Title:

A systematic review and meta-analysis of the reliability and validity of sensorimotor measurement instruments in people with chronic low back pain

Authors:

Katja Ehrenbrusthoff, MSc. [a,b]

[a]Health and Social Care Institute, Teesside University, Middlesbrough, Tees Valley, TS1 3BX, United Kingdom

[b]Hochschule für Gesundheit, Department of Applied Health Sciences, Gesundheitscampus 6-8, 44801Bochum, Germany, katja.ehrenbrusthoff@hs-gesundheit.de

Cormac G Ryan [a]**, PhD**

[a]Health and Social Care Institute, Teesside University, Middlesbrough, Tees Valley, TS1 3BX, United Kingdom, c.ryan@tees.ac.uk

Christian Grüneberg [b]**, Prof. Dr.**

[b]Hochschule für Gesundheit, Department of Applied Health Sciences, Gesundheitscampus 6-8, 44801Bochum, Germany, christian.grueneberg@hs-gesundheit.de

Denis J Martin[a]**, Prof.**

[a]Health and Social Care Institute, Teesside University, Middlesbrough, Tees Valley, TS1 3BX, United Kingdom, d.martin@tees.ac.uk

Corresponding Author:

Katja Ehrenbrusthoff, Department of Applied Health Sciences, Hochschule für Gesundheit Bochum, Gesundheitscampus 6-8, 44801 Bochum, Germany, +49(0)234 77727 626, katja.ehrenbrusthoff@hs-gesundheit.de

Abstract

Background: Deficits in the sensorimotor system and its peripheral and central processing of the affected body part might be a contributing factor to chronic low back pain (CLBP). Hence, sensorimotor assessment is important. Valid and reliable sensorimotor measurement instruments are needed.

Objective: To investigate the reliability and validity of sensorimotor measurement instruments for people with chronic low back pain (CLBP).

Design: Systematic review and meta-analysis.

Methods: The review was undertaken using the COSMIN guidelines. Databases were searched for studies investigating the clinimetric properties of sensorimotor tests in people with CLBP. The methodological study quality was rated by two independent reviewers using the COSMIN 4-point rating checklist.

Results: Ten studies were included covering six sensorimotor measurement instruments with findings for reliability/measurement error, known-groups validity and convergent validity. The methodological quality ranged from poor to good, with only one study rated as good. There was insufficient evidence of enough quality to assess reliability/measurement error or convergent validity. Two-point discrimination, laterality judgement and movement control tests had moderate evidence supporting their ability to distinguish between healthy people and those with CLBP.

Conclusions: Two-point discrimination, laterality judgment and movement control tests demonstrate the greatest level of known-groups validity for people with CLBP. However, as the reliability of these measurement tools have yet to be established, this validity data should be interpreted cautiously. Further research is warranted to investigate the clinimetric properties of these sensorimotor techniques.

Keywords

Chronic low back pain, sensorimotor test, systematic review, meta-analysis

Introduction

Chronic low back pain (CLBP) is a major public health problem, with a lifetime prevalence of ~84% (Denteneer et al., 2016, Murray et al., 2013). It is a leading cause of disability worldwide (Murray et al., 2013). Many factors contribute to the development and/or maintenance of CLBP (Denteneer et al., 2016). It has been proposed that deficits in the sensorimotor system (sensorimotor dysfunction) could be a contributing factor (Apkarian et al., 2011, Catley et al., 2014, Moseley and Flor, 2012). As such, there is growing interest in outcome measures and interventions that attempt to measure and improve sensorimotor function in people with CLBP (Ehrenbrusthoff et al., 2016, Elgueta-Cancino et al., 2015, Louw et al., 2015, Louw et al., 2016, Villafane et al., 2015, Vuilleumier et al., 2015).

Sensorimotor function encompasses all sensory and motor elements necessary for an individual to interact with their environment (Shumway-Cook and Woollacott, 2007). This includes the output from the nervous system contributing to motor function and any sensory input contributing to the interpretation of body position and movement (Hodges and Falla, 2015). A range of sensorimotor measurement instruments (SMIs) exist that attempt to measure the construct of sensorimotor dysfunction, defined as a process of altered motor behavior, and/or distorted interpretation or inaccurate input of afferent sensory information (Hodges and Falla, 2015, Pelletier et al., 2015). Some SMIs require expensive specialist equipment and highly skilled technical staff, such as functional magnetic resonance imaging (fMRI). Such techniques are beyond the capacity of routine clinical practice. Thus, there is a need for simple SMIs that are clinically practicable, to facilitate sensorimotor assessment and intervention.

There are a number of clinically practicable SMIs, such as two-point discrimination (TPD), laterality judgement and movement control tests (MCTs) (Catley et al., 2013, Luomajoki, 2012, Moseley, 2006). An essential prerequisite for any clinical test is that it demonstrates sound clinimetric properties (De Vet et al., 2011), particularly, reliability and validity (Atkinson and Nevill, 1998, De Vet et al., 2011). The clinimetric properties of some SMIs have been investigated in healthy people and an array of patient groups (Auld et al., 2011, Stanton et al., 2013, Wand et al., 2014a). The clinimetric properties of some of these SMIs have been explored in people with CLBP but the extent and the quality of the work has not been systematically reviewed. Such a review is needed to guide research and clinical practice in the field. Thus, the aim of this study was to systematically investigate the reliability and validity of simple SMIs in people with CLBP.

Methods

The search strategy was developed in accordance with COSMIN recommendations (Terwee et al., 2011) and the PRISMA guidelines (Moher et al., 2010). This systematic review is registered on PROSPERO (Registration number: CRD42015026880).

Structured search strategies were designed using search terms appropriate for each database. Standardised database subject headings such as MeSH terms (in MEDLINE) and Subject Headings (in CINAHL) were used in each database, as appropriate. For the MEDLINE search, the sensitive PubMed search filter proposed by COSMIN for measurement properties was used (Terwee et al., 2009). Search

terms and synonyms were searched separately in four main categories and finally combined into one search string per database. The categories complied with COSMIN guidelines (Terwee et al., 2009) and were defined as:

1. Construct: tactile acuity OR sensorimotor dysfunction OR cortical reorganization
2. Target population: chronic low back pain
3. Measurement instrument: sensorimotor test
4. Measurement properties: sensitive COSMIN search filter for measurement properties ~~for~~ in MEDLINE

Electronic searches of databases were conducted by one author (K.E.) until March 30th 2015 using MEDLINE via PubMed, CINAHL via EBSCO, Embase via Ovid and Central via Wiley. The search was updated with a time restriction from March 30th 2015 to April 30th 2016 to identify relevant studies published ad interim. A full description of the search strategies can be found in the supplementary data (Appendix 1: Search strategies for all databases). Identified records were screened by K.E. by title-abstract initially and then by full-text screening. Hand searching of key reference lists was also conducted.

Eligibility Criteria

Studies were included if: 1) their target population were individuals with CLBP, defined as pain between the 12th rib and the buttock creases, persisting for 3 months (Savigny et al., 2009), 2) the SMI investigated claimed to measure a component of sensorimotor dysfunction, 3) the SMI investigated was practicable without sophisticated/expensive instrumentation (e.g. an functional Magnetic Resonance Imaging (fMRI) machine) not easily accessible in a routine clinical setting. An

example of an unsophisticated and inexpensive piece of equipment would be a goniometer, 4) the aim was to investigate one or more measurement properties of the SMI under investigation, 5) they were designed to investigate reliability or validity of the SMI, in accordance with the COSMIN taxonomy (Mokkink et al., 2009), 6) the study was published as a full original article in English or German.

Studies were excluded if: 1) they were of an intervention based or single-case design, 2) the SMI investigated required extensive technical skills and/or equipment not found in routine clinical practice (e.g. fMRI, motion analysis systems).

Data Extraction

According to the COSMIN recommendations for data extraction, the generalisability box of the COSMIN tool was used to extract data on characteristics of the study sample (median/mean age, distribution of sex, important disease characteristics, setting, country, language, sampling strategy, percentage of missing responses). In addition, details of each SMI data collection protocol were summarised and the measurement property results per SMI were extracted separately (De Vet et al. 2011). The extraction process was carried out by the lead author (K.E).

Methodological Quality Evaluation

The COSMIN four-point scoring checklist (Terwee et al., 2012) was used to assess the methodological quality of included studies. The checklist is a validated tool comprising 10 sections, each assessing a separate measurement property (Mokkink et al., 2010a, Mokkink et al., 2010b).Two reviewers (C.R. and K.E.) with prior

experience in using the checklist rated each study. Each item for methodological quality within each section was scored from excellent to poor. The overall score for the measurement property within the study was defined as the lowest rating among all response options within one section, termed as “worst score counts” (Terwee et al., 2012). Where multiple measurement properties were assessed within one study, this study received multiple methodological quality evaluations.

Evaluation of measurement properties

In the studies included in the review, the results for each SMI measurement property were evaluated against the pre-defined quality for good measurement properties (Terwee et al., 2007), (see table 1 for details). For validity, we investigated the construct validity sub-categories known-groups validity and convergent validity. Known groups validity was defined as an instrument’s ability to discriminate between people with and without the target condition or between people having different manifestations of the target condition, respectively (De Vet et al., 2011). Convergent validity was defined as the expected relationship between instruments measuring related constructs (De Vet et al., 2011).

Table 1: Quality criteria for measurement properties

Property	Rating	Quality Criteria
Reliability		
Internal consistency	+	Cronbach's alpha(s) ≥ 0.70
	?	Cronbach's alpha not determined or dimensionally unknown
	-	Cronbach's alpha(s) < 0.70
Reliability	+	ICC / weighted Kappa ≥ 0.70 OR Pearson's r ≥ 0.80
	?	Neither ICC / weighted Kappa, nor Pearson's r determined
	-	ICC / weighted Kappa < 0.70 OR Pearson's r < 0.80
Measurement error	+	MIC > SDC OR MIC outside the LOA
	?	MIC not defined
	-	MIC ≤ SDC OR MIC equals or inside LOA
Validity		
Content validity	+	All items are considered to be relevant for the construct to be measured, for the target population, and for the purpose of the measurement AND the questionnaire is considered to be comprehensive
	?	Not enough information available
	-	Not all items are considered to be relevant for the construct to be measured, for the target population, and for the purpose of the measurement OR the

		questionnaire is considered not to be comprehensive
Construct validity - Structural validity	+	Factors should explain at least 50% of the variance
	?	Explained variance not mentioned
	-	Factors explain < 50% of the variance
- Hypothesis testing	+	Correlations with instruments measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses AND correlations with related constructs are higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlations with instruments measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses AND correlations with related constructs are lower than with unrelated constructs
Cross-cultural validity	+	No differences in factor structure OR no important DIF between language versions
	?	Multiple group factor analysis not applied AND DIF not assessed
	-	Differences in factor structure OR important DIF between language versions
Criterion validity	+	Convincing arguments that gold standard is “gold” AND correlation with gold standard ≥ 0.70

	?	No convincing arguments that gold standard is "gold"
	-	Correlation with gold standard < 0.70
Responsiveness		
	+	Correlation with changes on instruments measuring the same construct ≥ 0.50 OR at least 75% of the results are in accordance with the hypotheses OR AUC ≥ 0.70 AND correlations with changes in related constructs are higher than with unrelated constructs
	?	Solely correlations determined with unrelated constructs
	-	Correlation with changes on instruments measuring the same construct < 0.50 OR < 75% of the results are in accordance with the hypotheses OR AUC < 0.70 AND correlations with changes in related constructs are lower than with unrelated constructs

Legend: MIC = minimal important change, SDC = smallest detectable change, LOA = limits of agreement, ICC = intraclass correlation coefficient, DIF = differential item functioning, AUC = area under the curve, + = positive rating,? = indeterminate rating, - = negative rating

Table taken from COSMIN guidelines (Terwee et al., 2011)[1].

[1] Reprinted from the Journal of Clinical Epidemiology 2007;, Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW. Quality criteria were proposed for measurement properties of health status questionnaires, 60:34-42., Copyright (2007), with permission from Elsevier

Data synthesis: meta-analysis and best evidence synthesis

Where multiple studies with comparable study designs investigated the same SMI and measurement property, a meta-analysis was conducted. For known-groups validity, mean scores and standard deviations from healthy and patient groups were pooled using the statistical package RevMan (Version 5) by means of forest plots (fixed effects model) to establish a pooled difference between groups. Heterogeneity was quantified using the I^2 (Higgins et al., 2003). Following the COSMIN recommendations, studies with a poor methodological score were excluded from quantitative pooling (Mokkink et al., 2009). Where quantitative pooling was not appropriate, a 'best evidence synthesis" approach was used, (see Table 2) (Guyatt et al., 2011, Schünemann et al., 2011).

Table 2: Level of Evidence for the quality of the measurement property

Level	**Rating***	**Criteria**
strong	+++ or ---	Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality
moderate	++ or --	Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality
limited	+ or -	One study of fair methodological quality
conflicting	+/-	Conflicting findings
unknown	?	Only studies of poor methodological quality
Legend: * + = positive rating,? = indeterminate rating, - = negative rating		
Table taken from COSMIN guidelines (Terwee et al., 2011)[2]		

[2] Reprinted from the Journal of Clinical Epidemiology 2007; Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, Bouter LM, de Vet HCW. Quality criteria were proposed for measurement properties of health status questionnaires, 60:34-42., Copyright (2007), with permission from Elsevier

Results

Study Selection

Initially, 4,285 studies were identified, of which 407 were excluded as duplicates and another 3,839 were excluded following title and abstract screening. Fifty studies were included for full-text assessment from which nine studies evaluating six SMIs were included. In the updated search, 686 studies were initially identified, however, only one additional relevant study was included in the final study list. Thus, in total, 10 studies (Table 3) evaluating six SMIs were included (Figure 1), within which three measurement properties were investigated: reliability/measurement error, known-groups validity, and convergent validity. Details for the data collection protocols for each study are summarised in supplementary data (Appendix 2: Individual study data collection protocols).

