Clinimetric properties of sacroiliac joint mobility tests

Klerx, S. P.; Pool, J. J.M.; Coppieters, M. W.; Mollema, E. J.; Pool-Goudzwaard, A. L.

published in
Musculoskeletal Science and Practice
2020

DOI (link to publisher)
10.1016/j.msksp.2019.102090

document version
Publisher's PDF, also known as Version of record

Musculoskeletal Science and Practice 48 (2020) 102090

Contents lists available at ScienceDirect

Musculoskeletal Science and Practice

journal homepage: www.elsevier.com/locate/msksp

Clinimetric properties of sacroiliac joint mobility tests: A systematic review

S.P. Klerx [a], J.J.M. Pool [a], M.W. Coppieters [b,c], E.J. Mollema [a], A.L. Pool-Goudzwaard [c,d,*]

[a] HU University of Applied Sciences Utrecht, Institute of Movement Sciences, Utrecht, the Netherlands
[b] The Hopkins Centre, Menzies Health Institute Queensland, Griffith University, Brisbane and Gold Coast, Australia
[c] Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam Movement Sciences, Amsterdam, the Netherlands
[d] SOMT University of Physiotherapy, Amersfoort, the Netherlands

ARTICLE INFO

Keywords:
Rehabilitation
Musculoskeletal health
Disability
Manual therapy
Physiotherapy
Low back pain
Pelvic pain

ABSTRACT

Background: Previous systematic reviews revealed poor reliability and validity for sacroiliac joint (SIJ) mobility tests. However, these reviews were published nearly 20 years ago and recent evidence has not yet been summarised.
Objectives: To conduct an up-to-date systematic review to verify whether recommendations regarding the clinical use of SIJ mobility tests should be revised.
Study design: Systematic review.
Method: The literature was searched for relevant articles via 5 electronic databases. The review was conducted according to the PRISMA guidelines. COSMIN checklists were used to appraise the methodological quality. Studies were included if they had at least fair methodology and reported clinimetric properties of SIJ mobility tests performed in adult patients with non-specific low back pain, pelvic (girdle) pain and/or SIJ pain. Only tests that can be performed in a clinical setting were considered.
Results: Twelve relevant articles were identified, of which three were of sufficient methodological quality. These three studies evaluated the reliability of eight SIJ mobility tests and one test cluster. For the majority of individual tests, the intertester reliability showed slight to fair agreement. Although some tests and one test cluster had higher reliability, the confidence intervals around most reliability estimates were large. Furthermore, there were no validity studies of sufficient methodological quality.
Conclusion: Considering the low and/or imprecise reliability estimates, the absence of high-quality diagnostic accuracy studies, and the uncertainty regarding the construct these tests aim to measure, this review supports the previous recommendations that the use of SIJ mobility tests in clinical practice is problematic.

1. Introduction

Sacroiliac joint (SIJ) pain is described as a form of mechanical low back pain and is reported to affect between 15 and 30% of individuals with chronic, non-radicular low back pain (Irwin et al., 2007, Cohen et al., 2013, Laslett et al., 2005, Van der Wurff et al., 2006). The reported prevalence rates vary depending on the population studied and diagnostic criteria (Cohen et al., 2013). SIJ dysfunction is a broad term which refers to abnormal anatomy and/or biomechanical function of the SIJ. It may result from trauma, degenerative changes or inflammation, or may be pregnancy-related, and is believed, by some clinicians and researchers, to result in either hypermobility or hypomobility of the SIJ (Dreyfuss et al., 1994; van der Wurff et al., 2000a; Riddle and Freburger, 2002). In the literature, two categories of clinical tests to assess the SIJ can be distinguished: (1) pain provocation tests and (2) dysfunction (or mobility) tests. With respect to pain provocation tests, systematic reviews revealed poor reliability and validity (van der Wurff et al., 2000a, 2000b). One review (Hancock et al., 2007) showed better diagnostic accuracy values but still concluded that "the usefulness of these tests in clinic still remains unclear". A review on the diagnostic validity of the International Association for the Study of Pain (IASP) criteria to diagnose SIJ pain concluded that the thigh thrust test, the compression test, and three or more positive stress tests had discriminative power to diagnose SIJ pain. Considering the lack of a criterion standard for SIJ pain, the diagnostic validity of tests related to the IASP criteria for SIJ pain should be considered with care (Szadek et al., 2009). SIJ dysfunction tests mainly focus on altered SIJ mobility. However, SIJ movement and its clinical diagnosis remain highly

* Corresponding author. Department of Human Movement Sciences (Room MF 468) Faculty of Behavioural and Movement Sciences Vrije Universiteit Amsterdam, Van der Boechorststraat 7, 1081 BT, Amsterdam, the Netherlands.
E-mail address: a.l.pool-goudzwaard@vu.nl (A.L. Pool-Goudzwaard).

https://doi.org/10.1016/j.msksp.2019.102090
Received 19 June 2019; Received in revised form 23 September 2019; Accepted 8 November 2019
Available online 9 November 2019

controversial. Some authors report that SIJ movement can be palpated (Hungerford et al., 2007), whereas others state that there is virtually no SIJ movement (Sturesson et al., 1989, Sturesson et al., 2000a, 2000b, Godee et al., 2008, Kibsgard, 2015).

Systematic reviews for SIJ mobility tests identified a lack of reliable and valid tests (van der Wurff et al., 2000a, 2000b). The reviews concluded that it is questionable whether any SIJ mobility test would be of any value in clinical practice (van der Wurff et al., 2000a, 2000b). However, the latest reviews have been published nearly 20 years ago, and recent research has not yet been considered in a review. Moreover, new tests continue to be proposed (Shimpi et al., 2018). The aim of this study was therefore to conduct an up-to-date systematic review to verify whether recommendations regarding the use of SIJ mobility tests should be revised regarding clinimetric properties, including reliability, validity and responsiveness, in people with non-specific low back pain, pelvic girdle pain and/or SIJ pain.

2. Methods

2.1. Data sources and searches

The review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2019). The literature was searched for relevant articles from inception to May 2019 via 5 electronic databases: PubMed, CINAHL, Embase, SPORTDiscus and Cochrane Library. The search strings for each database were developed in collaboration with a medical information specialist (MEJ) and are listed in Appendix 1. No limit was set for year of publication or language. All databases were searched separately. The identified references were imported in Legacy RefWorks and duplicates were removed. The reference lists of eligible articles were checked to retrieve additional relevant studies.

