Validity of clinical tests for sacroiliac and lumbar joint dysfunction:

A systematic review of the literature.

Julie Mangen & Vasiliki Folia

European School of Physiotherapy, Hogeschool van Amsterdam, Tafelbergweg 51, Amsterdam, The Netherlands

Abstract
We conducted a systematic search of the literature in order to determine which physiotherapy tests concerning testing of the Sacroiliac Joint (SIJ) and Lumbar Joint (LJ) have the highest sensitivity, specificity, and predictive values for determining the presence of sacroiliac and/or lumbar joint dysfunction when compared with the golden standard of injection blocks (in the absence of a better standard criterion, the blocks continue to be considered as such (Broadhurst et al., 1998; van der Wurff et al., 2006). Six papers were identified for inclusion in the review of the sacroiliac joint, while for the lumbar joint the search yielded surprisingly poor results. In general, study designs rarely incorporated randomized, placebo controlled, double blinded study designs or confirmatory injection blocks. There was considerable inconsistency between studies in design and outcome measurement, making comparison difficult. For the SIJ, practitioners may consider using the distraction, compression, the sacral thrust and especially the posterior shear (P4 / thigh thrust) test. For the LJ no such conclusions concerning the best evidence based tests concerning non-neurogenic dysfunction or non-pregnant patients could be reached. Further research is required to determine the optimal tests and/or combination of tests to determine the cause of pain in the lumbosacral area and differentiate between SIJ and LJ dysfunction.

Keywords: sacroiliac, lumbar, joint, validity, examination.
1.1 Content list

This study is built up as following: a general introduction (1.2), an introduction to the SIJ and evidence based research (1.3), an introduction to the LJ and evidence based research (1.4), the methods (2.), the results (3.), a discussion about SIJ findings (4.1), a discussion about LJ findings (4.2), a discussion about differential diagnosis of the Lumbosacral area. Points of interest for the ESP students and teachers (5.1), and a manual of tests recommended from this systematic review (5.2).

The appendix 1 consist of the Manual of tests recommended

1.2 Introduction

The lumbar and sacroiliac joints have been studied broadly during the last years since the anatomical aspects of those two joints render sometimes the exact allocation of the problem difficult. While lumbar dysfunction seems to be more common for patients complaining over lumbar and sacroiliac joint pain (Schwarzer et al. 1995), sacroiliac joint pain (SIJ) may still be a source of low back pain (LBP), (Shaw et al., 1992) that according to the European Guidelines on the pelvic girdle pain (Albert et al, 2006) can occur separately or in conjunction with low back pain. According to Schwarzer (1995) the study of the SIJ has been hindered by the fact that there has not been a satisfactory standard of criteria by which its prevalence can be measured and against which various clinical examinations can be validated. For the clinician it is important to diagnose low back pain properly, and SIJ dysfunction in particular, in order to treat the problem in an appropriate way (Kirkaldy-Willis & Hill 1979). There is a wide variety of SIJ tests available to detect dysfunction, but none of them appears to be superior to others. The most common are: Gaenslen’s, Yeoman’s and Patrick’s test (Bernard et al., 1997). The same holds for the most common lumbar area joint tests, such as straight leg raising, Schober, Quadrant test (Magee, 2008). In three guidelines for the management of the low back pain (Dutch, European and New Zealand) there is not much evidence about the specificity and validity of some of these tests concerning their diagnosis (with exception of the straight leg raising for nerve root pain in LBP). In the area of the lumbar joint diagnostic imaging tests (including X-rays, CT and MRI) are not routinely indicated for acute nonspecific low back pain, if there are no clear indications of possible serious pathology or radicular syndrome (Albert et al, 2006). In the area of the imaging tools
concerning SIJ, conventional radiography, computer tomography and scintigraphy cannot be considered as appropriate diagnostic tools, since they have a poor sensitivity for detecting common pain provocation problems in the SI (such as arthritis) (Albert et al., 2006), while Braun et al. (2000) stated that other modalities such as MRI had a much higher sensitivity to detect early degenerative changes around the SIJ.

Thus, the common way for assessing an SI or lumbar pain problem is still the physical examination techniques, as stated in the guidelines.