The findings for each measurement property per SMI from the individual studies are quantified in supplementary data (Appendices 4-8).

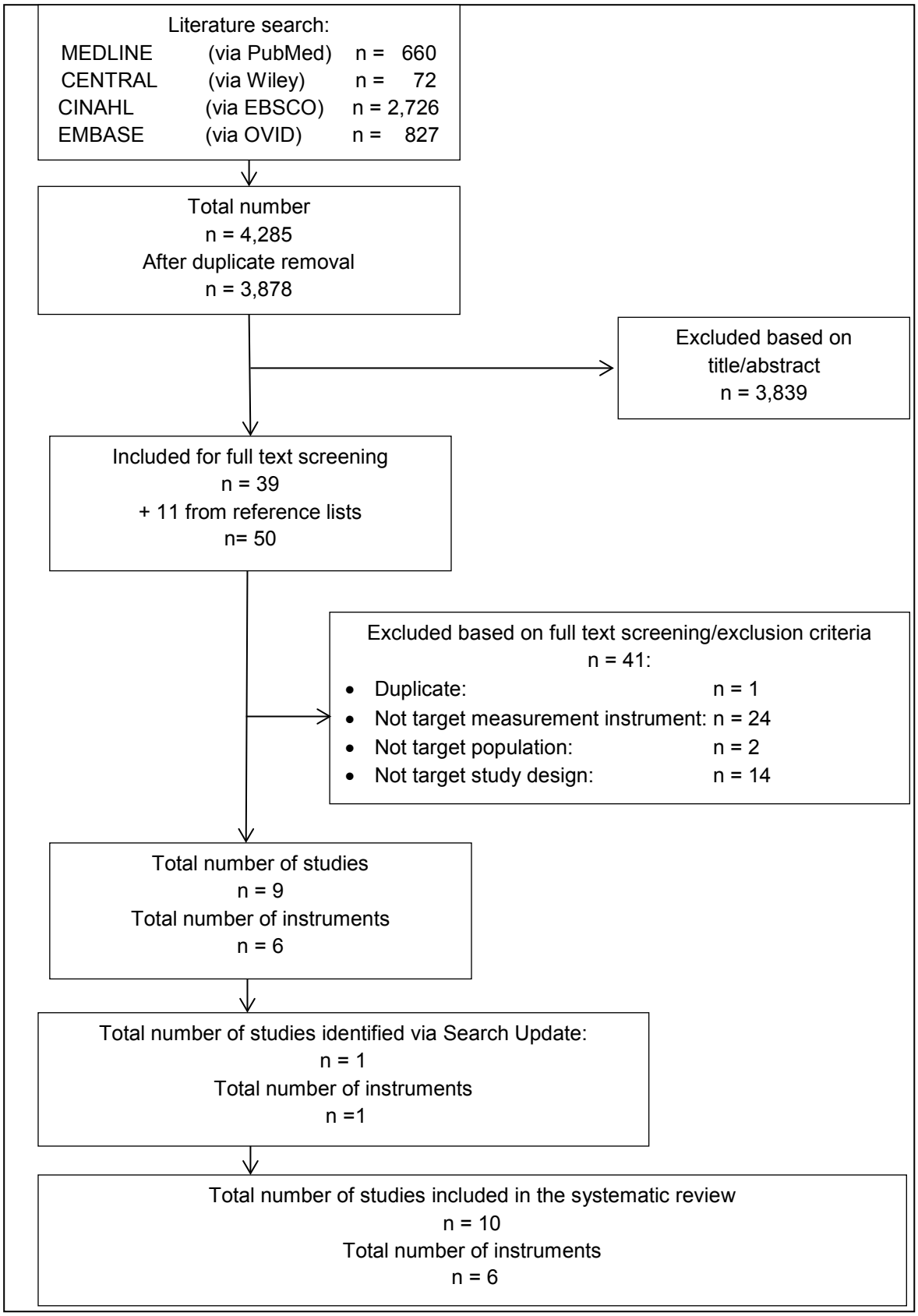


Figure 1: Flow chart of literature search and study selection process.

Table 3: Generalisability Box: Characteristics of included studies

		Patient Characteristics		
Author (Year)	**a) Instrument** **b) Measurement Property** **c) n**	**a) Mean Age (SD) (years)** **b) Distribution of Sex**	**a) Pain Severity Mean (SD)** **b) Disability Mean (SD)**	**a) Setting** **b) Country** **c) Language** **d) Sampling** **e) %of missing responses**
Linder et al (2015)	a) Laterality Judgement b) Known-Groups Validity c) CLBP: n=30 HC: n=30	a) CLBP: 44.9 (11.0) HC: 43.3 (9.6) b) CLBP: ♀ = 20 ♂ = 10 HC: ♀ = 20 ♂ = 10	a) VAS Scores: 55.3 (17.8) b) ODI Scores: 25.1 (13.1)	a) PT clinics b) Sweden c) Swedish d) CLBP: consecutive HC: convenience e) n=1 in CLBP group
Nishigami et al (2015)	a) TPD BID b) Known-Groups Validity c) CLBP: n=42 HC: n=17	a) CLBP normal BI: 65.1 (11.2) CLBP expanded BI: 56.7 (16.7) CLBP shrink BI: 62.0 (12.4) HC: 63.4 ± 12.2 c) CLBP: ♀ = 26 ♂ = 16 HC: ♀ = 8 ♂ = 9	a) VAS Scores normal BI: 48.3 (21.8) VAS Scores expanded BI: 42.5 (24.5) VAS Scores shrink BI: 42.0 (23.5) b) RMDQ Scores normal BI: 7.0 (2.4) RMDQ Scores expanded BI: 6.2 (3.4) RMDQ shrink BI: 6.8 (4.4)	a) Orthopedic clinic b) Japan c) Japanese d) Not stated e) Not stated

		Patient Characteristics		
Author (Year)	**a) Instrument b) Measurement Property c) n**	**a) Mean Age (SD) (years) b) Distribution of Sex**	**a) Pain Severity Mean (SD) b) Disability Mean (SD)**	**a) Setting b) Country c) Language d) Sampling e) %of missing responses**
Wand et al (2014)	a) FreBAQ b) Known-Groups Validity c) CLBP: n=51 HC: n=51	a) CLBP: 41.7 (14.0) HC: 38.7 (13.4) b) CLBP: ♀ = 21 ♂ = 30 HC: ♀ = 20 ♂ = 31	a) NRS Scores (0-100): 48.2 (17.8) b) RMDQ Scores: 10.1 (5.9)	a) Community PT practice; Department of pain management , The Sir Charles Gairdner Hospital, Perth, Western Australia b) English c) CLBP: convenience HC: convenience d) Not stated
Bowering et al (2014)	a) Laterality Judgement b) Known-Groups Validity c) Current back pain: n=117 History of back pain: n= 462 HC: n= 429	a) Complete sample: 37 (13) b) Complete sample: ♀ = 684; ♂ = 324	a) Not stated b) Not stated	a) Online study b) Australia c) English d) Convenience (online) e) 181 datasets excluded
Stanton et al (2013)	a) TPD b) Known Groups-Validity Convergent Validity c) CLBP: n=17 HC: n=18	a) CLBP: 45 (14) HC: 41 (11) b) CLBP: ♀ = 14 ♂ = 3 HC: ♀ = 11 ♂ = 7	a) Not stated b) Physical component of SF-36 (0-30): 19.7 (7.4)	a) Royal North Shore Hospital, Sydney b) Australia c) English d) CLBP data from Bray and Moseley (2011)

		Patient Characteristics		
Author (Year)	**a) Instrument b) Measurement Property c) n**	**a) Mean Age (SD) (years) b) Distribution of Sex**	**a) Pain Severity Mean (SD) b) Disability Mean (SD)**	**a) Setting b) Country c) Language d) Sampling e) %of missing responses**
				HC: convenience e) Not stated
Bray and Moseley (2011)	a) Laterality Judgement b) Known-Groups Validity c) CLBP: n=21 HC: n=14	a) CLBP: 44 (13) HC: 43 (7) b) CLBP: ♀ = 15; ♂ = 6 HC: ♀ = 9; ♂ = 5	a) VAS Scores (0-100): 37 (21) b) Not stated	a) Private PT practice b) United Kingdom (ethical approval) c) English d) Convenience sample e) Not stated
Luomajoki and Moseley (2011)	a) TPD Movement Control Tests b) Known-Groups Validity c) CLBP: n=45 HC: n=45	a) CLBP: 43 (15) HC: 41 (10) b) CLBP: ♀ = 25; ♂ = 20 HC: ♀ = 25; ♂ = 20	a) Not stated b) RMDQ Scores: 9 (5)	a) Private PT practice b) Switzerland c) German d) Convenience e) Not stated
Wand et al (2010)	a) TPD Graphesthesia b) Known-Groups Validity c) CLBP: n=19 HC: n=19	a) CLBP: 41 (12.5) HC: 34 (12.1) b) CLBP: ♀ = 11 ♂ = 8 HC: ♀ = 14 ♂ = 5	a) NRS Scores (0-10) usual pain: 3.9 (2.1) b) Physical component of SF-36 (0-30): 21.8 (5.0)	a) District General hospital, Perth, Western Australia b) Australia c) English d) CLBP: convenience HC: convenience e) 2 subjects with ambiguous pain scores and missing scores respectively, treated as missing values

		Patient Characteristics		
Author (Year)	**a) Instrument b) Measurement Property c) n**	**a) Mean Age (SD) (years) b) Distribution of Sex**	**a) Pain Severity Mean (SD) b) Disability Mean (SD)**	**a) Setting b) Country c) Language d) Sampling e) %of missing responses**
Moseley (2008)	a) TPD BID b) Known-Groups Validity c) CLBP: n= 6 HC: n=10	a) CLBP: 43.83 (11.12) HC: not stated b) CLBP: ♀ = 3 ♂ = 3 HC: ♀ = 5 ♂ = 5	a) VAS Scores: 47.2 (12.54) b) Not stated	a) Not stated b) UK c) English d) Consecutive e) Not stated
Luomajoki (2008)(	a) Movement Control Tests b) Known-Groups Validity c) LBP: n=108 HC: n= 102	a) LBP: 41 (15) HC: 37 (12) b) LBP: ♀ = 72 ♂ = 36 HC: ♀ = 58 ♂ = 44	a) Not stated b) RMDQ Scores: 8(5)	a) Outpatient PT clinics, Canton Aargau b) Switzerland c) German d) CLBP: consecutive HC: convenience e) Not stated
Separate Patient Characteristics of Studies investigating Reliability (Subsamples of CLBP patients)				
Wand et al (2014)	a) FreBAQ b) Reliability c) CLBP: n= 26	a) 42 (14) b) ♀ = 12; ♂ = 14	a) Back Pain Intensity (0-100): 47.7 (14.4) b) RMDQ Scores: 10.6 (6.0)	a) Community PT practice b) Australia c) English d) Not stated e) N=1 did not return second questionnaire; handled as missing item
Bray and Moseley (2011)	a) Laterality Judgement b) Reliability c) CLBP: n= 5	a) 46 (16) b) ♀ = 1; ♂ = 4	a) VAS Scores: 46 (23) b) ODI Scores: 25.1 (13.1)	a) Private PT practice b) UK c) English d) Convenience e) Not stated

		Patient Characteristics		
Author (Year)	**a) Instrument** **b) Measurement Property** **c) n**	**a) Mean Age (SD) (years)** **b) Distribution of Sex**	**a) Pain Severity Mean (SD)** **b) Disability Mean (SD)**	**a) Setting** **b) Country** **c) Language** **d) Sampling** **e) %of missing responses**
Linder et al (2015)	a) Laterality Judgement b) Measurement Error Reliability c) CLBP: n= 30	a) 44.9 (11) b) ♀ = 20; ♂ = 10	a) VAS Scores: 55.3 (17.8) b) ODI Scores: 25.1 (13.1)	a) PT clinics b) Sweden c) Swedish d) Convenience e) Not stated
Legend: CLBP= Chronic Low Back Pain; HC= Healthy Controls; ♂= male; ♀=female; TPD = Two-Point Discrimination; BID= Body Image Drawings; FreBAQ= Fremantle Back Awareness Questionnaire; ODI = Oswestry Disability Score; PT= Physiotherapy; SF-36 = 36-Item Short Form Health Survey (SF-36); RMDQ = Roland Morris Disability Questionnaire; NRS = Numerical Rating scale; Back Pain Intensity (0-100); Data are presented as Mean (SD) unless otherwise stated.				

Methodological quality evaluation of the studies

Across all 10 included studies, four methodological quality evaluations concerning reliability and/or measurement error were undertaken and received a poor methodological quality rating. Sixteen methodological quality evaluations of known-groups or convergent validity were conducted with one rated as good, eleven as fair and four as poor (see Table 4 and 5[3]).