2.2. Study selection

Two reviewers (SP and AP) independently selected the studies. The articles were first screened based on title and abstract, and then on full text. In case of disagreement, a third reviewer (JP) was consulted. Studies were included if they reported clinimetric properties of SIJ mobility tests performed in adult patients (i.e., 18 years and older) with non-specific low back pain, pelvic girdle pain and/or SIJ pain. Only tests that can be performed in a clinical setting were considered. Technical tests that could not be performed in clinical practice, such as radiostereometric analysis or rasterstereography were excluded. Studies that evaluated SIJ mobility tests in patients with a specific disorder, such as ankylosing spondylitis, were excluded. Systematic reviews were also excluded, although their reference lists were screened for relevant papers. Fig. 1 provides the flow diagram of the study.

2.3. Quality assessment and data extraction

Quality assessment and data extraction were conducted according to the 'Consensus-Based Standards for the Selection of Health Measurement Instruments' (COSMIN) Protocol for Systematic Reviews of Measurement Properties (Mokkink et al., 2016). The selected articles were

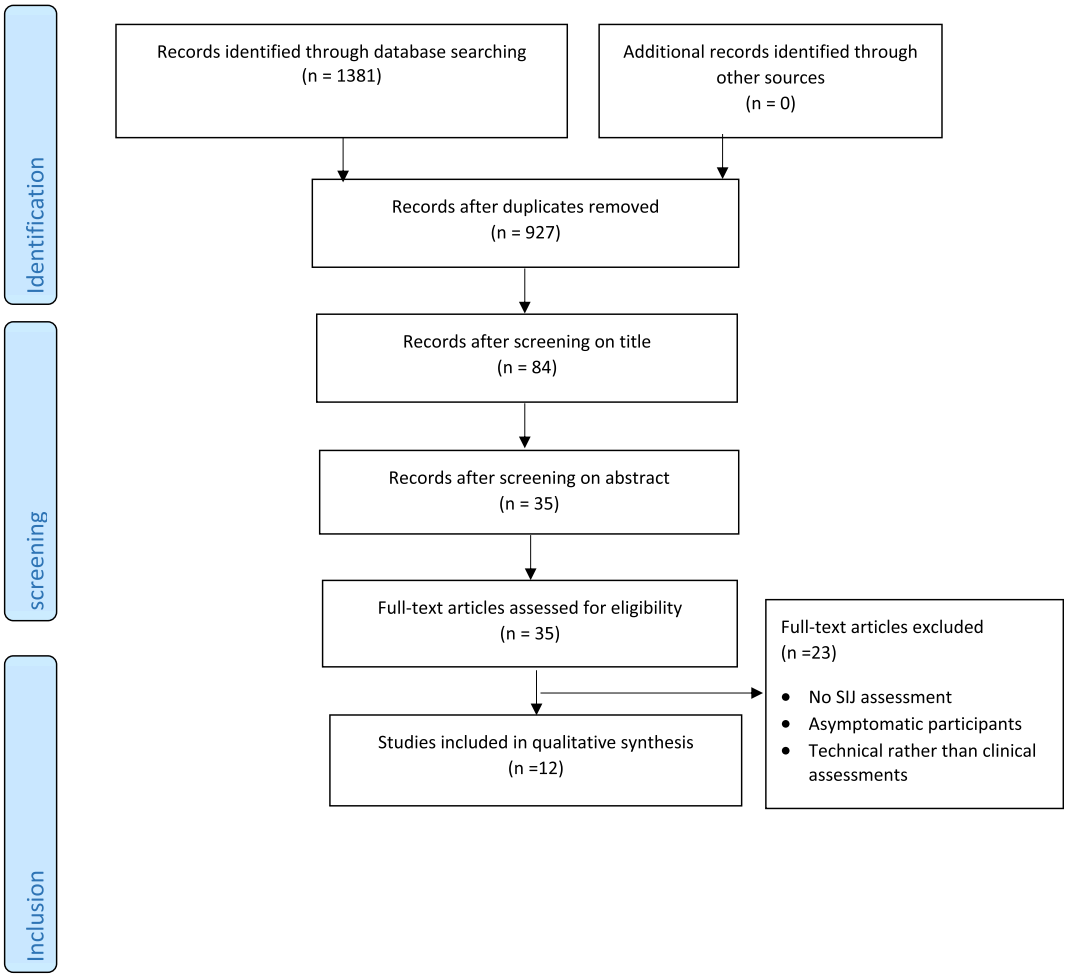

Fig. 1. PRISMA flow diagram showing the study selection process.

independently rated for quality by two investigators (SP and AP) using the COSMIN score-card (Mokkink et al., 2011). In case of disagreement, a third investigator (JP) was consulted to resolve the disagreement.

Various measurement properties (such as reliability, content validity, structural validity, criterion validity and responsiveness) can be assessed with the same COSMIN appraisal tool (Mokkink et al., 2011). Each measurement property is rated on several items. Each item is rated on a 4-point ordinal scale (excellent, good, fair, poor). The overall quality score is obtained by taking the lowest rating on any item scored. Only studies with a fair or higher overall quality score were selected for data extraction and data pooling. Since the COSMIN checklist is primarily developed for the evaluation of health-related patient-reported outcome instruments, adjustments in the scoring system are recommended to appraise clinical tests (Terwee et al., 2006). For example, appraisal of the sample size was omitted as sample sizes are typically much smaller in clinical studies compared to patient-reported outcome studies. Sample size is an item in all boxes, except for box G: generalisability. Therefore the methodological score could not (negatively) be influenced by a small sample size.

Data and study characteristics were independently extracted by two reviewers (SP and JP). In case of disagreement, a third investigator (AP) assisted in the resolution of the disagreement. Pooling of data and meta-analysis were considered in case of homogeneity of study population and outcome measures.

2.4. Data synthesis and analysis

We performed a best evidence synthesis per test for each measurement property. Since there is no consensus guideline available for reviews on measurement properties, we adopted the criteria for levels of evidence as presented by the Grades of Recommendation, Assessment, Development and Evaluation Working Group (GRADE Working Group), and adjusted them by taking methodological quality of the studies and adequacy of the measurement properties into account (Terwee et al., 2006). The criteria for levels of evidence are presented in Appendix 2.