The Hogeschool van Amsterdam (Hva), also known as Amsterdam School of Health Professions, uses in the ESP (European School of Physiotherapy) program module books to guide their students through the study. These module books are available on the ESP’s intranet, and include the contents of the modules as well as articles and guidelines about the topics taught. In the subject of Assessment in the first study year, the European guidelines about the diagnosis and treatment of pelvic girdle pain and the lumbar guidelines are introduced to the students, in which several tests meant to test pathology of these areas are introduced. Although necessary, this fact does not entail that all of them can be judged as valid tools for a student to include them in his/her “Physiotherapy practice toolbox”. This renders the students’ choice for the most appropriate assessment tools difficult. Moreover, many of the guidelines written for these two joints are quite extensive and it is not possible for a person to keep in mind all the information available for each time to use it. A useful flow chart for the algorithmic approach of distinguishing between an SI and / or lumbar pain is also missing.

Therefore, the purpose of this systematic review is to summarize the available evidence about the diagnostic accuracy of the physical diagnostic tests included in the module books of the HvA concerning the above mentioned joints. It is intended that comprehensive conclusions about the best documented (i.e., evidence based) tests can be found.

1.2 The mysterious Sacroiliac Joint and Evidence Based Research

Diagnosis and management of disorders of the lumbosacral spine is a source of much confusion and controversy. Before Mixter and Barr in 1934 introduced the herniated nucleus pulposus as a possible mechanism for low back pain, the Sacroiliac joint (SIJ) was one such theoretical mechanism (Vleeming et al., 1992). At present the model of the
herniated disc mediating all forms of low back pain (LBP) has been discounted (Vleeming et al., 1997).

A possible pain mechanism associated with sacro-iliac joint comes from a number of areas such as muscle imbalance, strains/sprains of the soft tissues in that area or tears. A literature review (Vleeming et al., 1995) dealing with the delineation between pain in the SIJ and lumbar area proves this attempt to be a difficult task since:

1) Pain patterns can be very similar in nature to symptoms of a herniated lumbar disc.
2) The nature of activity causing the injury can be very similar to the activities causing LBP.
3) There is no reliable and valid test procedure to determine the extent of SIJ dysfunction (Laslett et al., 1994).
4) There is a non-standardized approach in testing.

Most common cause of SIJ pain
Sacroilitis (inflammation in the sacrum) is the most common form of SIJ pain (Bernard et al. 1992), but the exact mechanism causing inflammation is not clear. Muscular dysfunction leading to a syndrome of chronic disuse or intense trauma or shearing blow of the SIJ could also cause degeneration to the connective sheath around the SIJ.

Incidence and prevalence of SIJ dysfunction
Bernard back in 1987 estimated that incidence of the SIJ dysfunction to be 22.5% (250 patients) in a sample of 1,200 patients reporting LBP. Cibulka (1992) studied 88 patients with chronic LBP. From this sample it was estimated that the 81% of them (71 patients) had some evidence of SIJ dysfunction. Concerning the prevalence of the SIJ dysfunction, Toussaint (1999) estimates a boundary of 19.3%-47.9% depending on the population being studied (for example, construction workers, pregnant women etc). An important fact to keep in mind is that females suffer more often from SIJ pain than males (Walker et al., 1992). Also, although there might be a notion that SIJ problems might be prevalent in one ethnicity more than in others, a quick review of the literature revealed no such relevant study being conducted apart from one that found that bridging of the SIJ is ethnicity independent (Dar et al., 2005).
In this paper we will review literature concerning men and women but excluding pregnant and post-partum women (women just right after they give birth), since SIJ pain in pregnancy is related to structural changes and altered distribution of forces. For example, in pregnancy SIJ pain can co-exist with other sources of LBP, a fact very rare in the rest of the population (Gutke et al., 2006).

Nowadays it seems that many researchers (Vleeming et al., 1995) and clinicians started again to look into the importance of the SIJ as a cause of pain in the lumbosacral area (Vleeming et al., 2007). It is believed that the dysfunction of this joint was overlooked for many years and commonly mis-diagnosed. This leads us to the obvious question: What are the useful clinical tests to evaluate SIJ? Unfortunately, no solid consensus as to the most appropriate tests exists.