[3] Reprinted from the Journal of Clinical Epidemiology, 63, Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW., International consensus on taxonomy, terminology, and definitions of measurement properties: results of the COSMIN study, 737-745., Copyright (2010), with permission from Elsevier

Table 4: Methodological quality evaluation: Reliability and Measurement Error Studies

	Design Requirements Reliability											**Statistical Methods**			
Instrument and Measure-ment Property	overall score	Item 1: Was the percentage of missing	Item 2: Was there a description of how	Item 3: Was the sample size included in	Item 4: Were at least two measurements	Item 5: Were the administrations	Item 6: Was the time stated?	Item 7: Were patients stable in the interim period on the construct to	Item 8: Was the time interval	Item 9: Were the test conditions similar for both measurements? e.g. type of administration	Item 10: Were there any important flaws in	Item 11: for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	Item 12: for dichotomous/ nominal/ordinal scores: Was kappa calculated?	Item 13: for ordinal scores: Was a weighted kappa calculated?	Item 14: for ordinal scores: Was the weighting scheme
Laterality Judgement Reliability															
Linder et al (2015)	poor	good	fair	poor	excell.	excell.	excell	good	excell.	fair	fair	excell.	n.a.	n.a.	n.a.
Bray and Moseley (2011)	poor	good	good	poor	excell.	excell.	excell	good	excell.	good	fair	good	n.a.	n.a.	n.a.
FreBaQ Reliability															
Wand et al (2014)	poor	good	fair	poor	excell.	good	excell.	poor	excell.	fair	fair	excell.	n.a.	n.a.	n.a.

	Design Requirements Measurement Error											**Statistical Methods**	
Instrument and Measure-ment Property	overall score	Item 1: Was the percentage of missing	Item 2: Was there a description of how	Item 3: Was the sample size included in	Item 4: Were at least two measurements	Item 5: Were the administrations	Item 6: Was the time interval stated?	Item 7: Were patients stable in the	Item 8: Was the time interval	Item 9: Were the test conditions similar for both measurements? e.g. type of administration	Item 10: Were there any important flaws in	Item 11:.for CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable	
Laterality Judgement Measure-ment Error													
Linder et al (2015)	poor	good	fair	poor	excell.	excell.	excell.	good	excell.	fair	fair	good	
FreBaQ= Fremantle Back Awareness Questionnaire; CTT= Classical Test Theory; excell. = excellent; n.a.= not applicable													

Table 5: Methodological quality evaluation: Known-Groups and Convergent Validity Studies

	Design Requirements										**Statistical Methods**
Instrument and Measurement Property	overall score	Item 1: Percentage of missing items given?	Item 2: Description of how missing items were handled?	Item 3: Adequate sample size included in the analysis?	Item 4: Formulation of hypotheses regarding Correlations or mean differences a priori?	Item 5: Was the expected direction of correlations or mean differences included in the	Item 6: Was the expected absolute or relative magnitude of correlations or mean	Item 7: For CV: was an adequate description of the	Item 8: For CV: Were the measurement properties of the	Item 9: Were there any important flaws in the design or methods of the study?	Item 10: Were design and statistical methods adequate for the hypotheses to be tested?
TPD KGV											
Nishigami et al (2015) TPD measured as side-to-side differences	fair	good	fair	good	fair	good	good	n.a.	n.a.	fair	excell.
Stanton et al (2013)	fair	good	fair	fair	good	good	good	n.a	n.a.	fair	excell.
Luomajoki and Moseley (2011)	fair	good	fair	good	excell.	excell.	good	n.a.	n.a.	excell.	good
Wand et al (2010)	fair	good	fair	fair	excell.	excell.	good	n.a.	n.a.	fair	excell.

	Design Requirements										Statistical Methods
Instrument and Measurement Property	overall score	Item 1: Percentage of missing items given?	Item 2: Description of how missing items were handled?	Item 3: Adequate sample size included in the analysis?	Item 4: Formulation of hypotheses regarding Correlations or mean differences a priori?	Item 5: Was the expected direction of correlations or mean differences included in the	Item 6: Was the expected absolute or relative magnitude of correlations or mean	Item 7: For CV: was an adequate description of the	Item 8: For CV: Were the measurement properties of the	Item 9: Were there any important flaws in the	Item 10: Were design and statistical methods adequate for the hypotheses to be tested?
Moseley (2008)	poor	good	good	poor	fair	good	good	n.a.	n.a.	fair	poor
TPD CV											
Stanton et al (2013)	poor	good	fair	poor	excell.	excell.	good	excell.	fair	poor	excell.
MCT KGV											
Luomajoki Moseley (2011)	fair	good	fair	good	excell.	excell.	good	n.a.	n.a.	fair	good
Luomajoki et al (2008)	fair	good	fair	excell.	good	good	good	n.a.	n.a.	fair	good
Graphestesia KGV											
Wand et al (2010)	fair	good	fair	fair	excell.	excell.	good	n.a.	n.a.	fair	excell.
Laterality Judgement KGV											
Linder et al (2015)	fair	good	fair	good	fair	good	good	n.a.	n.a.	excell. #	excell.
Bowering et al (2014)	good	excell.	excell.	excell.	good	excell.	good	n.a.	n.a.	excell.	excell.

	Design Requirements										Statistical Methods
Instrument and Measurement Property	overall score	Item 1: Percentage of missing items given?	Item 2: Description of how missing items were handled?	Item 3: Adequate sample size included in the analysis?	Item 4: Formulation of hypotheses regarding Correlations or mean differences a priori?	Item 5: Was the expected direction of correlations or mean differences included in the	Item 6: Was the expected absolute or relative magnitude of correlations or mean	Item 7: For CV: was an adequate description of the	Item 8: For CV: Were the measurement properties of the	Item 9: Were there any important flaws in the	Item 10: Were design and statistical methods adequate for the hypotheses to be tested?
Bray and Moseley (2011)	fair	good	fair	fair	excell.	excell.	good	n.a.	n.a.	fair	excell.
Laterality Judgement CV											
Stanton et al (2013)	poor	good	fair	poor	excell.	excell.	good	excell.	fair	poor	excell.
BIDs KGV											
Nishigami et al (2015)	fair	good	fair	good	fair	Good	good	n.a.	n.a.	fair	excell.
Moseley (2008)	poor	good	good	poor	fair	good	good	n.a.	n.a.	fair	poor
FreBaQ KGV											
Wand et al (2014)	fair	good	fair	excell.	fair	good	good	n.a.	n.a.	excell.	excell.

TPD = Two-Point Discrimination; KGV= Known-Groups Validity; MCT= Movement Control Test; CV=Convergent Validity; BID= Body Image Drawings;

FreBaQ= Fremantle Back Awareness Questionnaire; excell.= excellent; n.a.= not applicable

Reviewer Agreement

The inter-rater agreement for methodological quality between raters was good (Altman, 1991) (absolute agreement = 73%, Cohen's Kappa κ = 0.62 (95% CI 0.54, 0.70). Initial disagreement was resolved by consensus (see supplementary data Appendix 3: Reviewer Agreement).

Measurement properties

Reliability and Measurement Error

Studies were identified that investigated the reliability of laterality judgement and the FreBaQ.

Laterality Judgement: reliability

Two studies (Bray and Moseley, 2011, Linder et al., 2015) (CLBP = 10 and CLBP=25/22), both of poor methodological quality, investigated the reliability of laterality judgement. Intraclass correlation coefficients (ICCs) for response time and accuracy were provided (see supplementary data: Appendix 5).

ICC values ranged from 0.51 to 0.91 for reaction time and from 0.69 to 0.92 for accuracy. Thus, the level of reliability could be considered acceptable for accuracy, but not for reaction time against the predefined acceptable level (ICC ≥0.70).

As this body of evidence consisted of only poor quality studies, the evidence for the reliability of laterality judgement was classified as unknown.

Linder et al. (2015) investigated measurement error reporting the coefficients of variation (CV) for reaction time and accuracy of repeated measurements between time point one and two (CLBP= 25) and between time point two and three (CLBP= 22). For reaction time, the CV reduced from 19.6 % to 6.2 % whilst for accuracy it remained stable at 6.46 % to 6.77%. Data on minimally important change (MIC) were not provided. As this body of evidence consists of only one poor methodological quality study, the evidence for the measurement error of laterality judgement was classified as unknown.

Fremantle Back Awareness Questionnaire (FreBaQ): reliability

One study (Wand et al., 2014b), of poor methodological quality (n=26), investigated the test-retest reliability of the FreBaQ over a period of one week in people with CLBP (see supplementary data: Appendix 8). The test-retest performance was $ICC_{2.1}$ (95% CI) =0.652(0.307-0.848) (Agreement) and $ICC_{2.1}$ (95% CI)=0.667(0.317-0.857) (Consistency). This was below the COSMIN threshold of ≥0.70.

However, as this body of evidence consists of only one poor quality study, the evidence for the test-retest reliability of the FreBaQ was classified as unknown.

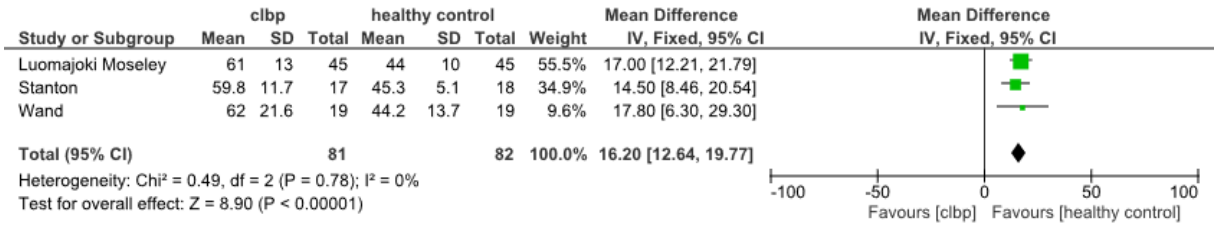
Known-Groups Validity

Studies investigating the known-groups validity of TPD, graphaesthesia, laterality judgement, MCTs and the FreBaQ were found.

Two Point Discrimination (TPD): Known-Groups Validity

Five studies (CLBP=129, HC=91) investigated the known-groups validity of TPD were found. Four were of fair quality (Loumajoki and Moseley, 2011, Nishigami et al., 2015, Stanton et al., 2013, Wand et al., 2010) and one of poor methodological quality (Moseley, 2008) (see supplementary data: Appendix 4). Four of these studies were broadly consistent regarding their measurement protocol, in particular regarding the region of the lower back assessed and the horizontal TPD measurement approaches (Luomajoki and Moseley, 2011, Moseley, 2008, Stanton et al., 2013, Wand et al., 2010). The other study used side-to-side difference of TPD threshold as the outcome measure (Nishigami et al., 2015) and categorised patients according to their perceived body image. Of the four comparable studies, three identified a statistically wider TPD threshold for people with CLBP whilst the one study of poor methodological quality noted no difference between patients and healthy controls. When the three fair quality studies were statistically pooled people with CLBP demonstrated a statistically wider TPD threshold of 16 mm than healthy people (figure 2). No evidence of heterogeneity was found when the data was statistically pooled (Higgins and Green, 2011). Thus, there was moderate evidence from three studies of fair methodological quality that TPD possesses known-groups validity.

Figure 2: Forest Plot comparing TPD regarding its known-groups validity to distinguish between people with CLBP and healthy controls



Legend: IV = inverse variance; CI= confidence interval. Note: only data from horizontal TPD measurements were included in this analysis, leaving out values from vertical measurements from Luomajoki and Moseley (2011).

Nishigami et al. (2015) also reported a statistically significant difference between groups regarding the side-to-side differences of TPD thresholds (CLBP=42, HC=17). Thus, there is limited evidence from one study of fair methodological quality that TPD side-to-side difference possesses known-groups validity.

Laterality Judgment: Known Groups Validity

Three studies (Bowering et al., 2014, Bray and Moseley, 2011, Linder et al., 2015) (168=CLBP, 473=HC) assessed the known-groups validity of laterality judgement, two of fair (Bray and Moseley, 2011, Linder et al., 2015) and one (Bowering et al., 2014) of good methodological quality (see supplementary data: Appendix 5). Two studies (Bowering et al., 2014, Bray and Moseley, 2011) found a significant difference regarding laterality judgement accuracy, but only one concerning reaction time (Bowering et al., 2014), between people with CLBP and healthy controls, whilst one study found neither a statistical difference for reaction time nor accuracy (Linder et al., 2015). When the three studies were quantitatively pooled, people with CLBP were, on average, 9% less accurate and 0.1 seconds slower than healthy controls, both statistically significant. (figure 3a and 3b). The level of heterogeneity was substantial with I^2 values of 65% and 90% for reaction time and accuracy respectively, therefore, the findings should be interpreted cautiously (Higgins and Green, 2011). Hence, there was moderate evidence from one study of good and two

of fair methodological quality that demonstrated known-groups validity in laterality judgement accuracy and reaction time.