3. Results

The searches identified 1381 articles. After removal of duplicates, 927 articles remained. After screening titles 84 articles remained, and after screening abstract 35 articles remained. Following full-text assessment, 12 studies met the inclusion criteria and were assessed for methodological quality (see Fig. 1). Eleven studies (Potter and Rothstein, 1985, Herzog et al., 1989, Lindsay et al., 1995, Meijne et al., 1999, van Kessel-Cobelens et al., 2008, Riddle and Freburger, 2002, Tong et al., 2006, Hungerford et al., 2007, Arab et al., 2009, Curnow et al., 2010, Shimpi et al., 2018) described the reliability and two studies (Cibulka and Koldehoff, 1999, Shimpi et al., 2018) described the validity of SIJ mobility tests. Reliability and validity were the only measurement properties reported in the identified studies.

Of the twelve studies that met the inclusion criteria, three reliability studies (Meijne et al., 1999, van Kessel-Cobelens et al., 2008, Arab et al., 2009) were of fair methodological quality. The other studies documenting reliability (Potter and Rothstein, 1985, Herzog et al., 1989, Lindsay et al., 1995, Riddle and Freburger, 2002, Tong et al., 2006, Hungerford et al., 2007, Curnow et al., 2010, Shimpi et al., 2018) and validity (Cibulka and Koldehoff, 1999, Shimpi et al., 2018) were of poor methodological quality and were therefore not considered for data extraction. Box B (Reliability), Box F (Hypothesis testing), Box H (Criterion validity) and Box G (Generalisability) from the COSMIN score card (Mokkink et al., 2011) were used. Table 1 summarises the methodological scores. Appendix 3 provides an overview of the COSMIN quality appraisal of the different studies.

The three studies (Meijne et al., 1999, van Kessel-Cobelens et al., 2008, Arab et al., 2009) with fair methodology described the reliability of eight SIJ mobility tests and of one cluster of four SIJ mobility tests. All three studies reported the inter-tester reliability (Meijne et al., 1999, van Kessel-Cobelens et al., 2008, Arab et al., 2009), and two also reported the intra-tester reliability (Meijne et al., 1999, Arab et al., 2009). The statistics used were percentage agreement, Cohen's kappa coefficient with 95% confidence intervals, bias-adjusted kappa (BAK) and prevalence-adjusted bias-adjusted kappa (PABAK). Because adjustment of kappa values for bias or prevalence is discouraged (Hoehler, 2000), BAK and PABAK values were not considered in this review. We also focussed on inter-tester reliability, rather than intra-tester reliability. The level of evidence was considered low due to the fact that only one study with at least fair methodological quality existed per test, with the exception of the Gillet test. For the Gillet test, the level of evidence was conflicting. No study assessed validity. Below is an overview of the tests from the three included studies that scored fair on methodological quality.

Table 1
Overview of the methodological screening using COSMIN.

Study	Methodological quality per clinimetric property			
	Box B Reliability	Box F Hypothesis testing	Box H Criterion validity	Box G Generalisability
(Meijne et al., 1999)	FAIR	N/A	N/A	FAIR
(van Kessel-Cobelens et al., 2008)	FAIR	N/A	N/A	GOOD
(Tong et al., 2006)	POOR	N/A	N/A	FAIR
(Arab et al., 2009)	FAIR	N/A	N/A	FAIR
(Shimpi et al., 2018)	POOR	POOR	N/A	GOOD
(Hungerford et al., 2007)	POOR	N/A	N/A	FAIR
(Cibulka and Koldehoff, 1999)	N/A	N/A	POOR	POOR
(Riddle and Freburger, 2002)	POOR	N/A	N/A	FAIR
(Curnow et al., 2010)	POOR	N/A	N/A	FAIR
(Potter and Rothstein, 1985)	POOR	N/A	N/A	FAIR
(Herzog et al., 1989)	POOR	N/A	N/A	POOR
(Lindsay et al., 1995)	POOR	N/A	N/A	POOR

N/A: not applicable.

3.1. Individual SIJ mobility tests

The inter-tester reliability of eight SIJ mobility tests was assessed. The tests were the: (1) click-clack test (i.e., with the patient sitting, movement of the left and right posterior superior iliac spine (PSIS) is assessed when the patient moves the trunk from lordosis to kyphosis); (2) standing flexion test (i.e., with the patient standing, SIJ movement is assessed while the patient bends forward); (3) seated flexion test (i.e., with the patient sitting, SIJ movement is assessed while the patient bends forward); (4) Gillet test (i.e., with the patient standing, SIJ movement is assessed while the patient pulls the opposite knee to the chest); (5) prone knee flexion test (i.e., with the patient prone, leg length discrepancy is compared between test positions); (6) heel-bank test (i.e., with the patient sitting, SIJ movement is assessed while the patient places one foot on the treatment table); (7) abduction test (i.e., with the patient in side-lying, a discrepancy in load transfer is assessed); and (8) thumb-posterior superior iliac spine (PSIS) test (i.e., with the patient sitting, the position of the PSIS were measured on a horizontal line in relation to each other).

The kappa values were reported for the left and right SIJ, rather than for the affected or unaffected side. When there was a discrepancy in level of agreement for the left and right SIJ, our interpretation of the reliability was based on the lowest value. One test (seated flexion test (Arab et al., 2009)) had substantial agreement (κ: 0.61 to 0.80); one test (standing flexion test (Arab et al., 2009)) had moderate agreement (κ:

Table 2
Inter-tester reliability of individual SIJ mobility tests.

SIJ mobility test	Inter-tester reliability	Interpretation
Standing flexion test (Arab et al., 2009)	κ (95% CI) left SIJ: 0.55 (0.20–0.90)	Moderate agreement
	κ (95% CI) right SIJ: 0.51 (0.08–0.95)	Moderate agreement
Seated flexion test (Arab et al., 2009)	κ (95% CI) left SIJ: 0.64 (0.32–0.96)	Substantial agreement
	κ (95% CI) right SIJ: 0.75 (0.42–1.08)	Substantial agreement
Prone knee flexion test (Arab et al., 2009)	κ (95% CI) left SIJ: 0.33 (−0.18–0.85)	Fair agreement[a]
	κ (95% CI) right SIJ: 0.58 (0.25–0.91)	Moderate agreement[a]
Gillet test (Arab et al., 2009)	κ (95% CI) left SIJ: 0.34 (−0.06–0.7)	Fair agreement[a]
	κ (95% CI) right SIJ: 0.41 (0.03–0.87)	Moderate agreement[a]
Gillet test (Meijne et al., 1999)	κ (95% CI) M1: 0.03 (−0.01–0.15)	Slight agreement
	κ (95% CI) M2: 0.14 (−0.06–0.34)	Slight agreement
Heel-bank test (van Kessel-Cobelens et al., 2008)	κ left SIJ: 0.39	Fair agreement[a]
	κ right SIJ: 0.06	Slight agreement[a]
Abduction test (van Kessel-Cobelens et al., 2008)	κ left SIJ: 0.50	Moderate agreement[a]
	κ right SIJ: 0.37	Fair agreement[a]
Thumb-PSIS test (van Kessel-Cobelens et al., 2008)	κ: 0.20	Slight agreement
Click-clack test (van Kessel-Cobelens et al., 2008)	κ: 0.00	Slight agreement

[a] When there was a discrepancy in the level of agreement for the left and right SIJ, the lowest value was used for the interpretation in the manuscript. K (95% CI): Kappa coefficient with 95% confidence interval; SIJ: Sacroiliac joint; M1 and M2: measurement time 1 and 2.