Evidence based problems concerning the testing of the SIJ dysfunction

As it can be concluded from the above, it is very difficult to establish a precise procedure for SIJ testing. According to Albert et al. (2000) the classification system for SIJ pain must include clinical tests that can facilitate separation of pelvic from low back pain. They also must have a high inter-examiner reliability, sensitivity and specificity.

Till so far the main types of tests used in the pelvic region are:

1) Topographic-palpation tests: They detect possible anomalies in the pelvic alignment.
2) Pain provoking tests: They aim at stressing structures in order to reproduce the patient’s symptoms. Östgaard et al. (1991) argues that only pain provoking tests yield the required objectivity and reproducibility. In the literature so far there is much of debate concerning the predictive value of the tests used for SIJ dysfunction. Dreyfuss et al., (1996) conclude that there are no such tests, while Broadhurst et al. (1998) and Laslett & Williams (1994) indicate some possible predictive tests for SIJ dysfunction.

1.3 Lumbar Spine and another mysterious dysfunction: unspecific Low Back Pain

In comparison to the aforementioned SI joint, it is obvious that the lumbar joint is another joint that has complicated structure and causes lots controversy around the most common dysfunction of it: low back pain. One problem in defining low back pain is the various symptoms that patients might report. Malliou et al. (2006) define that the description of low back pain (LBP) can be in terms of its intensity (how much pain), its quality (e.g.
whether dull, etc.), and its anatomical location. Furthermore definitions based on the anatomical regions of the pain, or the duration of symptoms are used.

Under the terms of the European Guidelines (2004), LBP is defined as ‘pain and discomfort, localized below the costal margin and above the inferior gluteal folds, with or without leg pain’. Thus, they choose to use anatomical orientation in defining the unspecific low back pain.

Low back pain can also be categorized as specific or non-specific, and acute, sub-acute or chronic low back pain (Vleeming et al, 2007), with acute or chronic referring to the duration of pain and not its severity (Malliou et al., 2006).

What is of interest is the distinction between specific and non-specific low back pain. Specific lumbar pain is when the pain is attributed to recognizable, known specific pathology (e.g. infection, tumour, osteoporosis, ankylosing spondylitis, fracture, etc) (Vleeming et al, 2007), whereas non-specific back pain defined as low back pain that does not have a specified physical cause (Bekkering et al, 2003).

Low back pain can also be classified according to its persistence as either acute (0–6 weeks’ duration), sub-acute (7–12 weeks’ duration) or chronic (> 12 weeks’ duration) (Bekkering et al, 2003). Malliou et al (2006) add that episodes of pain that recur within periods less than 3 months in duration are not classified as acute or chronic but can be best defined as recurrent back pain.

**Incidence and prevalence of LJ dysfunction**

In the last decades, the incidence of low back pain has increased dramatically, at a rate disproportionate to all other health conditions (Frymoyer et al., 1992). The KNGF Guidelines (Bekkering et al, 2003) state that at least once in their lives, 60% to 90% of the population will experience low back pain, and the pain may be continuous or intermittent, with the first episode usually occurring between the ages of 20 and 55 years. The factors contributing to this disability are psychosocial and work environmental factors, exceeding physical factors (Frymoyer et al., 1992). In about 85% of people, pain cannot be attributed to pathology or neurological encroachment (Albert et al, 2006). Environmental risk factors for low back pain are heavy physical work, frequent bending, twisting, lifting, pulling and pushing, repetitive work, static postures and vibrations; while psychosocial risk factors include stress, distress, anxiety, depression, cognitive
functioning, and pain behavior job dissatisfaction and mental stress at work (Albert et al, 2006).

Another major factor is the duration of the current episode (Frymoyer et al., 1992). If the complaints of patients continue the psychosocial factors will have more impact on the disabilities due to low back pain than the biomedical or biomechanical factors (Bekkering et al, 2003).

It must be noted that the lumbar disability seems to be one of the few well documented dysfunctions related to psychosocial problems, reported as ‘yellow flags’ in the guidelines.