Figure 3a: Forest Plot Laterality Judgement Reaction Time

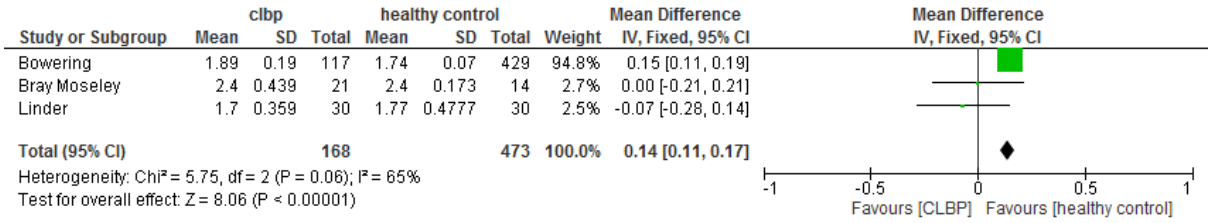
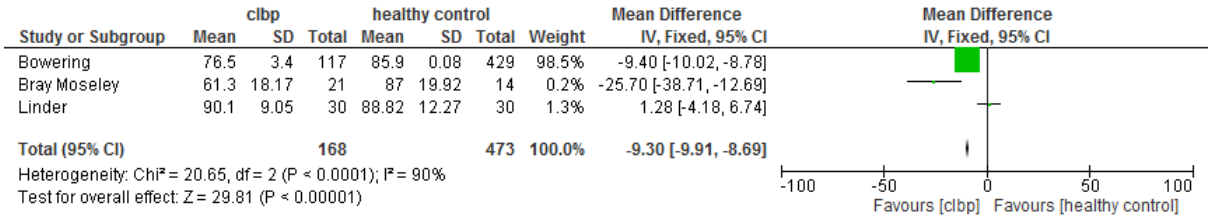


Figure 3b: Forest Plot Laterality Judgement Accuracy



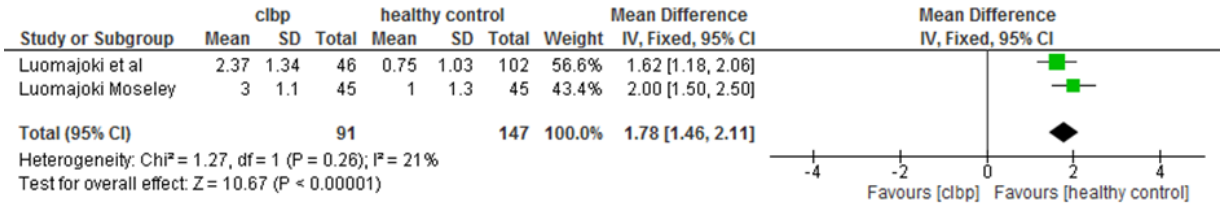
Legend: IV = inverse variance; CI= confidence interval. For Linder et al. (2015) combined data for left/right laterality judgements were used for quantitative pooling; for Bowering et al (2014) only the data from “current back pain” patients were used. The data from Bowering et al (2014) was calculated from a graph using the software DigitizeIt®.

Movement Control Tests (MCTs): Known-Groups Validity

Two studies of fair methodological quality (Luomajoki et al., 2008, Luomajoki and Moseley, 2011) (CLBP=91, HC=147) investigated the known-groups validity of MCTs

(see supplementary data: Appendix 6). Both studies independently, and when quantitatively pooled (figure 4), found a statistically poorer performance in the MCT performance by people with CLBP compared to healthy controls. The degree of heterogeneity was low and, likely, of no importance (Higgins and Green, 2011). Thus, there is moderate evidence from two studies of fair methodological quality that MCTs demonstrate known-groups validity.

Figure 4: Forest Plot comparing movement control tests regarding their known-groups validity to distinguish between people with CLBP and healthy controls



Legend: IV = inverse variance; CI= confidence interval

Graphesthesia: Known-Groups Validity

One study (Wand et al., 2010) (CLBP=19, HC=19) of fair methodological quality evaluated the known-groups validity of graphesthesia (see supplementary data: Appendix 7). Performance, as adjudged by letter recognition error rates, was poorer in patients (mean difference; 6.1, 95% CI: 1.3 to 11.0). Thus, there was limited evidence from one study of fair methodological quality that graphesthesia demonstrates known-groups validity.

FreBaQ: Known-Groups Validity

One study, which investigated the FreBaQ reliability also (Wand et al., 2014b) looked at the FreBaQ known-groups validity (see supplementary data: Appendix 8). Due to the larger sample size gathered for this specific question (n=52 CLBP, n=52 HC) the study's methodological quality was rated fair (see supplementary data: Appendix 8). Data were presented as medians and inter-quartile ranges. The FreBaQ score (median [range]) for the patient and control group was 11 [0-26] and 0 [0–6], respectively, indicating poorer performance by patients ($p<0.05$). Thus, there is limited evidence from one study of fair methodological quality that the FreBaQ demonstrates known-groups validity.

Convergent Validity

Studies that investigated the convergent validity of TPD, laterality judgement and BID were found.

Two Point Discrimination and Laterality judgment: Convergent Validity

One study (n=17) of poor methodological quality investigated the convergent validity of TPD against laterality judgement in people with CLBP (Stanton et al., 2013) (see supplementary data: Appendices 4 and 5). An increase in the TPD threshold by 1 mm was associated with a decrease in accuracy of 0.6% (β= - 0.06, 95% CI: 0.80 to 0.43). However, as this body of evidence consists of only one poor quality study, the

quality of evidence for convergent validity of TPD and laterality judgment was classified as unknown.

Two-Point Discrimination and Body Image Drawings: Convergent Validity

Two studies, one of fair (Nishigami et al., 2015) (CLBP= 42) and one of poor methodological quality (Moseley, 2008) (CLBP=6) reported aspects of convergent validity for BIDs compared to TPD in a qualitative manner. Nishigami et al. (2015) displayed TPD values for participants with CLBP who drew either a normal, expanded or shrunken body image. These data suggested that TPD was increased in patients who reported an expanded BID. Similarly, Moseley (Moseley, 2008) reported an increase of the TPD threshold corresponding to the zone of the absence or disruption of the BIDs. However, as neither study quantified the relationship between TPD and BIDs, no evidence statement can be drawn regarding the convergent validity of BIDs.

Discussion

The aim of this study was to systematically investigate the reliability and validity of SMIs in people with CLBP. Ten studies investigating the following six SMIs were included; TPD, laterality judgment, MCTs, BIDs, FreBaQ, and graphesthesia. The methodological quality ranged from poor to good with only one study rated good (Bowering et al., 2014). The SMIs with the strongest support in the literature were TPD, laterality judgment and MCTs. There was moderate evidence to support the known-groups validity of these three SMIs. However, in general, there was a lack of high-quality studies investigating the clinimetric properties of SMIs for people with

CLBP. Hence, data collected using these techniques should be interpreted cautiously.

Only three studies assessed reliability comprising two SMIs, the FreBaQ and laterality judgement. All three studies were graded as methodologically poor, primarily due to small sample sizes. Thus, the level of evidence for the reliability and/or measurement error of FreBaQ and laterality judgement was unknown. For all other outcome measures, there were no studies to inform an evidence statement regarding reliability or measurement error. Regarding convergent validity, there were four poor quality studies, investigating the convergent validity of either TPD, laterality judgement or BIDs. Thus, the level of convergent validity for these SMIs was considered unknown, and for the other outcome measures, there were no studies to inform an evidence statement. The state of the evidence only allowed for statements to be made with respect to the known-groups validity of the SMIs.

TPD, laterality judgement and MCTs demonstrated the strongest evidence of known-groups validity. Regarding TPD, our meta-analysis demonstrated a mean difference between healthy controls and people with CLBP of 16mm. This is broadly in keeping with a previous meta-analysis by Catley et al. (2014) which compared TPD performance between healthy controls and people with CLBP reporting a mean difference of 11.7mm (95% CI:5.5 mm to 17.8 mm). The differences between our results and those of Catley et al. (21) are explained by the exclusion of the vertical TPD measurements from Luomajoki and Moseley (Loumajoki and Moseley, 2011) from the present meta-analysis and by the exclusion of the results from Moseley et al. (2008) as this study was rated of poor methodological quality in our review. Meta-analysis for laterality judgement and MCTs demonstrated evidence for known-groups

validity of both outcome measures. However, there are no previous meta-analyses with which to compare our findings.

There was less evidence regarding graphesthesia, the FreBAQ and BIDs. One study of fair methodological quality implied a degree of known-groups validity for graphesthesia which was in line with results from studies investigating this technique in other clinical conditions, such as Parkinson disease (Jobst et al., 1997) and corticobasal degeneration (Drago et al., 2010). The results from one study of fair methodological quality demonstrated a degree of known-groups validity for the FreBaQ. A further study (Wand et al., 2016), published after the search cut-off date for the present review, investigated the psychometric properties of the FreBAQ by means of a Rasch analysis in 255 people with CLBP demonstrating adequate internal consistency with a person reliability index of 0.74 and a Cronbach's Alpha Value of 0.80. As adequate internal consistency is an essential prerequisite for questionnaires which intend to measure a single underlying construct by multiple items (Terwee et al., 2007), the results provide a basis for further psychometric evaluation. There was very limited evidence upon which to make any recommendations regarding the use of BIDs in people with CLBP.

Review limitations

The search was restricted to full peer-reviewed published articles to enhance quality control, thus relevant conference papers/grey literature may have been excluded. Only one author undertook the screening and selection process increasing the risk of inadvertently excluding relevant studies. Additionally, only one reviewer extracted the

data from the included studies, which increases the risk of errors in the extraction process.

There were variations in the data collection protocols reported in the reviewed studies. In some cases, the variations were quite marked. For example, Nishigami et al. (2015) measured TPD performance as side-to-side differences, which was very different to the TPD protocols used in the other included studies. However, the protocol differences between the other studies were more subtle (see supplementary data: Appendix 2). These variations in data collection protocols may have reduced the comparability of the studies. In addition, the studies included in this review tended to have small sample sizes, which can lead to over-inflated effect sizes. This may have influenced the results of our meta-analyses.

The level of heterogeneity was substantial when the studies for laterality judgment were pooled for both reaction time and accuracy, thus these meta-analyses should be interpreted cautiously.

A key issue affecting the strengths of recommendations which can be drawn from this systematic review was the quality of included studies. Only one included study was rated as being of good methodological quality. In addition, the degree of reliability could not be established for any of the measures investigated in this review. Reliability is an important prerequisite of validity. Thus, the validity data presented should be interpreted cautiously. There is a need for higher quality studies investigating the clinimetric properties of SMIs for people with CLBP, with particular attention to recruiting adequate sample sizes.

Conclusion

There was a lack of high quality studies investigating the clinimetric properties of SMIs in people with CLBP. The methodological quality of the studies were predominately rated as poor or fair, with a small sample size frequently the reason for a low rating. The strongest body of evidence currently exists for TPD, laterality judgment and MCT with respect to known-groups validity. However, as the reliability of these measurement tools have yet to be established, this validity data should be interpreted cautiously. There is an urgent need to undertake high quality studies investigating the clinimetric performance of SMIs for people with CLBP in order to guide clinical and research practice. Given the state of the evidence, data collected using these SMIs should be treated cautiously.

Acknowledgements

The authors would like to acknowledge Ms M.B. (Librarian) for her advice and help in the development of the literature search.

Funding

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

References

Altman DG. Practical statistics for medical research. New York;London;: Chapman and Hall; 1991.

Apkarian AV, Hashmi JA, Baliki MN. Pain and the brain: specificity and plasticity of the brain in clinical chronic pain. Pain. 2011;152:S49.

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

Auld ML, Boyd RN, Moseley GL, Johnston LM. Tactile assessment in children with cerebral palsy: a clinimetric review. Physical & occupational therapy in pediatrics. 2011;31:413-39.

Bowering KJ, Butler DS, Fulton IJ, Moseley GL. Motor imagery in people with a history of back pain, current back pain, both, or neither. Clinical Journal of Pain. 2014;30:1070-5.

Bray H, Moseley GL. Disrupted working body schema of the trunk in people with back pain. Br J Sports Med. 2011;45:168-73.

Catley MJ, O'Connell NE, Berryman C, Ayhan FF, Moseley GL. Is tactile acuity altered in people with chronic pain? A systematic review and meta-analysis. The Journal of Pain. 2014;15:985-1000.

Catley MJ, Tabor A, Wand BM, Moseley GL. Assessing tactile acuity in rheumatology and musculoskeletal medicine—how reliable are two-point discrimination tests at the neck, hand, back and foot? Rheumatology. 2013.

De Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide: Cambridge University Press; 2011.

Denteneer L, Van Daele U, De Hertogh W, Truijen S, Stassijns G. 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.

Drago V, Foster PS, Edward D, Wargowich B, Heilman KM. Graphesthesia: A test of graphemic movement representations or tactile imagery? Journal of the International Neuropsychological Society. 2010;16:190-3.

Ehrenbrusthoff K, Ryan CG, Grueneberg C, Wolf U, Krenz D, Atkinson G, et al. The intra- and inter-observer reliability of a novel protocol for two-point discrimination in individuals with chronic low back pain. Physiological measurement. 2016;37:1074-88.

Elgueta-Cancino E, Schabrun S, Danneels L, Hodges P. Validation of a Clinical Test of Thoracolumbar Dissociation in Chronic Low Back Pain. Journal of Orthopaedic & Sports Physical Therapy. 2015:1-37.

Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. Journal of clinical epidemiology. 2011;64:380-2.

Higgins J, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses [journal article as teaching resource, deposited by John Flynn]. British medical journal. 2003;327:557-60.

Higgins JP, Green S. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2011.

Hodges P, Falla D. Interaction between pain and sensorimotor control. Grieve's Modern Musculoskeletal Physiotherapy: Elsevier, UK; 2015.

Jobst EE, Melnick ME, Byl NN, Dowling GA, Aminoff MJ. Sensory perception in parkinson disease. Archives of Neurology. 1997;54:450-4.

Linder M, Michaelson P, Roijezon U. Laterality judgments in people with low back pain - A cross-sectional observational and test-retest reliability study. Man Ther. 2015.

Loumajoki H, Moseley G. Tactile acuity and lumbopelvic motor control in patients with back pain and healthy controls. Br J Sports Med. 2011;45:437 - 40.

Louw A, Farrell K, Wettach L, Uhl J, Majkowski K, Welding M. Immediate effects of sensory discrimination for chronic low back pain: A case series. New Zealand Journal of Physiotherapy. 2015;43:58-63.