0.41 to 0.60); two tests (prone knee flexion test (Arab et al., 2009) and abduction test (van Kessel-Cobelens et al., 2008)) had fair agreement (κ: 0.21 to 0.40); two tests (heel-bank test (van Kessel-Cobelens et al., 2008), thumb PSIS (van Kessel-Cobelens et al., 2008)) had slight agreement (κ: 0.01 to 0.20); one test (click-clack test (van Kessel-Cobelens et al., 2008)) had equal to chance agreement (κ: 0.00); and one test (Gillet test) had slight agreement in one study (Meijne et al., 1999) and fair agreement in another study (Arab et al., 2009). Table 2 provides a detailed overview of the inter-tester reliability values and interpretation. Noteworthy were the large confidence intervals around the reliability estimates, even for the tests with the better reliability scores.

3.2. Test clusters

Arab et al. (Arab et al., 2009) examined a test cluster of four SIJ mobility tests (Gillet test, standing flexion test, seated flexion test and the prone knee flexion test). The inter-tester reliability was assessed for one, two, three and four positive tests. The kappa values with 95% confidence intervals are presented in Table 3. Compared to the individual tests, the test cluster showed higher reliability, with the highest agreement for two positive tests. As for the individual tests, the confidence intervals around most reliability estimates were large.

The only test that was evaluated in more than one study was the Gillet test (Meijne et al., 1999; Arab et al., 2009). Pooling of the data and meta-analysis could not be performed due to the heterogeneity of the study population.

4. Discussion

The purpose of this study was to systematically review the clinimetric properties of SIJ mobility tests in the peer-reviewed literature to verify whether recommendations regarding the clinical usability of SIJ mobility tests had to be revised. The latest recommendation is that SIJ mobility tests should not be used because of poor clinimetric properties, but this recommendation is based on reviews published almost 20 years ago (van der Wurff et al., 2000a, 2000b). Based on the three included reliability studies (Meijne et al., 1999, van Kessel-Cobelens et al., 2008, Arab et al., 2009) with sufficient methodological quality and the absence of studies that measured validity, we conclude that the current recommendation that SIJ mobility tests should not be utilized in clinical practice remains unchanged.

Table 3
Inter-tester reliability of a clusters of four SIJ mobility tests.

Test cluster	Inter-tester reliability	Interpretation
One positive test	κ (95% CI) left SIJ: 0.56 (0.16–0.95)	Moderate agreement[a]
	κ (95% CI) right SIJ: 0.60 (0.29–0.91)	Substantial agreement[a]
Two positive tests	κ (95% CI) left SIJ: 0.81 (0.49–1.07)	Almost perfect agreement[a]
	κ (95% CI) right SIJ: 0.78 (0.57–1.06)	Substantial agreement[a]
Three positive tests	κ (95% CI) left SIJ: 0.44 (0.08–0.79)	Moderate agreement[a]
	κ (95% CI) right SIJ: 0.60 (0.18–1.02)	Substantial agreement[a]
Four positive tests	κ (95% CI) left SIJ: 0.33 (−0.36 to 1.04)	Fair agreement[a]
	κ (95% CI) right SIJ: 0.77 (0.35–1.2)	Substantial agreement[a]

[a] When there was a discrepancy in the level of agreement for the left and right SIJ, the lowest value was used for the interpretation in the manuscript. The test cluster consisted of the Gillet test, standing flexion test, sitting flexion test and prone knee flexion test. K (95% CI): Kappa coefficient with 95% confidence interval; SIJ: Sacroiliac joint.

4.1. Reliability of individual SIJ mobility tests

Meijne et al. (1999) and Van Kessel et al. (2008) demonstrated low reliability values for individual SIJ mobility tests. These values suggest that different clinicians may reach different conclusions when assessing the same patient. The results from Arab et al. (2009) appeared more promising at first sight as they reported fair and moderate reliability for individual SIJ mobility tests. For example, the seated flexion test showed substantial reliability. However, the corresponding confidence intervals were very wide, which questions the reproducibility of these reliability values. Furthermore, reliability should be assessed in patients with a suspicion of a certain condition. By assessing the reliability for the left and the right SIJ, rather than for the affected and unaffected SIJ, the reliability values might have been inflated by the assessment of unaffected joints.

4.2. Reliability of test clusters

The inter-tester reliability for the cluster of tests was fair to substantial (Arab et al., 2009). The reliability was the highest for the cluster with two positive tests out of four SIJ mobility tests. Although these reliability values were higher than for individual tests and of a clinically useful magnitude, there remains uncertainty regarding the reliability. First, the test cluster consists of individual SIJ mobility tests with uncertain reliability due to the large confidence intervals; Second, as for the individual tests, the reliability of the test clusters was reported for the left and right SIJ rather than for the affected and unaffected side which may have inflated the reported reliability; Third, there was no (a-priori) rationale why the cluster with two positive tests would have the highest reliability. This is problematic as the reliability was assessed for a large number of SIJ mobility test combinations (4 clusters for SIJ mobility tests and 12 clusters combining SIJ mobility and pain provocation tests). No statistical considerations were made for the large number of reliability measurements in a small sample (N = 25).

4.3. Validity of SIJ mobility tests

Our literature search did not identify validity studies for SIJ mobility tests that were of sufficient methodological quality to be included in this review. Therefore, no recommendations could be made regarding the diagnostic accuracy of SIJ mobility tests.

With respect to face validity, there is considerable variation in the anatomical shape of the joint surfaces of the SIJ (Bakland and Hansen, 1984, Vleeming et al., 2012). Furthermore, SIJ motion is very small (for example: 0.2° posterior rotation; 0.6° rotation around the helical axis and 0.3 mm translation (Sturesson et al., 2000a)). Measuring SIJ mobility using manual palpation might therefore be impossible. (Walker, 1992, Sturesson et al., 2000a, 2000b; Freburger, 2001). Furthermore, the innominate bone may deform during loading (Pool--Goudzwaard et al., 2012), and soft tissue structures may cause palpatory illusions (Lewit and Liebenson, 1993, Sturesson et al., 2000a, 2000b).