Prognosis / diagnosis and Evidence based problems concerning the testing of the LBP

In general, 80%-90% of cases, the patients’ complaints reduce spontaneously within four to six weeks, so a natural course of low back pain is favorable (Bekkering et al, 2003). Nevertheless, recurrent low back pain is common, and a smaller group of patients (about 20%) are prone to the development of chronic low back pain, presenting complex psychosocial and occupational problems (Deyo et al., 1996).

According to Tousignant et al. (2005), LBP management is a challenge for clinicians. Malliou et al. (2006) ally to this statement, affirming that it is very difficult to categorize the chronic low back pain (CLBP) patients, as there are many factors used to characterize the symptoms, such as pain intensity, pain duration, number of recurrent episodes and more as stated in the introduction.

Subsequently, the three essential components of assessing the pain in CLBP patients are the description of the CLBP, the impact of the CLBP on the patient’s life and the responses of the CLBP. (Malliou et al., 2006)

Agreeable to Tousignant et al. (2005), the three levels of dysfunction (level of impairments, activity limitations and participation restrictions) must be addressed in the assessment, taking the assessment of the range of motion as a key in the identification of impairment.

In order for these levels though to be assessed in a valid and reliable way, several clinimetric tools are needed, but unfortunately evidence based research on such an important theme is lacking.
2. Methods

Search Strategy
The literature searching strategy was computer based and included the following databases: PubMed, PEDro, CINAHL, and Cochrane library. The keywords used were the names of the tests being included in the module books and guidelines (c.f footnote of next page), including also the following combinations of words: ‘sacroiliac joint/lumbar joint examination’, ‘assessment of the lumbar/sacroiliac joint’, ‘validity measures’. In addition, the references found in relevant studies were also examined. Literature had to be in the English language and published in an international journal.

Study Selection
Studies were selected for inclusion in this review if they assessed the sensitivity, the specificity and/or predictive values of physical examination tests of the SI joint, when compared to the golden standard of a SI joint block (injection). The same was attempted for the Lumbar joint. The articles discussing the reliability of the two joints testing were excluded unless they included data concerning the validity of the tests. The population of interest had to be patients that have mechanical low back pain (of likely SI joint origin). Pregnant patients or articles describing patients with degenerative joint disease or other malignancies (systemic diseases) were excluded. The inclusion and exclusion criteria for the articles are stated in Table 1.1.
Table 1. Criteria for inclusion/exclusion of articles dealing with the sacroiliac joint

<table>
<thead>
<tr>
<th>Inclusion Criteria</th>
<th>Exclusion criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full reports (Abstracts, letters were not selected)</td>
<td>No review studies were included</td>
</tr>
<tr>
<td>All relevant clinical tests of the module books/European guidelines were included</td>
<td>Presentation of only test positive findings</td>
</tr>
<tr>
<td>Publication date &gt;1990</td>
<td>Publication date &lt;1990</td>
</tr>
<tr>
<td>English language</td>
<td>Case studies</td>
</tr>
<tr>
<td>Discussion of the reliability and validity of at least one of the relevant SIJ and/or lumbar tests</td>
<td>Studies investigating sacroiliac/lumbar joint dysfunction in pregnant women or patients with systemic diseases</td>
</tr>
<tr>
<td>Use as golden standard injection blocks</td>
<td></td>
</tr>
</tbody>
</table>

Assessment of studies

The studies were independently selected and rated by the two reviewers. The methodological quality of the studies was assessed through the PEDro rating scale.

The tests entitling our research were: Gaenslen’s test (also referred as pelvic torsion test), Provocation of sacrotuberal and sacrospinal ligament, Sit to supine test, Gapping test, Approximation test, Sacral apex pressure test, Palpation of the long dorsal SI ligament, Palpation of the symphysis, Active SLR and modifications, Schober Test, Quadrant Test, Segmental Instability test, Tests for "Malingering", Test on intermittent claudication, Foramina compression test, Distraction test, Test for vascular sign, Vertebral artery test, Patrick’s sign (also referred as the Faber test), Menell’s test, and the PPPP test (posterior pelvic pain provocation, also referred as Thigh Thrust, SI aggravation test or Thigh thrust test). Other possible tests being reported in the literature were checked for potential good scientific evidence.