Louw A, Zimney K, Puentedura EJ, Diener I. The efficacy of pain neuroscience education on musculoskeletal pain: A systematic review of the literature. Physiotherapy Theory and Practice. 2016;32:332-55.

Luomajoki H. Sechs Richtige: Mit der Testbatterie die lumbale Bewegungskontrolle untersuchen. manuelletherapie. 2012;16:220-5.

Luomajoki H, Lean C, de Bruin E, Ariaksinen O. Movement Control tests of the low back; evaluation of the difference between patients with low back pain and healthy controls. BMC Musculoskeletal Disorders. 2008;9.

Luomajoki H, Moseley GL. Tactile acuity and lumbopelvic motor control in patients with back pain and healthy controls. British Journal of Sports Medicine. 2011;45:437-40.

Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. International Journal of Surgery. 2010;8:336-41.

Mokkink LB, Terwee CB, Gibbons E, Stratford PW, Alonso J, Patrick DL, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) checklist. BMC medical research methodology. 2010a;10:82.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist manual. Amsterdam: VU University Medical Centre. 2009.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and

definitions of measurement properties for health-related patient-reported outcomes. Journal of clinical epidemiology. 2010b;63:737-45.

Moseley GL. Graded motor imagery for pathologic pain. Neurology. 2006;67:2129-34.

Moseley GL. I can't find it! Distorted body image and tactile dysfunction in patients with chronic back pain. Pain. 2008;140:239-43.

Moseley GL, Flor H. Targeting Cortical Representations in the Treatment of Chronic Pain A Review. Neurorehabilitation and neural repair. 2012;26:646-52.

Murray CJ, Richards MA, Newton JN, Fenton KA, Anderson HR, Atkinson C, et al. UK health performance: findings of the Global Burden of Disease Study 2010. The lancet. 2013;381:997-1020.

Nishigami T, Mibu A, Osumi M, Son K, Yamamoto S, Kajiwara S, et al. Are tactile acuity and clinical symptoms related to differences in perceived body image in patients with chronic nonspecific lower back pain? Manual Therapy. 2015;20:63-7.

Pelletier R, Higgins J, Bourbonnais D. Is neuroplasticity in the central nervous system the missing link to our understanding of chronic musculoskeletal disorders? BMC musculoskeletal disorders. 2015;16:1.

Pin TW. Psychometric Properties of 2-Minute Walk Test: A Systematic Review. Archives of physical medicine and rehabilitation. 2014;95:1759-75.

Savigny P, Watson P, Underwood M. Early management of persistent non-specific low back pain: summary of NICE guidance. BMJ: British Medical Journal. 2009;338.

Schünemann H, Hill S, Guyatt G, Akl EA, Ahmed F. The GRADE approach and Bradford Hill's criteria for causation. Journal of epidemiology and community health. 2011;65:392-5.

Shumway-Cook A, Woollacott MH. Motor control: translating research into clinical practice: Lippincott Williams & Wilkins; 2007.

Stanton TR, Lin CWC, Bray H, Smeets RJEM, Taylor D, Law RYW, et al. Tactile acuity is disrupted in osteoarthritis but is unrelated to disruptions in motor imagery performance. Rheumatology (United Kingdom). 2013;52:1509-19.

Terwee C, Jansma E, Riphagen I, Vet HW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Quality of Life Research. 2009;18:1115-23.

Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology. 2007;60:34-42.

Terwee CB, de Vet H, Prinsen C, Mokkink L, Terwee CB. Protocol for systematic reviews of measurement properties. 2011.

Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. 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.

Villafane JH, Zanetti L, Isgro M, Cleland JA, Bertozzi L, Gobbo M, et al. Methods for the assessment of neuromotor capacity in non-specific low back pain: Validity and applicability in everyday clinical practice. Journal of back and musculoskeletal rehabilitation. 2015;28:201-14.

Vuilleumier PH, Biurrun Manresa JA, Ghamri Y, Mlekusch S, Siegenthaler A, Arendt-Nielsen L, et al. Reliability of Quantitative Sensory Tests in a Low Back Pain Population. Regional anesthesia and pain medicine. 2015;40:665-73.

Wand B, Di Pietro F, George P, O'Connell NE. Tactile thresholds are preserved yet complex sensory function is impaired over the lumbar spine of chronic non-specific low back pain patients. A preliminary investigation. Physiotherapy. 2010;96:317-23.

Wand BM, Catley MJ, Luomajoki HA, O'Sullivan KJ, Di Pietro F, O'Connell NE, et al. Lumbar tactile acuity is near identical between sides in healthy pain-free participants. Manual Therapy. 2014a.

Wand BM, Catley MJ, Rabey MI, O'Sullivan PB, O'Connell NE, Smith AJ. Disrupted self-perception in people with chronic low back pain. Further evaluation of The Fremantle Back Awareness Questionnaire. J Pain. 2016.

Wand BM, James M, Abbaszadeh S, George PJ, Formby PM, Smith AJ, et al. Assessing self-perception in patients with chronic low back pain: Development of a back-specific body-perception questionnaire. Journal of back and musculoskeletal rehabilitation. 2014b.

Supplementary Data (e-Supplements):

Appendix 1: Search Strategies for all databases

Search Strategy Medline with search filter for measurement properties

#1 "Sensory acuity"[tw] OR "Sensory perception"[tw] OR "sensory threshold"[tw] OR "Tactile acuity"[tw] OR "Tactile threshold"[tw] OR "tactile perception"[tw] OR "tactile discrimination"[tw] OR "tactual discrimination"[tw] OR "pressure sensitivity"[tw] OR "pressure sensibility"[tw] OR "proprioceptive"[tw] OR "acuity"[tw] OR "touch sensitivity"[tw] OR "tactile sensation*"[tw] OR "tactile sensitivity"[tw] OR "tactile sensibility"[tw] OR "depth-sense threshold"[tw] OR "perception threshold"[tw] OR "Discrimination sensation"[tw] OR "discriminative sensations"[tw] OR "Touch perception"[tw] OR "sensorimotor performance"[tw] OR "sensorimotor competence"[tw] OR "distorted body image"[tw] OR "Body schema"[tw] OR "physical self-awareness"[tw] OR "primary somatosensory cortex"[tw] OR "primary sensory cortex"[tw] OR "sensory-motor incongruence"[tw] OR "S1"[tw] OR "S1 representation"[tw] OR "Cortical reorganisation"[tw] OR "Cortical reorganization"[tw] OR "Consciousness"[tw] OR Neuroimaging[tw] OR "Neuronal plasticity"[tw] OR "cortical body map"[tw] OR "Touch Perception"[MeSH Terms] OR "Touch/physiology"[MeSH Terms] OR "Recognition (Psychology)"[MeSH Terms] OR "Pain Perception"[MeSH Terms] OR "Discrimination Learning"[MeSH Terms] OR "Discrimination (Psychology)/physiology*"[MeSH Terms] OR "Perception/physiology"[MeSH Terms] OR "proprioception/physiology"[MeSH Terms] OR "Pain Threshold"[MeSH Terms] OR "Pattern Recognition, Physiological"[MeSH Terms] OR "Brain mapping"[MeSH Terms]

#2 back pain[tw] OR backache[tw] OR lumbago[tw] OR sciatic[tw] OR sciatica[tw] OR "low back disorder "[tw] OR "low back pain"[tw] OR "Chronic low back pain"[tw] OR "lower back pain"[tw] OR "non-specific low back pain"[tw] OR "NSCLBP"[tw] OR "back injury"[tw] OR "lumbar spine dysfunction"[tw] OR "Back Pain"[MeSH Terms] OR Sciatica[MeSH Terms] OR "Low Back Pain"[MeSH Terms] OR "Low Back Pain/physiopathology*"[MeSH Terms]

#3 Sensorymotor*[tw] OR Sensorimotor*[tw] OR "Sensory-motor*"[tw] OR Sensomotor[tw] OR Sensomotoric[tw] OR "sensori-motor"[tw] OR "sensory-perceptual-motor*"[tw] OR "sensory discrimination"[tw] OR "tactile stimulation"[tw] OR "Tactile assessments"[tw] OR "tactile perceptual tasks"[tw] OR "tactile tests"[tw] OR "sensory tests"[tw] OR "sensory testing"[tw] OR "somatosensory task"[tw] OR "somatosensory testing"[tw] OR "Test"[tw] OR "Testing"[tw] OR task[tw] OR "Somatosensory Cortex"[MeSH Terms] OR "Motor Cortex"[MeSH Terms] OR "Physical Stimulation"[MeSH Terms] OR "Two point discrimination"[tw] OR "Two-point discrimination"[tw] OR "Two-point-discrimination"[tw] OR "two-point thresholds"[tw] OR "TPD threshold"[tw] OR "2-point discrimination"[tw] OR "2-PD"[tw] OR "TPD"[tw] OR "T.P.D"[tw] OR "discrimination threshold" OR Graphaestesia[tw] OR graphesthesia[tw] OR graphestesia[tw] OR "lumbopelvic motor control"[tw] OR "Movement Control"[tw] OR "Lumbopelvic control"[tw] OR "Movement test"[tw] OR "Motor Activity/physiology*"[MeSH Terms] OR "Lumbosacral Region"[MeSH Terms] OR "Motor Skills*"[MeSH Terms] OR "Movement/physiology"[MeSH Terms] OR "Movement Disorders"[MeSH Terms] OR "Body image drawing"[tw] OR "motor imagery"[tw] OR "motor imagery task"[tw] OR "Body schema"[tw] OR "body-perception"[tw] OR "Body image assessment"[tw] OR "Body image perception"[tw] OR "body image"[MeSH Terms] OR "motor imagery"[tw] OR "left/right judgment"[tw]

OR "left/right judgement"[tw] OR "Recognition (Psychology)"[MeSH Terms] OR "Functional Laterality"[MeSH Terms]

#4 instrumentation[sh] OR methods[sh] OR Validation Studies[pt] OR Comparative Study[pt] OR "psychometrics"[MeSH] OR psychometr*[tiab] OR clinimetr*[tw] OR clinometr*[tw] OR "outcome assessment (health care)"[MeSH] OR outcome assessment[tiab] OR outcome measure*[tw] OR "observer variation"[MeSH] OR observer variation[tiab] OR "Health Status Indicators"[MeSH] OR "reproducibility of results"[MeSH] OR repoducib*[tiab] OR "discriminant analysis"[MeSH] OR reliab*[tiab] OR unreliab*[tiab] OR valid*[tiab] OR coefficient[tiab] OR homogeneity[tiab] OR homogeneous[tiab] OR "internal consistency"[tiab] OR (cronbach*[tiab] AND (alpha[tiab] OR alphas[tiab])) OR (item[tiab] AND (correlation*[tiab] OR selection*[tiab] OR reduction*[tiab])) OR agreement[tiab] OR precision[tiab] OR imprecision[tiab] OR "precise values"[tiab] OR test-retest[tiab] OR (test[tiab] AND retest[tiab]) OR (reliab*[tiab] AND (test[tiab] OR retest[tiab])) OR stability[tiab] OR interrater[tiab] OR inter-rater[tiab] OR intrarater[tiab] OR intra-rater[tiab] OR intertester[tiab] OR inter-tester[tiab] OR intratester[tiab] OR intra-tester[tiab] OR interobserver[tiab] OR inter-observer[tiab] OR intraobserver[tiab] OR intra-observer[tiab] OR intertechnician[tiab] OR inter-technician[tiab] OR intratechnician[tiab] OR intra-technician[tiab] OR interexaminer[tiab] OR inter-examiner[tiab] OR intraexaminer[tiab] OR intra-examiner[tiab] OR interassay[tiab] OR inter-assay[tiab] OR intraassay[tiab] OR intra-assay[tiab] OR interindividual[tiab] OR inter-individual[tiab] OR intraindividual[tiab] OR intra-individual[tiab] OR interparticipant[tiab] OR inter-participant[tiab] OR intraparticipant[tiab] OR intra-participant[tiab] OR kappa[tiab] OR kappa's[tiab] OR kappas[tiab] OR repeatab*[tiab] OR ((replicab*[tiab] OR repeated[tiab]) AND (measure[tiab] OR measures[tiab] OR

findings[tiab] OR result[tiab] OR results[tiab] OR test[tiab] OR tests[tiab])) OR generaliza*[tiab] OR generalisa*[tiab] OR concordance[tiab] OR (intraclass[tiab] AND correlation*[tiab]) OR discriminative[tiab] OR "known group"[tiab] OR factor analysis[tiab] OR factor analyses[tiab] OR dimension*[tiab] OR subscale*[tiab] OR (multitrait[tiab] AND scaling[tiab] AND (analysis[tiab] OR analyses[tiab])) OR item discriminant[tiab] OR interscale correlation*[tiab] OR error[tiab] OR errors[tiab] OR "individual variability"[tiab] OR (variability[tiab] AND (analysis[tiab] OR values[tiab])) OR (uncertainty[tiab] AND (measurement[tiab] OR measuring[tiab])) OR "standard error of measurement"[tiab] OR sensitiv*[tiab] OR responsive*[tiab] OR ((minimal[tiab] OR minimally[tiab] OR clinical[tiab] OR clinically[tiab]) AND (important[tiab] OR significant[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR (small*[tiab] AND (real[tiab] OR detectable[tiab]) AND (change[tiab] OR difference[tiab])) OR meaningful change[tiab] OR "ceiling effect"[tiab] OR "floor effect"[tiab] OR "Item response model"[tiab] OR IRT[tiab] OR Rasch[tiab] OR "Differential item functioning"[tiab] OR DIF[tiab] OR "computer adaptive testing"[tiab] OR "item bank"[tiab] OR "cross-cultural equivalence"[tiab]

#5 #1 AND #2 AND #3 AND #4

#6 #5 NOT ("addresses"[pt] OR "biography"[pt] OR "case reports"[pt] OR "comment"[pt] OR "directory"[pt] OR "editorial"[pt] OR "festschrift"[pt] OR "interview"[pt] OR "lectures"[pt] OR "legal cases"[pt] OR "legislation"[pt] OR "letter"[pt] OR "news"[pt] OR "newspaper article"[pt] OR "patient education handout"[pt] OR "popular works"[pt] OR "congresses"[pt] OR "consensus development conference"[pt] OR "consensus development conference, nih"[pt] OR "practice guideline"[pt]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])

Search Strategy Embase

#1 (Sensory acuity or Sensory perception or sensory threshold or Tactile acuity or Tactile threshold or tactile perception or tactile discrimination or tactual discrimination or pressure sensitivity or pressure sensibility or proprioceptive or acuity or touch sensitivity or tactile sensation$ or tactile sensitivity or tactile sensibility or depth-sense threshold or perception threshold or Discrimination sensation or discriminative sensations or Touch perception or sensorimotor performance or sensorimotor competence or distorted body image or Body schema or physical self-awareness or primary somatosensory cortex or primary sensory cortex or sensory-motor incongruence or S1 or S1 representation or Cortical reorganization or Cortical reorganisation or Neuroimaging or Neuronal plasticity or cortical body map).mp. or exp touch/ or exp recognition/ or exp nociception/ or exp proprioception/ or exp pain threshold/ or exp brain mapping/ or exp discrimination learning/ or exp perceptive threshold/

#2 (back pain or backache or lumbago or sciatic or sciatica or low back disorder or low back pain or chronic low back pain or lower back pain or non-specific low back pain or NSCLBP or back injury or lumbar spine dysfunction).mp.