4.4. Limitations

A limitation of this review is that the findings were based on a limited number of studies. This was due to the fact that the majority of the relevant papers (9 out of 12) were of insufficient methodological quality. The reasons for the poor scores of the reliability studies on the COSMIN checklist (Mokkink et al., 2011) included inappropriate time intervals between measurements, insufficient description of demographic and/or disease characteristics, inappropriate selection of statistics, uncertainty about blinding of testers, and how missing items were handled. For the two validity studies, the reference standard was not adequate. The included reliability studies (Meijne et al., 1999, van Kessel-Cobelens et al., 2008, Arab et al., 2009) were of fair, but not high, methodological quality. Most were underpowered for the amount of tests and combinations examined, although the sample size criterion was waived as recommended when using the COSMIN appraisal tool for clinical tests rather than patient-reported outcome measures. Due to the small number of studies which assessed different SIJ mobility tests in different populations, no pooling of data or meta-analysis could be performed.

4.5. Reflections and conclusions

There is no new evidence for the validity of SIJ mobility tests when considering literature of at least fair methodological quality. Furthermore, there is only low quality and conflicting evidence for inter-rater reliability and no evidence for responsiveness of tests because no studies examined this property. Considering the absence of high-quality diagnostic accuracy studies, and the questionable face validity, the literature on SIJ mobility tests should be interpreted prudently. The reliability of individual SIJ mobility tests and test clusters is questionable or uncertain. New techniques have been proposed to assess SIJ mobility, however they were only evaluated in healthy participants (Bussey et al., 2004, Bussey et al., 2009, Adhia et al., 2012). Considering the theoretical construct of palpation for joint mobility, it is it is uncertain that these new tests will be diagnostically accurate and reliable. The findings of this review were based on studies with only fair methodological quality, but we believe it is unlikely that an increase in methodological quality will lead to better clinimetric properties of SIJ mobility tests. We therefore conclude that the findings of this review are in line with the earlier recommendations (van der Wurff et al., 2000a, 2000b) and that SIJ mobility tests in clinical practice or curricula of educational programs remains problematic.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.msksp.2019.102090.

Appendix A. Search strings per database

PubMed: ("Sacroiliac Joint" [Mesh] OR "Sacroiliac Joint" [tiab] OR "Sacroiliac Joints" [tiab] OR sij [tiab] OR "si joint" [tiab] OR "sacro iliac joint" [tiab] OR "sacro iliac joints" [tiab]) AND ("Range of Motion, Articular" [Mesh] OR "range of movement" [tiab] OR "flexibility" [tiab] OR "mobility" [tiab] OR motion [tiab] OR flexion [tiab] OR stiffness [tiab] OR dysfunction [tiab]) AND ("Diagnosis" [Mesh] OR "Diagnosis" [sh] OR assessment [tiab] OR assessing [tiab] OR assess [tiab] OR examination [tiab] OR exam [tiab] OR examinate [tiab] OR examinating [tiab] OR examine [tiab] OR diagnose [tiab] OR diagnosis [tiab] OR diagnoses [tiab] OR diagnosing [tiab] OR diagnostic [tiab] OR test [tiab] OR tests [tiab] OR testing [tiab]) AND ("Reproducibility of Results" [Mesh] OR "Sensitivity and Specificity" [Mesh] OR valid*[tiab] OR reliab*[tiab] OR reproduc*[tiab] OR sensitiv* [tiab] OR specific*[tiab] OR significan*[tiab] OR responsive*[tiab] OR unreliab*[tiab])

Embase: ('sacroiliac Joint'/exp OR ('sacroiliac Joint' OR 'sacroiliac Joints' OR sij OR 'si joint' OR 'sacro iliac joint' OR 'sacro iliac joints'):ti,ab) AND ('range of motion'/exp OR 'joint characteristics and functions'/exp OR ('range of movement' OR flexibility OR mobility OR motion OR flexion OR stiffness OR dysfunction):ti,ab) AND ('diagnosis'/exp OR (assessment OR assessing OR assess OR examination OR exam OR examinate OR examinating OR examine OR diagnose OR diagnosis OR diagnoses OR diagnosing OR diagnostic OR test OR tests OR testing):ti,ab) AND ('measurement precision'/exp OR 'sensitivity and specificity'/exp OR 'statistical parameters'/exp OR (valid* OR reliab* OR reproduc* OR sensitiv* OR specific* OR significan* OR responsive* OR unreliab*):ti,ab).

CINAHL: (MH "Sacroiliac Joint" OR MH "Sacroiliac Joint Dysfunction" OR "Sacroiliac Joint" OR "Sacroiliac Joints" OR sij OR "si joint" OR "sacro iliac joint" OR "sacro iliac joints") AND (MH "Movement+" OR "range of movement" OR flexibility OR mobility OR motion OR flexion OR stiffness OR dysfunction) AND (MH "Diagnosis+" OR assessment OR assessing OR assess OR examination OR exam OR examinate OR examinating OR examine OR diagnose OR diagnosis OR diagnoses OR diagnosing OR diagnostic OR test OR tests OR testing) AND (MH "Measurement Issues and Assessments+" OR MH "Reproducibility of Results" OR valid* OR reliab* OR reproduc* OR sensitiv* OR specific* OR significan* OR responsive* OR unreliab*)

SPORTDiscus: (DE "SACRUM" OR DE "SACROILIAC joint"OR "Sacroiliac Joint" OR "Sacroiliac Joints" OR sij OR "si joint" OR "sacro iliac joint" OR "sacro iliac joints" OR sacrum) AND (DE "JOINTS – Range of motion" OR DE "JOINT stiffness" OR DE "JOINTS – Hypermobility" OR "range of movement" OR flexibility OR mobility OR motion OR flexion OR stiffness OR dysfunction) AND (DE "DIAGNOSIS" OR DE "PHYSICAL diagnosis" OR DE "DIAGNOSTIC examinations" OR assessment OR assessing OR assess OR examination OR exam OR examinate OR examinating OR examine OR diagnose OR diagnosis OR diagnoses OR diagnosing OR diagnostic OR test OR tests OR testing) AND (valid* OR reliab* OR reproduc* OR sensitiv* OR specific* OR significan* OR responsive* OR unreliab*)

Cochrane Library: ("Sacroiliac Joint" or "Sacroiliac Joints" or sij or "si joint" or "sacro iliac joint" or "sacro iliac joints":ti,ab,kw) AND ("Range of Motion" or "range of movement" or "flexibility" or "mobility" or motion or flexion or stiffness or dysfunction:ti,ab,kw) AND ("Diagnosis" or assessment or assessing or assess or examination or exam or examinate or examinating or examine or diagnose or diagnosis or diagnoses or diagnosing

or diagnostic or test or tests or testing:ti,ab,kw) AND ("Reproducibility of Results" or valid* or reliab* or reproduc* or sensitiv* or specific* or significan* or responsive* or unreliab*:ti,ab,kw).

