronic low back pain disorders: maladaptive movement and motor control impairments as underlying mechanism. Man Ther 2005;10(4):242–55.
Panjabi MM. What happens in the motion segment? Bull Hosp Jt Dis 1996;55(3):149–53.
Panjabi MM. A hypothesis of chronic back pain: ligament subfailure injuries lead to muscle control dysfunction. Eur Spine J 2006;15(5):668–76.
Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehab 1998;12(3):187–99.
Rasmussen-Barr E, Äng B, Arvidsson I, Nilsson-Wikmar L. Graded exercise for recurrent low-back pain – a randomized, controlled trial with 6-, 12-, and 36-month follow-ups. Spine (Phila Pa 1976) 2009;34(3):221–8.
Reeves NP, Narendra KS, Cholewicki J. Spine stability: the six blind men and the elephant. Clin Biomech (Bristol, Avon) 2007;22(3):266–74.
Roussel NA, Nijs J, Truijen S. Low back pain: clinimetric properties of the trendelenburg test, active straight leg raise test, and breathing pattern during active straight leg raising. J Manipulative Physiol Ther 2007;30(4):270–8.
Sahrmann S. Diagnosis and treatment of movement impairment syndromes; 2002.
Schabrun SM, Hodges PW. Muscle pain differentially modulates short interval intracortical inhibition and intracortical facilitation in primary motor cortex. J Pain 2012;13(2):187–94.
Sedaghat N, Latimer J, Maher C, Wisbey-Roth T. The reproducibility of a clinical grading system of motor control in patients with low back pain. J Manipulative Physiol Ther 2007;30(7):501–8.
Simopoulos TT, Manchikanti L, Singh V, Gupta S, Hameed H, Diwan S, et al. A systematic evaluation of prevalence and diagnostic accuracy of sacroiliac joint interventions. Pain Physician 2012;15(3):E305–44.
Stanton TR, Latimer J, Maher CG, Hancock MJ. A modified Delphi approach to standardize low back pain recurrence terminology. Eur Spine J 2011;20(5):744–52.
Tidstrand J, Horneij E. Inter-rater reliability of three standardized functional tests in patients with low back pain. BMC Musculoskelet Disord 2009;2(10):58–65.
Tsao H, Hodges PW. Immediate changes in feedforward postural adjustments following voluntary motor training. Exp Brain Res 2007;181(4):537–46.
Tsao H, Hodges PW. Persistence of improvements in postural strategies following motor control training in people with recurrent low back pain. J Electromyogr Kinesiol 2008;18(4):559–67.
Van Dillen LR, Sahrmann SA, Norton BJ, Caldwell CA, Fleming DA, Mcdonnell MK, et al. Reliability of physical examination used for classification of patients with low back pain. Phys Ther 1998;78(9):979–88.
Waddell G. The back pain revolution. Churchill Livingstone; 2004.
Walker BF. The prevalence of low back pain: systematic review of the literature from 1966 to 1998. J Spinal Disord 2000;13(3):205–17.
Woolsey N, Sahrmann S, Dixon L. Triaxial movement of the pelvis during prone knee flexion. Phys Ther 1988:68827.

Please cite this article in press as: Carlsson H, Rasmussen-Barr E, Clinical screening tests for assessing movement control in non-specific low-back pain. A systematic review of intra- and inter-observer reliability studies, Manual Therapy (2012), http://dx.doi.org/10.1016/j.math.2012.08.004