#3 (Sensorymotor* or Sensorimotor* or sensory-motor* or sensori-motor* or Sensomotor or Sensomotoric or sensory-perceptual-motor* or sensory discrimination or tactile stimulation or Tactile assessments or tactile perceptual tasks or tactile tests or sensory tests or sensory testing or somatosensory task or somatosensory testing or test or testing or task or Two point discrimination or two-point thresholds or TPD threshold or 2-point discrimination or 2-PD or TPD or "T.P.D" or discrimination threshold or graphaestesia or graphesthesia or graphestesia or lumbopelvic motor control or movement control or Lumbopelvic control or Movement test or

Lumbosacral Region or Body image drawing or motor imagery or motor imagery task or Body schema or body-perception or Body image assessment or Body image perception or "left/right judgment" or "left/right judgement").mp. or exp motor dysfunction/ or exp "movement (physiology)"/ or exp recognition/ or exp body image/ or exp somatosensory cortex/ or exp motor cortex/ or exp motor activity/ or exp motor performance/

#4 #1 AND #2 AND #3

#5 4 NOT (editorial or letter or conference abstract or conference paper or conference proceeding or conference review).pt. not (exp animal/ not exp human/)

Search Strategy CINAHL

S1 TX Sensory acuity OR TX Sensory perception OR TX sensory threshold OR TX Tactile acuity OR TX Tactile threshold OR TX tactile perception OR TX tactile discrimination OR TX tactual discrimination OR TX pressure sensitivity OR TX pressure sensibility OR TX proprioceptive OR TX acuity OR TX touch sensitivity OR TX tactile sensation* OR TX tactile sensitivity OR TX tactile sensibility OR TX depth-sense threshold OR TX perception threshold OR TX Discrimination sensation OR TX discriminative sensations OR TX Touch perception OR TX sensorimotor performance OR TX sensorimotor competence OR TX distorted body image OR TX Body schema OR TX physical self-awareness OR TX primary somatosensory cortex OR TX primary sensory cortex OR TX sensory-motor incongruence OR TX "S1" OR TX "S1 representation" OR TX Cortical reorganization OR TX Cortical reorganization OR TX Consciousness OR TX Neuroimaging OR TX Neuronal plasticity OR TX cortical body map OR MH "Touch/PH" OR MH "Recognition Psychology" OR MH "Pain+" OR MH "Perception+" OR MH "Perception/PH" OR MH "Proprioception+/PH" OR MH "Pain Threshold" OR MH "Brain Mapping"

S2 TX back pain OR TX backache OR TX lumbago OR TX sciatic OR TX sciatica OR TX low back disorder OR TX low back pain OR TX Chronic low back pain OR TX lower back pain OR TX non-specific low back pain OR TX NSCLBP OR TX back injury OR TX lumbar spine dysfunction OR MH "Back Pain+" OR MH "Sciatica" OR MH "Low Back Pain" OR MH "Low Back Pain/PP"

S3 TX Sensorymotor* OR TX Sensorimotor* OR TX Sensory-motor* OR TX Sensomotor OR TX Sensomotoric OR TX sensori-motor OR TX sensory-perceptual-motor* OR TX sensory discrimination OR TX tactile stimulation OR TX Tactile assessments OR TX tactile perceptual tasks OR TX tactile tests OR TX sensory tests

OR TX sensory testing OR TX somatosensory task OR TX somatosensory testing OR TX Test OR TX Testing OR TX task OR MH "Physical Stimulation+" OR TX Two point discrimination OR TX Two-point discrimination OR TX Two-point-discrimination OR TX two-point thresholds OR TX TPD threshold OR TX 2-point discrimination OR TX 2-PD OR TX TPD OR TX T.P.D OR TX discrimination threshold OR TX Graphaestesia OR TX graphesthesia OR TX graphestesia OR TX lumbopelvic motor control OR TX Movement Control OR TX Lumbopelvic control OR TX Movement test OR MH "Motor Activity+/PH" OR MH "Motor Skills+" OR MH "Movement+/PH" OR MH "Movement Disorders+" OR TX Body image drawing OR TX motor imagery OR TX motor imagery task OR TX Body schema OR TX body-perception OR TX Body image assessment OR TX Body image perception OR MH "Recognition Psychology" OR MH "Body Image+" OR TX motor imagery OR TX left/right judgment OR TX left/right judgement

S4 **S1 AND S2 AND S3**

S5 **S4 NOT** (PT biography OR PT case study OR PT commentary OR PT directories OR PT editorial OR PT interview OR PT legal case OR PT letter OR PT consumer/patient teaching materials OR PT practice guidelines) NOT ((MH "Animals+") NOT (MH "Human"))

Search Strategy CENTRAL

#1 "Sensory acuity" OR "Sensory perception" OR "sensory threshold" OR "Tactile acuity" OR "Tactile threshold" OR "tactile perception" OR "tactile discrimination" OR "tactual discrimination" OR "pressure sensitivity" OR "pressure sensibility" OR proprioceptive OR acuity OR "touch sensitivity" OR "tactile sensation*" OR "tactile sensitivity" OR "tactile sensibility" OR "depth-sense threshold" OR "perception threshold" OR "Discrimination sensation" OR "discriminative sensations" OR "Touch perception" OR "sensorimotor performance" OR "sensorimotor competence" OR "distorted body image" OR "Body schema" OR "physical self-awareness" OR "primary somatosensory cortex" OR "primary sensory cortex" OR "sensory-motor incongruence" OR "S1" OR "S1 representation" OR "Cortical reorganisation" OR "Cortical reorganization" OR Consciousness OR Neuroimaging OR "Neuronal plasticity" OR "cortical body map"

#2 MeSH descriptor: [Touch Perception] explode all trees

#3 MeSH descriptor: [Touch] explode all trees and with qualifier(s): [Physiology - PH]

#4 MeSH descriptor: [Recognition (Psychology)] explode all trees

#5 MeSH descriptor: [Pain Perception] explode all trees

#6 MeSH descriptor: [Discrimination Learning] explode all trees

#7 MeSH descriptor: [Discrimination (Psychology)] explode all trees and with qualifier(s): [Physiology - PH]

#8 MeSH descriptor: [Perception] explode all trees and with qualifier(s): [Physiology - PH]

#9 MeSH descriptor: [Proprioception] explode all trees and with qualifier(s): [Physiology - PH]

#10 MeSH descriptor: [Pain Threshold] explode all trees

#11 MeSH descriptor: [Pattern Recognition, Physiological] explode all trees

#12 MeSH descriptor: [Brain Mapping] explode all trees

#13 #1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12

#14 "back pain" OR backache OR lumbago OR sciatic OR sciatica OR "low back disorder" OR "low back pain" OR "Chronic low back pain" OR "lower back pain" OR "non-specific low back pain" OR "NSCLBP" OR "back injury" OR "lumbar spine dysfunction"

#15 MeSH descriptor: [Back Pain] explode all trees

#16 MeSH descriptor: [Sciatica] explode all trees

#17 MeSH descriptor: [Low Back Pain] explode all trees

#18 MeSH descriptor: [Low Back Pain] explode all trees and with qualifier(s): [Physiopathology - PP]

#19 #14 OR #15 OR #16 OR #17 OR #18

#20 Sensorymotor* OR Sensorimotor* OR "Sensory-motor*" OR Sensomotor OR Sensomotoric OR "sensori-motor" OR "sensory-perceptual-motor*" OR "sensory discrimination" OR "tactile stimulation" OR "Tactile assessments" OR "tactile perceptual tasks" OR "tactile tests" OR "sensory tests" OR "sensory testing" OR "somatosensory task" OR "somatosensory testing" OR "Test" OR "Testing" OR task

OR "Two point discrimination" OR "Two-point discrimination" OR "Two-point-discrimination" OR "two-point thresholds" OR "TPD threshold" OR "2-point discrimination" OR "2-PD" OR "TPD" OR "T.P.D" OR "discrimination threshold" OR Graphaestesia OR graphesthesia OR graphestesia OR "lumbopelvic motor control" OR "Movement Control" OR "Lumbopelvic control" OR "Movement test" OR "Body image drawing" OR "motor imagery" OR "motor imagery task" OR "Body schema" OR "body-perception" OR "Body image assessment" OR "Body image perception" OR "motor imagery" OR "left/right judgment" OR "left/right judgement"

#21 MeSH descriptor: [Somatosensory Cortex] explode all trees

#22 MeSH descriptor: [Motor Cortex] explode all trees

#23 MeSH descriptor: [Physical Stimulation] explode all trees

#24 MeSH descriptor: [Motor Activity] explode all trees and with qualifier(s): [Physiology - PH]

#25 MeSH descriptor: [Lumbosacral region] explode all trees

#26 MeSH descriptor: [Motor Skills] explode all trees

#27 MeSH descriptor: [Movement Disorders] explode all trees

#28 MeSH descriptor: [Body Image] explode all trees

#29 MeSH descriptor: [Functional Laterality] explode all trees

#30 MeSH descriptor: [Movement] explode all trees and with qualifier(s): [Physiology - PH]

#31 MeSH descriptor: [Recognition (Psychology)] explode all trees

#32 #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27 OR #28 OR #29 OR #30 OR #31

#33 #13 AND #19 AND #32

Appendix 2: Individual study data collection protocols (adapted from (Pin, 2014)

Author (Year)	**a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal**	**Patient position**	**Instructions/ Feedback**	**Administration**
Linder et al (2015)	a. Laterality Judgement b. PT c. Program: Recognise Online™ – Difficulty setting "Vanilla" – Images displayed against plain background, randomly rotated at 0°, 90° or 180° to either the left or right distributed equally regarding laterality and rotation. d. 2 instructional images before each part of testing	seated comfortably with elbows at 90°, palms facing downwards, either left or right hand used on keys A and D or left/right arrows	Instructions: Session 1: Oral instruction by PT Session 2 and 3: Written instructions Feedback: NA	60 trunk images; Participants should determine, whether the depicted trunk was moved, laterally, flexed or rotated to the left or right. After 5 s new image if no selection was made.
Bowering et al (2014)	a. Laterality Judgement b. NA c. Program: Recognise Online™ – Images of the back and control images, which contained the back with another body part; "Neutral Images" - displayed against plain background, randomly rotated at +90°, -90° or 180° to either the left or right, distributed equally	Seated on a comfortable chair in front of the computer; hand on keys A and D	Instructions: Written instructions Feedback: Not provided during task	40 images; 2 identical testing blocks with 2 minute break between; after 8 s a new image was presented if no selection was made.