Appendix B. Best evidence synthesis criteria for levels of evidence

Level of evidence	Criteria
High quality evidence	Multiple studies with methodological quality rated as at least Good OR One study with methodological quality rated as Excellent AND Consistent findings ((+) OR (−))
Moderate quality evidence	Multiple studies with methodological quality rated as Fair OR one study with methodological quality rated as Good AND Consistent findings ((+) OR (−))
Low quality evidence	One study with methodological quality rated as Fair AND Consistent findings ((+) OR (−))
Conflicting evidence	Multiple studies AND conflicting findings
No evidence	Only studies with methodological quality rated as Poor OR Only studies with results rated as (?)

Appendix C. Methodological screening

Study	Study characteristics			Methodological Quality		
	Participant characteristics	Test instrument Measurement of movement pattern	Method Purpose	Box B Reliability	Box H Criterion validity Box F Hypotheses testing	Box Generalisability
(Meijne et al., 1999)	N = 38 M = 38 F = 0 Mean age 23 SD 2.24 Participants with and without low back pain.	Gillet test Palpatory and visual estimation	Purpose: Determine the intra and interreliability of the Gillet test.	FAIR Other minor methodological flaws in the design or execution of the study: The use of BAK and PABAK, the use of a convenience sample and heterogeneity of the population.	N/A	FAIR Important disease characteristics are partly described.
(van Kessel-Cobelens et al., 2008)	N = 62 M = 62 F = 0 Mean age: 30.9 SD 4.2 1.Participants with PGP (n = 22) Mean age 30.7 SD 3.2 2.Participants that were pregnant with no PGP (n = 20), age mean 30.5 SD 3.6 3. participants that were not pregnant and had no back pain (n = 20) Age mean 30.6 SD 5.5.	1 Thumb PSIS 2 Click clack 3 Heel bank test 4 Abduction test Palpatory and visual estimation	Purpose: Investigate the intertester reliability of tests to detect asymmetry in mobility of the SIJ.	FAIR Other minor methodological flaws in the design or execution of the study: Between the inclusion and the testing days was a period of 3 weeks, heterogeneity of the population and the construct of the test that was measured (measuring SIJ mobility with experience in effort).	N/A	GOOD It can be deduced (e.g. from the affiliations of the authors) and in which language the PRO instrument was evaluated
(Tong et al., 2006)	N = 24 M = 9 F = 15 Mean age: 68.3 SD 13.2 Participants with low back pain.	Cluster of tests consisting of 1 Seated flexion test 2 Standing Stork 3 Standing flexion test Palpatory and visual estimation	Purpose: Compare the intertester reliability of three methods of combining the results to decide the side of SIJ dysfunction, sacral bone position and innominate bone position.	POOR Other important methodological flaws in the design or execution of the study: Head researcher not blinded. Exclusion from 1 s examiner. Not clear if all second examiners examined. Method description not clear.	N/A	FAIR Percentage of missing responses NOT described. Important disease characteristics and setting are partly described.
(Arab et al., 2009)	N = 25 M = 15 F = 10	Single and cluster of tests consisting of: 1. Standing flexion test	Purpose: Evaluate the inter en intra reliability	FAIR Other minor methodological	N/A	FAIR Important disease

(continued on next page)

(*continued*)

Study	Study characteristics			Methodological Quality		
	Participant characteristics	Test instrument Measurement of movement pattern	Method Purpose	Box B Reliability	Box H Criterion validity Box F Hypotheses testing	Box Generalisability
	Age Range 20-65 Mean 43 SD 10 Patients with pain below L5, over the posterior aspect of the SIJ, around PSIS and buttock with or without leg pain.	2. Sitting flexion test 3. Prone knee flexion test 4. Gillet test Palpatory and visual estimation	of single and composites of tests.	flaws in the design or execution of the study: The use of BAK and PABAK. A large amount of (combination of) tests examined on a small sample size.		characteristics are partly described
(Shimpi et al., 2018)	N = 45 23 participants with SIJ dysfunction Mean age: 29.4 SD 4.5 M = 10 F = 13 22 healthy Mean age: 27.8 SD 6.1 M = 9 F = 13	Shimpi prone SIJ test	Purpose: To assess reliability and validity of a new clinical test for the assessment of patients with SIJ movement dysfunction.	POOR Other important methodological flaws in the design or execution of the study: Assumable that the testers were not blinded for present or absence of pain from the participants. Number of participants with only mobility restrictions or only pain not presented. ICC not suitable as analysis for the measurements with outcomes in the study.	**Box F** POOR Other important methodological flaws in the design or execution of the study: The tests used for hypotheses testing are frequently presented as poor in literature. The values for validity are too low to use as a reference standard. The calculation of the tests is not adequate. The validity values were calculated by averaging every value.	GOOD It can be deduced (e.g. from the affiliations of the authors) and in which language the PRO instrument was evaluated.
(Hungerford et al., 2007)	N = 33 M = 14 F = 19 Mean age: 36.2 SD 13.4 Participants with and without lumbopelvic pain.	Standing stork test Palpatory and visual estimation	Purpose: Determine whether experienced therapist could reliable detect the pattern of motion occurring between the innominate bone and the sacrum (intrapelvic motion).	POOR Only percentage agreement calculated.	N/A	FAIR Percentage of missing responses NOT described. Important disease characteristics and setting are partly described.
(Cibulka and Koldehoff, 1999)	N = 219 M = 125 F = 94 Mean age: 29 SD 15.1 Range 12-84 Participants with and without low back pain.	1 Standing flexion test 2 Sitting posterior-superior iliac spines palpation 3 Supine long sitting test 4 Prone knee flexion test Palpatory and visual estimation	Purpose: Determine the clinical usefulness for a cluster of sacro iliac joint tests by finding the sensitivity, specificity and positive and negative predictive value.	N/A	**BOX H** POOR Other important methodological flaws in the design or execution of the study: Missing items not described, spreading population not described. NOT assessed if all items refer to relevant aspects of the construct to be measured. NOT assessed if all items are relevant for the study population OR target population not involved. Criterion used can NOT be considered	POOR Distribution of sex NOT described.