Author (Year)	a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal	Patient position	Instructions/ Feedback	Administration
	regarding laterality and rotation. d. 2 instructional images displaying a left or right hand; Participants had to press the “a” key for left and “d” for right			
Bray and Moseley (2011)	a. Laterality Judgement b. NA c. Program: Recognise™; 28 photographs of a male model in various positions; trunk rotated right between 5° and 90°; photographs were digitally mirrored to construct identical pictures of the same model in various degrees of left rotation ;56 pictures integrated into Recognise™ d. Each trial preceded by practice trial of 80 pictures	Participants positioned themselves so that they were comfortable; Index and middle finger of the dominant hand placed on key “a” for left and “d” for right.	NA	56 photographs randomly displayed; 40 images per trial;2 blocks with 3minute break between; Participants had to sit quietly
Stanton et al (2013)	Laterality Judgement Data obtained from Bray and Moseley (2011)			
Nishigami et al (2015)	a. TPD b. NA	NA	Subjects were instructed to say “one” when they perceived one point and	Bilateral assessment; Caliper position: perpendicular with the spine, transverse process of the

Author (Year)	a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal	Patient position	Instructions/ Feedback	Administration
	c. Plastic ruler with 1 mm precision d. NA		"two" when they perceived two.	most severe pain level and same opposite level centered between the two tips of the caliper Pressure: until first blanching of the skin; Testing Order: Ascending: Starting at 0 mm, 5 mm steps until subjects identified "2 points"; Descending: Starting from 10 cm; 5 mm steps until subjects identified "1 point"; Repetitions: 2 ascending, 2 descending tests; Values of 1 ascending and descending run per side were averaged; Determination of TPD threshold: Side-to side-difference: TPD value higher pain side – TPD value lower pain side
Stanton et al (2013)	a. TPD b. NA c. Plastic caliper ruler d. Sensory testing with monofilaments to assess potential hypoesthesia	NA	NA	Bilateral assessment; Caliper position: horizontally on both sides of the back; between the first lumbar vertebra and iliac crest; Pressure: Supra-sensory threshold; non-noxious; Testing Order: Ascending: Starting from 10 mm; 5 mm increments Descending: Starting from 100 mm, 5 mm increments; Repetitions: 1 ascending, 1 descending test;

Author (Year)	a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal	Patient position	Instructions/ Feedback	Administration
				Determination of TPD threshold: Mean of ascending/descending test per side; Average of right and left mean actual TPD threshold
Luomajoki et al (2011)	a. TPD b. Physiotherapist c. Plastic caliper ruler d. NA	NA	NA	Bilateral Assessment; Caliper position: Horizontally and vertically between the first lumbar vertebra and iliac crest; Pressure: NA; Testing Order: Ascending: Starting from 10 mm; 5 mm increments Descending: Starting from 100 mm, 5 mm increments; Repetitions: 1 ascending, 1 descending test; Catch trials to prevent guessing (expanding the calipers instead of contracting or vice versa); Determination of TPD threshold: Average of ascending/descending test
Wand et al (2010)	a. TPD b. NA c. Lafayette two-point aesthesiometer, (Lafayette Instruments, Lafayette, IN, USA), 1 mm precision d. NA	Positioned comfortably in prone lying on an examination table with back exposed Pillow under stomach to flatten the lumbar spine	Subjects were instructed to say 'one' when they felt one point and 'two' when they felt two points.	Bilateral Assessment; Caliper position: Parallel to the spine; L3 transverse process in the centre of the caliper; Pressure: until first blanching of the skin; Testing Order: Ascending: Starting at 0 mm, 2 mm steps until subjects identified "2 points"; Descending: Starting from a start point "well

Author (Year)	a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal	Patient position	Instructions/ Feedback	Administration
		and to standardized lumbar position.		above the initial ascending threshold value" ; 2 mm steps until patients identified "1 point"; Repetitions: NA; Determination of TPD threshold: Testing continued around initial values using ascending and descending sequences until a consistent response was obtained; Catch trials used to prevent guessing
Moseley (2008)	a. TPD b. NA c. Mechanical caliper with 1mm precision d. NA	prone	Subjects were instructed to say 'one', when one point was felt, 'two', when two points were felt.	Bilateral Assessment; caliper position: 16 levels from the fourth thoracic vertebra to the bottom of the gluteal folds; Medial point was 1, 2, 3 cm from midline; Pressure: until first blanching of the skin; Testing Order: Ascending: Starting at 0 mm, gradually increasing until patients identified "2 points"; Descending: NA; Level was randomised and counterbalanced; side was alternated until 6 measures (3 each side) were obtained; Repetitions: 3 measures per level and side; Determination of TPD threshold: Average of 1 ascending and descending series

Author (Year)	a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal	Patient position	Instructions/ Feedback	Administration
				Average of 3 measures per level and side were used for analysis
Luomajoki et al (2008)	a. MCT b. PTs, on average 7 years working experience; not blinded to subject's group c. NA d. Subjects did not know the tests before	Depending on test starting position; Subjects wore underwear to allow inspection of the entire spine, hips and lower extremities	Verbal standardized instructions specified for each test; If the subject did not understand how to perform the test, the examiner explained and demonstrated it again;	Testing order: Waiter's bow, Pelvic Tilt, One leg stance, Single knee extension, Quadruped position: • Rocking backward • Rocking forward Prone lying active knee flexion; Repetitions: 3 trials permitted; Rating Protocol: Clear movement dysfunction was rated as "not correct" and scored "1"; correct movements were scored"0"; If the movement control improved by instruction and correction, it was considered not a relevant movement dysfunction.
Luomajoki and Moseley (2011)	a. TPD b. Trained PT c. NA d. NA	Depending on test starting position;	Subjects were given a picture in which a model demonstrated the target alignment of the pelvis and lumbar spine as a reference	Testing Order: Battery of six tests, referencing Luomajoki et al (2008); Repetitions: NA; Reference to Luomajoki et al (2008); Rating Protocol: Scores ranging from 6-0; 6 demonstrated the poorest movement control performance

Author (Year)	a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal	Patient position	Instructions/ Feedback	Administration
Wand et al (2010)	a. Graphesthesia b. NA c. Blunt end of monofilament d. Subjects were first shown a wall chart o upper case alphabet; letters and were instructed that this would be the way the letters would be drawn. They were then shown a diagram of the lumbar spine depicting the orientation and location of the letters	positioned comfortably in prone lying on an examination table with back exposed Pillow under stomach to flatten the lumbar spine and to standardised lumbar position.	Subjects were asked to identify the letters drawn on the back	Testing order: Letters were drawn on three sites centred on the tips of the L1, L3 and L5 transverse processes; did not extend across the midline; the height of the letters was such that there was no overlap in the area of skin in which the letters were drawn between the 3 sites; 20 random letters at each level of the 3 sites; 3 sites were tested in random order; Error rate out of 60 was calculated for each side of the back
Moseley et al (2008)	a. BID b. NA c. a line drawing showing the posterior surface of the back with only the top and bottom of the picture drawn d. NA	Subjects stood in front of a waist high bench	'Concentrate on your back. Add to this drawing by following the outline of your own back as you track it in your mind. Concentrate on where you feel your back to be. Also draw in the vertebra that you can feel. Do this without touching your back. Your drawing should relate to your own sense of your back. Don't draw any part you can't sense.	On request after instruction

Author (Year)	**a. Instrument b. Assessor (Profession) c. Equipment d. Rehearsal**	**Patient position**	**Instructions/ Feedback**	**Administration**
			Do not draw what you think your back looks like – draw what it feels like."	
Nishigami et al (2015)	a. BID b. NA c. a line drawing showing the posterior surface of the back with only the top and bottom of the picture drawn (acc. to Moseley et al (2008)) d. NA	Subjects were asked to sit in a chair	"Concentrate on your back. Add to this drawing by following the outline of your own back as you track it in your mind. Concentrate on where you feel your back to be. Also draw in the vertebra that you can feel. Do this without touching your back. Do not draw any part you cannot sense. Do not draw what you think your back looks like- draw what it feels like."	On request after instruction
Wand et al (2014)	a. FreBAQ b. NA c. NA d. NA	NA	Written instructions	First test: On site; Second test: Take home copy; filled out and posted one week later
Legend: NA = information not available; PT = physiotherapist; TPD= Two-Point Discrimination; MCT = Movement Control Test; BID= body image drawings; FreBAQ= Fremantle Back Awareness Questionnaire				

Appendix 3: Reviewer Agreement

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
TPD Known-Groups Validity					
Luomajoki and Moseley (2011)	1	2	2	1	2
	2	2	1	0	1
	3	1	2	1	2
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	0	2
	7	n.a. †	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	3	3	0	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	10	2	3	1	2
Stanton et al (2013)	1	2	2	1	2
	2	1	1	1	1
	3	1	1	1	1
	4	3	2	0	2
	5	3	2	0	2
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	0	1	0	1
	10	3	3	1	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
Nishigami et al (2015)	1	2	2	1	2
	2	1	1	1	1
	3	2	2	1	2
	4	3	1	0	1
	5	3	2	0	2
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	1	1	1	1
	10	3	3	1	3
Moseley (2008)	1	2	2	1	2
	2	2	2	1	2

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	3	0	0	1	0
	4	2	1	1	1
	5	3	2	0	2
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	1	1	1	1
	10	2	0	0	0
Wand et al (2010)	1	2	2	1	2
	2	3	1	0	1
	3	1	1	1	1
	4	1	3	0	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	5	3	3	1	3
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	3	1	0	1
	10	3	3	0	3
TPD Convergent Validity					
Luomajoki and Moseley (2011)	1	2	2	1	2
	2	2	1	0	1
	3	2	2	1	2
	4	3	3	1	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	5	3	3	1	3
	6	2	2	1	2
	7	3	2	0	3
	8	1	1	1	1
	9	3	1	0	1
	10	2	3	0	2
Stanton et al (2013)	1	2	2	1	2
	2	1	1	1	1
	3	0	0	1	0
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	1	2

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	7	3	3	1	3
	8	1	1	1	1
	9	0	1	0	0
	10	3	3	1	3
Wand et al (2010)	1	2	2	1	2
	2	3	1	0	1
	3	0	0	1	0
	4	1	3	0	3
	5	3	3	1	3
	6	2	2	1	2
	7	3	3	1	3
	8	1	0	0	0

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	9	3	1	0	1
	10	3	3	1	3
MCT Known-Groups Validity					
Luomajoki and Moseley (2011)	1	2	2	1	2
	2	2	1	0	1
	3	1	2	0	2
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	9	3	1	0	1
	10	2	3	0	2
Luomajoki et al (2008)	1	2	2	1	2
	2	1	1	1	1
	3	3	3	1	3
	4	2	2	1	2
	5	2	2	1	2
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	1	1	1	1

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	10	3	2	0	2
MCT Convergent Validity					
Luomajoki and Moseley (2011)	1	2	2	1	2
	2	2	1	0	1
	3	2	2	1	2
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	1	2
	7	3	2	0	3
	8	1	1	1	1
	9	3	1	0	1

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	10	2	3	0	2
Graphesthesia Known-Groups Validity					
Wand et al (2010)	1	2	2	1	2
	2	3	1	0	1
	3	1	1	1	1
	4	1	3	0	3
	5	3	3	1	3
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	9	3	1	0	1
	10	3	3	1	3
Graphesthesia Convergent Validity					
Wand et al (2010)	1	2	2	1	2
	2	3	1	0	1
	3	2	2	1	2
	4	1	3	0	3
	5	3	3	1	3
	6	2	2	1	2
	7	3	3	1	3
	8	1	1	1	1

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	9	3	1	0	1
	10	3	3	1	3
Laterality Judgement Known-Groups Validity					
Stanton et al (2013)	1	2	2	1	2
	2	1	1	1	1
	3	0	0	1	0
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	8	n.a.	n.a.	1	n.a.
	9	0	1	0	0
	10	3	3	1	3
Bowering et al (2014)	1	2	3	0	3
	2	3	3	1	3
	3	3	3	1	3
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	3	3	1	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	10	3	3	1	3
Bray and Moseley (2011)	1	2	2	1	2
	2	2	1	0	1
	3	0	1	0	1
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	1	1	1	1
	10	3	3	1	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
Linder et al (2015)	1	2	2	1	2
	2	1	1	1	1
	3	1	2	0	2
	4	1	1	1	1
	5	2	2	1	2
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	3	3	1	3
	10	3	3	1	3
Laterality Judgement					

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
Convergent Validity					
Stanton et al (2013)	1	2	2	1	2
	2	1	1	1	0
	3	0	0	1	0
	4	3	3	1	3
	5	3	3	1	3
	6	2	2	1	2
	7	3	3	1	3
	8	1	1	1	1
	9	0	1	0	0
	10	3	3	1	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
Laterality Judgement Reliability					
Linder et al (2015)	1	2	2	1	2
	2	1	1	1	1
	3	1	0	0	0
	4	3	3	1	3
	5	2	3	0	3
	6	3	3	1	3
	7	2	1	0	2
	8	3	3	1	3
	9	0	1	0	1
	10	3	1	0	1

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	11	3	2	0	3
	12	n.a.	n.a.	1	n.a.
	13	n.a.	n.a.	1	n.a.
	14	n.a.	n.a.	1	n.a.
Bray and Moseley (2011)	1	2	2	1	2
	2	2	1	0	1
	3	0	0	1	0
	4	3	3	1	3
	5	3	3	1	3
	6	3	3	1	3
	7	2	2	1	2
	8	3	3	1	3

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	9	2	2	1	2
	10	1	1	1	1
	11	2	2	1	2
	12	n.a.	n.a.	1	n.a.
	13	n.a.	n.a.	1	n.a.
	14	n.a.	n.a.	1	n.a.
Laterality Judgement Measurement Error					
Linder et al (2015)	1	2	2	1	2
	2	1	1	1	1
	3	2	0	0	0

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	4	3	3	1	3
	5	3	3	1	3
	6	3	3	1	3
	7	1	2	0	2
	8	3	3	1	3
	9	1	1	1	1
	10	1	1	1	1
	11	2	2	1	2
BIDs Known Groups Validity					
Nishigami et al (2015)	1	2	2	1	2
	2	1	1	1	1

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	3	2	2	1	2
	4	3	1	0	1
	5	3	2	0	2
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	1	1	1	1
	10	3	3	1	3
Moseley (2008)	1	2	2	1	2
	2	2	2	1	2
	3	0	0	1	0
	4	2	1	0	1
	5	3	2	0	2

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	6	2	2	1	2
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	1	1	1	1
	10	2	0	0	0
FreBaQ Known Groups Validity					
Wand et al (2014)	1	2	2	1	2
	2	1	1	1	1
	3	3	3	1	3
	4	1	1	1	1

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	5	2	2	1	2
	6	2	2	1	3
	7	n.a.	n.a.	1	n.a.
	8	n.a.	n.a.	1	n.a.
	9	3	3	1	3
	10	3	3	1	3
FreBaQ Reliability					
Wand et al (2014)	1	2	2	1	2
	2	1	1	1	1
	3	0	0	1	0
	4	3	3	1	3
	5	2	2	1	2

Author (year)	COSMIN Item	Methodological Quality Rating Reviewer 1	Methodological Quality Rating Reviewer 2	Rater Agreement 1=agreement 0=disagreement	Final Methodological Quality Rating (consensus)
	6	3	3	1	3
	7	0	2	0	0
	8	3	3	1	3
	9	1	1	1	1
	10	3	1	0	1
	11	3	3	1	3
	12	n.a.	n.a.	1	n.a.
	13	n.a.	n.a.	1	n.a.
	14	n.a.	n.a.	1	n.a.