(*continued on next page*)

(*continued*)

Study	Study characteristics			Methodological Quality		
	Participant characteristics	Test instrument Measurement of movement pattern	Method Purpose	Box B Reliability	Box H Criterion validity Box F Hypotheses testing	Box Generalisability
					as adequate Gold Standard.	
(Riddle and Freburger, 2002)	N = 65 M = 23 F = 42 Mean age: 47.4 SD 14.0 Range 18-81 Participants with low back pain and discomfort in the buttock area.	1 Standing flexion test 2 Prone knee flexion test 3 Supine long sitting test 4 Sitting PSIS test Palpatory and visual estimation	Purpose: Determining the reliability of individual and composite of motion tests for the SIJ with the side of presumed SIJ dysfunction and the type of asymmetry taken into account.	POOR Other important methodological flaws in the design or execution of the study: First analysis: observation outcomes only on + or -, not on agreement on side of + or -. Second analysis therapists did not necessarily have to agree on the type of asymmetry present (ie, anteriorly or posteriorly rotated innominate), just the side that was involved.	N/A	FAIR Methods for patient selection was unclearly described. Percentage of missing responses NOT described. The setting is partly described.
(Curnow et al., 2010)	N = 74 M = 19 F = 55 Age median M: 51 Mean age M: 49.8 SD 16.3 F: Age median: 51 F: Mean age: 50.8 SD 13 Participants could be symptomatic (mild low back pain) or asymptomatic.	Standing stork in three positions Palpatory and visual estimation	Purpose: Examine the reliability of the stork test and test if changing the stance of the starting position alters the outcome.	POOR Only one measurement. Other important methodological flaws in the design or execution of the study: In analysis of the DVD there is not described with who and how many people did the analysis or if there was consultation.	N/A	FAIR Important disease characteristics are partly described. Percentage of missing responses NOT described.
(Potter and Rothstein, 1985)	N = 17 M = 10 F = 7 Age: 24-58 Mean age: 39 Median age: 36 F: Mean age: 39.6 Range 33-45 M: Mean age: 38.7 Range: 24-58 Participants with lumbosacral symptoms	1. Sitting flexion test 2 Standing flexion test 3. Supine-long sitting test 4. Gillet test 5. Prone knee flexion test Palpatory and visual estimation	Purpose: Intertester reliability of the selected tests	POOR Test conditions were NOT similar. Only percentage agreement calculated.	N/A	FAIR Important disease characteristics and setting are partly described. Methods for patient selection was unclearly described. Percentage of missing responses NOT described.
(Herzog et al., 1989)	N = 11 Participants with a sacro iliac joint problem Sex:not described Age:not described	Gillet motion palpation procedure Palpatory and visual estimation	Purpose: Asses intra and inter test reliability of the Gillet test.	POOR Only percentage agreement calculated.	N/A	POOR Median or mean age is NOT described. Distribution of sex is NOT described. Important disease characteristics, median or mean, sex are NOT described.
(Lindsay et al., 1995)	N = 8 Participants from a cross-country ski	1 Active kinetic tests: contralateral and ipsilateral	Purpose: To examine the interrater	POOR Other important methodological		POOR Median or mean age is NOT

(*continued on next page*)

(*continued*)

Study	Study characteristics			Methodological Quality		
	Participant characteristics	Test instrument Measurement of movement pattern	Method Purpose	Box B Reliability	Box H Criterion validity Box F Hypotheses testing	Box Generalisability
	club with lumbosacral problems Sex: not described Age: not described	2 Mobility tests: forward bending sitting and standing, 3 Passive accessory glides: anterior-posterior and cranial-caudal Palpatory and visual estimation	reliability of the lumbosacral tests.	flaws in the design or execution of the study: There was no standardisation for the tests. Testers reviewed and summarised their outcomes after testing. No inclusion or exclusion criteria were described.		described. Distribution of sex is NOT described. Important disease characteristics, median or mean, sex are NOT described. Method for patient selection not described.

N= Power, M: male, F: female AS: Ankylosing spondylitis, PPP: post-partum pelvic girdle pain, SD: standard deviation, N/A: not applicable.

References

Adhia, D.B., Bussey, M.D., Mani, R., Jayakaran, P., Aldabe, D., Milosavljevic, S., 2012. Inter-tester reliability of non-invasive technique for measurement of innominate motion. Man. Ther. 17 (1), 71–76. https://doi.org/10.1016/j.math.2011.09.005.

Arab, A.M., Abdollahi, I., Joghataei, M.T., Golafshani, Z., Kazemnejad, A., 2009. Inter- and intra-examiner reliability of single and composites of selected motion palpation and pain provocation tests for sacroiliac joint. Man. Ther. 14 (2), 213–221. https://doi.org/10.1016/j.math.2008.02.004.

Bakland, O., Hansen, J., 1984. The axial sacroiliac joint. Anat. Clin. (6), 29–36.

Bussey, M.D., Bell, M.L., Milosavljevic, S., 2009. The influence of hip abduction and external rotation on sacroiliac motion. Man. Ther. 14 (5), 520–525. https://doi.org/10.1016/j.math.2008.08.009.

Bussey, M.D., Yanai, T., Milburn, P., 2004. A non-invasive technique for assessing innominate bone motion. Clin. Biomech. 19 (1), 85–90.

Cibulka, M.T., Koldehoff, R., 1999. Clinical usefulness of a cluster of sacroiliac joint tests in patients with and without low back pain. J. Orthop. Sport. Phys. Ther. 29 (2), 83–89. https://doi.org/10.2519/jospt.1999.29.2.83 discussion 90-2.

Cohen, S.P., Chen, Y., Neufeld, N.J., 2013. Sacroiliac joint pain: a comprehensive review of epidemiology, diagnosis and treatment. Expert Rev. Neurother. 13 (1), 99–116. https://doi.org/10.1586/ern.12.148.

Curnow, D., Cobbin, D., Wyndham, J., 2010. Reliability of the stork test: is starting stance important? Chiropract. J. Aust. 40 (4), 137–141.

Dreyfuss, P., Dryer, S., Griffin, J., Hoffman, J., Walsh, N., 1994. Positive sacroiliac screening tests in asymptomatic adults. Spine 19 (10), 1138–1143.

Freburger, J.K., 2001. Using published evidence to guide the examination of the sacroiliac joint region. Phys. Ther. 81 (5), 1135–1143.

Godee, A., Hegedus, E., Sizer, P., Brismee, J., Linberg, A., Cook, C., 2008. Three-dimensional movement of the sacroiliac joint:A systematic review of the literature and assessment of clinical utility. J. Man. Manip. Ther. 16 (1), 25–38.

Hancock, M.J., Maher, C.G., Latimer, J., Spindler, M.F., McAuley, J.H., Laslett, M., Bogduk, N., 2007. Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. Eur. Spine J. (16), 1539–1550.

Herzog, W., Read, L.J., Conway, P.J., Shaw, L.D., McEwen, M.C., 1989. Reliability of motion palpation procedures to detect sacroiliac joint fixations. J. Manip. Physiol. Ther. 12 (2), 86–92.

Hoehler, F.K., 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J. Clin. Epidemiol. 53 (5), 499–503.

Hungerford, B.A., Gilleard, W., Moran, M., Emmerson, C., 2007. Evaluation of the ability of physical therapists to palpate intrapelvic motion with the stork test on the support side. Phys. Ther. 87 (7), 879–887.

Irwin, R., Watson, T., Minick, R., Ambrosius, W., 2007. Age, body mass index, and gender differences in sacro iliac joint pathology. Am. J. Phys. Med. Rehabil. 86 (1), 37–44.

Kibsgard, T.J., 2015. Radiostereometric analysis of sacroiliac joint movement and outcomes of pelvic joint fusion. Acta Orthop. 86, 1–43.

Laslett, M., Aprill, C., McDonald, B., Young, S., 2005. Diagnosis of sacroiliac joint pain: validity of individual provocation tests and composites of tests. Man. Ther. 10 (3), 207–218.

Lewit, K., Liebenson, C., 1993. Palpation-problems and implications. J. Manip. Physiol. Ther. 16 (9), 586–590.

Lindsay, D.M., Meeuwise, W.H., Mooney, M.E., Summersides, J., 1995. Interrater reliability of manual therapy assessment techniques. Physiother. Can. 47 (3), 173–180.

Meijne, W., Neerbos, K., Aufdemkampe, G., Wurff, P., 1999. Intraexaminer and interexaminer reliability of the gillet test. J. Manip. Physiol. Ther. 22 (1), 4–9.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D., 2019. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. http://www.equator-network.org/reporting-guidelines/prisma/. (Accessed 27 March 2019).

Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., Bouter, L.M., de Vet, H.C.W., 2019. COSMIN checklist manual 2011, VU University Medical Center,department of EMGO Amsterdam. http://www.cosmin.nl/images/upload/files/COSMIN%20checklist%20manual%20v9.pdf. (Accessed 25 November 2016).

Mokkink, L.B., Prinsen, C.A.C., Bouter, L.M., de Vet, Henrica, C.W., Terwee, C.B., 2016. The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) and how to select an outcome measurement instrument. Braz. J. Phys. Ther. Rev. Bras. Fisioterapia 20 (2), 43–51.

Pool-Goudzwaard, A., Gnat, R., Spoor, K., 2012. Deformation of the innominate bone and mobility of the pubic symphysis during asymmetric moment application to the pelvis. Man. Ther. 17 (1), 66–70. https://doi.org/10.1016/j.math.2011.09.002.

Potter, N.A., Rothstein, J.M., 1985. Intertester reliability for selected clinical tests of the sacroiliac joint. Phys. Ther. 65 (11), 1671–1675.

Riddle, D.L., Freburger, J.K., 2002. Evaluation of the presence of sacroiliac joint region dysfunction using a combination of tests: a multicenter intertester reliability study. Phys. Ther. 82 (8), 772–781.

Shimpi, A., Hatekar, R., Shyam, A., Sancheti, P., 2018. Reliability and validity of a new clinical test for assessment of sacroiliac joint dysfunction. Hong Kong Physiother. J. 38 (1), 13–22.

Sturesson, B., Selvik, G., Uden, A., 1989. Movements of the sacroiliac joints. A roentgen stereophotogrammetric analysis. Spine 14 (2), 162–165.

Sturesson, B., Uden, A., Vleeming, A., 2000a. A radiostereometric analysis of movements of the sacroiliac joints during the standing hip flexion test. Spine 25 (3), 364–368.

Sturesson, B., Uden, A., Vleeming, A., 2000b. A radiostereometric analysis of the movements of the sacroiliac joints in the reciprocal straddle position. Spine 25 (2), 214–217.

Szadek, K., Wurff, v, Tulder, vM., Zuurmond, W., Perez, R.S.G.M., April 2009. Diagnostic validity of criteria for sacroiliac joint pain: a systematic review. J. Pain 10 (4), 354–368.

Terwee, C., Mokkink, L., Steultjens, M., Dekker, J., 2006. Performance-based methods for measuring the physical function of patients with osteoarthritis of the hip of knee: a systematic review of measurement properties. Rheumatology 45 (7), 890–902.

Tong, H.C., Heyman, O.G., Lado, D.A., Isser, M.M., 2006. Interexaminer reliability of three methods of combining test results to determine side of sacral restriction, sacral base position, and innominate bone position. J. Am. Osteopath. Assoc. 106 (8), 464–468.

Van der Wurff, P., Buijs, E., Groen, G., 2006. A multitest regimen of pain provocation tests as an aid to reduce unnecessary minimally invasive sacroiliac joint procedures. Arch. Phys. Med. Rehabil. 87 (1), 10–14.

van der Wurff, P., Hagmeijer, R., Meyne, W., 2000a. Clinical tests of the sacroiliac joint: a systematic methodological review. part 1: reliability. Man. Ther. 5 (1), 30–36.

van der Wurff, P., Meyne, W., Hagmeijer, R., 2000b. Clinical tests of the sacroiliac joint: a systematic methodological review. part 2: validity. Man. Ther. 5 (2), 89–96.

van Kessel-Cobelens, A.M., Verhagen, A.P., Mens, J.M., Snijders, C.J., Koes, B.W., 2008. Pregnancy-related pelvic girdle pain: intertester reliability of 3 tests to determine asymmetric mobility of the sacroiliac joints. J. Manip. Physiol. Ther. 31 (2), 130–136. https://doi.org/10.1016/j.jmpt.2007.12.003.

Vleeming, A., Schuenke, M., Masi, A., Danneels, L., Willard, F., 2012. The sacroiliac joint: An overview of its anatomy, function and potential clinical implications. J. Anat. 221 (6), 537–567.

Walker, J.M., 1992. The sacroiliac joint: a critical review. Phys. Ther. 72 (12), 903–916.