Legend: TPD= Two-Point Discrimination; n.a. = not applicable; MCT= Movement Control Test; BIDs=Body Image Drawing; FreBaQ=Fremantle Back Awareness Questionnaire; Reviewer 1=KE; Reviewer 2=CR;
Coding methodological quality rating: 0 = poor; 1=fair; 2=good; 3=excellent

Appendix 4: Summary of TPD measurement properties (known-groups and convergent validity)

Author (Year)	**a) Instrument** **b) Design** **c) n**	**Measurement Property Result**		**Mean Difference (95%CI) [p-value]**	**COSMIN Score**
		Mean (SD) [mm]			
		CLBP	HC		
Luomajoki and Moseley (2011)	a) TPD b) Known-Groups Validity c) CLBP = 45 HC = 45	61 (13)	44 (10)	**17*** (12.14 to 21.86) [p < 0.01]	fair
Stanton et al (2013)	a) TPD b) Known-Groups Validity c) CLBP =17 HC =18	59.8 (11.7)	45.3 (5.1)	**14.50** (8.34 to 20.65) [p < 0.0001]	fair
		Related Construct			
	a) TPD b) Convergent Validity c) CLBP = 17	Laterality Reconstruction Accuracy (%): β (95%CI) -0.6 (-0.80 to -0.43)		not stated	poor
		CLBP Mean (SD) [mm]	HC Mean (SD) [mm]		
Moseley (2008)	a) TPD b) Known-Groups Validity c) CLBP = 6 HC = 10	48.83 (1.83)	47 (8)	1.83 (-5.38 to 9.037) [p = 0.59]	poor
Wand et al (2010)	a) TPD b) Known-Groups Validity	62.0 (21.6)	44.2 (13.7)	**17.80** (5.90 to 29.70) [p = 0.0045]	fair

Author (Year)	**a) Instrument** **b) Design** **c) n**	**Measurement Property Result**		**Mean Difference (95%CI) [p-value]**	**COSMIN Score**
	c) CLBP = 19 HC = 19				
Nishigami et al (2015)	a) TPD (mean difference between sides) b) Known-Groups Validity c) CLBP = 42 HC = 17	normal BI: 4.5 (5.5) expanded BI: 13.3 (6.8) shrink: 9.4 (7.0)	5.5 (3.8)	CLBP normal BI - HC: -1.00 (-4.27 to 2.27) [p = 0.54] CLBP expanded BI - HC: **7.80** (3.75 to 11.85) [p = 0.0005] shrink BI - HC: 3.90 (-0.23 to 8.03) [p = 0.06]	Fair
TPD=Two-Point Discrimination; SD=Standard Deviation; 95%CI= 95%Confidence Interval; CLBP=Chronic Low Back Pain; HC=Healthy Controls; BI= Body Image; *Bold figures indicate statistically significant differences					

Appendix 5: Summary of Laterality Judgement measurement properties (known-groups and convergent validity, reliability and measurement error)

Author (Year)	**a) Instrument b) Design c) n**	**Measurement Property**		**Mean Difference (95%CI) [p-value]**	**COSMIN Score**
		Result Mean (SD)			
		CLBP	**HC**		
Linder et al (2015)	a) Laterality Judgement b) Known-Groups Validity c) CLBP = 30 HC = 30	ACC Right Trunk Rotation Mean (SD): S1: 88.7 (8.0) S2: 90.5 (9.8) S3: 90.6 (8.3) ACC Left Trunk Rotation Mean (SD): S1: 89.0 (9.1) S2: 89.1 (11.2) S3: 92.7 (7.4) RT right Trunk Rotation Mean (SD): S1: 1.89 (0.37) S2: 1.60 (0.34) S3: 1.57 (0.26) RT left Rotation Mean (SD): S1: 1.93 (0.43)	ACC Right Trunk Rotation Mean (SD): S1: 87.8 (20.2) S2: 87.9 (12.8) S3: 91.8 (6.6) ACC Left Trunk Rotation Mean (SD): S1:86.0 (10.4) S2: 89.0 (11.4) S3: 90.4 (7.9) RT right Trunk Rotation Mean (SD): S1: 2.01 (0.52) S2: 1.80 (0.53) S3:1.54 (0.34) RT left Rotation Mean (SD): S1: 2.01 (0.55)	ACC Right Trunk Rotation S1: 0.90 (-3.84 to 5.64) [p=0.90] S2: 2.600 (3.29 to 8.49) [p=0.38] S3: -1.20 (-5.08 to 2.68) [p=0.54] ACC Left Trunk Rotation: S1: 3.00 (-2.05 to 8.05) [p=0.24] S2: 0.10 (5.74 to 5.94) [p=0.97] S3: 2.30 (-1.66 to 6.26) [p=0.25]	fair

Author (Year)	a) Instrument b) Design c) n	Measurement Property		Mean Difference (95%CI) [p-value]	COSMIN Score
		Result Mean (SD)			
		CLBP	**HC**		
		S2:1.64 (0.40) S3: 1.57 (0.33)	S2: 1.72 (0.47) S3: 1.55 (0.42)	RT right Trunk Rotation S1: -0.08 (-0.34 to 0.176) [p= 0.53] S2: -0.08 (0.31 to 0.15) [p= 0.50] S3: 0.02 (-0.18 to 0.22) [p= 0.84] RT left Rotation Mean : S1: -0.08 (-0.34 to 0.18) [p= 0.53] S2: -0.08 (-0.31 to 0.15) [p= 0.48] S3: 0.02 (-0.18 to 0.22) [p= 0.84]	

Author (Year)	**a) Instrument b) Design c) n**	**Measurement Property**		**Mean Difference (95%CI) [p-value]**	**COSMIN Score**
		Result Mean (SD)			
		CLBP	**HC**		
Bowering et al (2014)*	a) Laterality Judgement b) Known-Groups Validity c) Current back pain = 117 History of back pain = 462 HC = 429	RT Trunk Rotation 1.89 (0.19) ACC Trunk Rotation: 76.5 (3.4)	RT Trunk Rotation 1.74 (0.07) ACC Trunk Rotation: 85.9 (0.08)	RT Trunk Rotation: **0.14†** (0.11 to 0.17) [p < 0.0001] ACC Trunk Rotation: **-9.40** (-10.02 to -8.78) [p < 0.0001]	good
Bray and Moseley (2011)	a) Laterality Judgement b) Known-Groups Validity c) CLBP = 21 HC = 14	RT Trunk Rotation 2.4 (0.35) ACC Trunk Rotation bilateral pain: 53.4 (19.55) unilateral pain: 67.2 (15.27)	RT Trunk Rotation: 2.4 (0.33) ACC Trunk Rotation: 87 (19.92)	RT Trunk Rotation 0 (-0.24 to 0.24) [p=0] ACC Trunk Rotation: bilateral pain - HC: **-33.60** (-47.43 to -19.77) [p < 0.0001] unilateral pain - HC: **-19.80** (-31.91 to -7.69) [p = 0.0022]	fair

Author (Year)	**a) Instrument b) Design c) n**	**Measurement Property**	**Mean Difference (95%CI) [p-value]**	**COSMIN Score**
		Result Mean (SD)		
		CLBP | **HC**		
		Related Construct		
Stanton et al (2013)	a) Laterality Judgement b) Convergent Validity c) CLBP = 17	TPD: β (95%CI) -0.6 (-0.80 to -0.43)	n.a.	poor
			Result Reliability (95%CI)	
Bray and Moseley (2011)	a) Laterality Judgement b) Test-Retest Reliability: Mean Time interval (Range) [days]: 1 (1-7) c) CLBP = 10		RT Trunk Rotation: $ICC_{2,1}$= 0.872 (0.731 - 0.951) ACC Trunk Rotation: $ICC_{2,1}$= 0.920 (0.831 - 0.970)	poor

Author (Year)	**a) Instrument b) Design c) n**	**Measurement Property**		**Mean Difference (95%CI) [p-value]**	**COSMIN Score**
		Result Mean (SD)			
		CLBP	**HC**		
				Result Reliability (95%CI)	
Linder et al (2015)	a) Laterality Judgement b) Test-Retest Reliability Mean Time interval (Range) [days]: Session 1 to 2: 2.2 (2-5) Session 2-3: 2.4 (1-11) c) Session 1 and 2: CLBP = 25 Session 2 and 3: CLBP = 22			Session 1 to 2: RT Trunk Rotation: ICC2,1= 0.51 (0.15 - 0.75) ACC Trunk Rotation: ICC2,1= 0.71 (0.44 - 0.86) Session 2 to 3: RT Trunk Rotation: ICC2,1= 0.91 (0.79 - 0.96) ACC Trunk Rotation: ICC2,1= 0.69 (0.39 - 0.86)	poor

Author (Year)	**a) Instrument b) Design c) n**	**Measurement Property**		**Mean Difference (95%CI) [p-value]**	**COSMIN Score**
		Result Mean (SD)			
		CLBP	**HC**		
				Coefficient of Variation (95%CI):	
Linder et al (2015)	a) Laterality Judgement b) Measurement Error[ŧ] c) CLBP =22			Session 1 to 2: RT Trunk Rotation: 19.6 (13.79-25.64) ACC Trunk Rotation: 6.46 (4.52-8.32) Session 2 to 3: RT Trunk Rotation: 6.23 (4.35-8.14) ACC Trunk Rotation: 6.77 (4.72-8.86)	poor
ACC= Accuracy; RT= Reaction Time; S1,S2, S3 = Session 1,2,3; SD=Standard Deviation; $ICC_{2,1}$ = Intraclass Correlation Coefficient; two-way random model; 95%CI= 95%Confidence Interval; CLBP=Chronic Low Back Pain; HC=Healthy Controls; TPD= Two-Point Discrimination; n.a.= not applicable;*Data from Bowering et al (2014) were derived via DigitizeIt® and p-values were calculated online via GraphPad Quick Calcs software; ǂBold figures indicate statistically significant differences; [ŧ]= minimal important change (MIC) data was not provide					

Appendix 6: Summary of Movement Control Test known-groups validity

Author (Year)	**a) Instrument b) Design c) n**	**Measurement Property Result**			
		Result Mean (SD) [0-6]		Mean Difference (95%CI) [p-value]	COSMIN Score
		CLBP	HC		
Luomajoki and Moseley (2011)	a) Movement Control Test b) Known-Groups Validity c) CLBP =45 HC = 45	3 (1.1)	1 (1.3)	**2.00*** (1.50 to 2.50) [p < 0.0001]	fair
Luomajoki et al (2008)	a) Movement Control Test b) Known-Groups Validity c) LBP =102 CLBP=46 HC = 102	CLBP 2.37 (1.34)	HC 0.75 (1.03)	**1.62** (1.22 to 2.02) [p < 0.0001]	fair
SD= Standard Deviation; CLBP= chronic low back pain; HC= healthy controls; 95%CI = 95% Confidence Interval; [0-6] = six tests are included in the test battery, the higher the score the worse the test performance; *Bold figures indicate statistically significant differences					

Appendix 7: Summary of Graphesthesia known-groups validity

Author (Year)	**d) Instrument** **e) Design** **f) n**	**Measurement Property Result**			
		Result Mean (SD) [Error Rate]		Mean Difference (95%CI) [p-value]	COSMIN Score
		CLBP	HC		
Wand et al (2010)	a) Graphesthesia b) Known-Groups Validity c) CLBP =19 HC = 19	25.5 (8.0)	19.3 (6.8)	**6.1*** (1.3 to 11.0) [p = 0.01]	fair
SD= Standard Deviation; 95%CI= 95% Confidence Interval; CLBP=Chronic low back pain; HC=Healthy Controls; *Bold figures indicate statistically significant differences					

Appendix 8: Summary of FreBaQ known-groups validity and reliability

Author (Year)	**a) Instrument** **b) Design** **c) n**	**Measurement Property** **Result**		**Mean Difference (95%CI)** **[p-value]**	**COSMIN Score**
		Mean [0-36] *Median (Range)*			
		CLBP	HC		
Wand et al (2014)	a) FreBAQ b) Known-Groups Validity c) CLBP =51 HC = 51	10.8 *11 (0-26)*	0.5 *0 (0-6)*	Mann-Whitney Test **11*** [p < 0.001]	fair
				Result Reliability (95%CI)	
Wand et al (2014)	a) FreBAQ b) Test-Retest Reliability (Mean Time interval: 1 week) c) CLBP =26			$ICC_{2,1}$= 0.652 (0.307-0.848) (Agreement) $ICC_{2,1}$ = 0.667 (0.317-0.857) (Consistency)	poor

FreBAQ = Fremantle Back Awareness Questionnaire; CLBP= Chronic low back pain; HC= Healthy Controls; $ICC_{2.1}$= intra class correlation coefficient (two-way random effect model with single measures); MIC=minimal important change
*Bold figures indicate statistically significant differences