All of the studies were subjectively rated by the authors. The reason for choosing PEDro scale was due to the fact that there is no other better standardized evaluation for methodology ratings to the best knowledge of the authors. As stated clearly in the PEDro
official site it should not be used as a measure of the “validity” of a study’s conclusions. Though, it is a good way to categorize studies and obtain an overview about the standardized procedures each researcher followed. Moreover, use of the PEDro scale led to a score out of 10 points for each study and allowed for a method quality comparison between the included studies. The PEDro score is determined by counting the number of checklist criteria that are satisfied in the trial report. PEDro scale includes 11 items and according to this scaling an article under review can score from 0 points (fulfilling no quality criteria of the PEDro Scale) to 10 points (fulfilling all the quality criteria of the PEDro Scale), since the first item on the PEDro list is not an item that can be graded (c.f, PEDro official website). Data were reported in the form of tables illustrating the predictive values, the specificity, sensitivity of the different tests (Tables 3 & 4) as well as a table of the methodology assessment sheet (Table 2).

Procedure
The reviewers were blinded for author(s), source of publication, results and conclusions in order to minimize potential reviewer bias. Per study the sensitivity, specificity and positive predictive value is reported.

Sensitivity is the proportion of true-positive patients with disease who test positive. A test that can correctly identify every person having the target dysfunction has a sensitivity of 1.0. Specificity is the proportion of true-negative findings of patients without the dysfunction. A test that can correctly identify every person who does not have the target dysfunction has a specificity of 1.0. The positive predictive value of a test is defined as the proportion of patients with a positive test who actually have the dysfunction. A positive predictive value of 1.0 would include all of the people who tested positive and actually had the target dysfunction. The negative predictive value of a test is defined as the proportion of patients with a negative test who really do not have the dysfunction. A negative predictive value of 1.0 will include all the people who tested negative and indeed do not have the dysfunction (Roach et al., 1997).
Statistical analysis

For the comparison of the tests in the different factors reported (validity and predictive values) descriptive analysis of the data was performed using Excel program and SPSS version 15.0.

3. Results

Search results for SIJ

An initial search yielded twenty eight papers dealing with SI joint dysfunction (excluding three systematic review papers). The majority of this literature was referring to studies concerning the reliability of testing the different diagnostic tools for SI joint. Two of them were studying pregnant women as target population, while only six of them were following the inclusion criteria set by the authors. From the literature outcome five papers fitted our previously mentioned inclusion criteria. From those the study of Maigne (1996) might have included injection blocks but did not include calculations of validity. Thus, the remaining literature was 4 articles (Table 2).

Table 2. Studies included in the present review- descriptive information concerning the study population and the methodology quality of each paper.

<table>
<thead>
<tr>
<th>First Author, Year</th>
<th>Total number of patients studied</th>
<th>Gender F/M</th>
<th>Age</th>
<th>PEDro score (Mean values of 2 independent reviewers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dreyfuss, 1996</td>
<td>85</td>
<td>61/24*</td>
<td>18-87 (median 44.5)</td>
<td>3</td>
</tr>
<tr>
<td>Broadhurst, 1998</td>
<td>40</td>
<td>30/10*</td>
<td>35 (Average)</td>
<td>8</td>
</tr>
<tr>
<td>Slipman, 1998</td>
<td>50</td>
<td>31/29*</td>
<td>43 (Average)</td>
<td>2</td>
</tr>
<tr>
<td>Laslett, 2005</td>
<td>48</td>
<td>32/16*</td>
<td>42 (Median)</td>
<td>7</td>
</tr>
</tbody>
</table>

*successful technically injections in the number of patients reported
**Methodological Quality**

Table 2 contains the methodological quality information and some descriptive information concerning the selected studies. As it can be seen half of the studies’ quality is low as rated by the PEDro scale. The study of Broadhurst et al. (1998), and the study of Laslett et al. (2005) had the highest methodological quality score. The study of Broadhurst et al. (1998) was the only one that employed placebo and active injection groups and used also double blinding of the researchers. The studies of Maigne et al. (1996) (not included) and Laslett et al. (2005) blinded the doctor performing the injections.

**Predictive value**

The positive predictive value and the negative predictive value are reported in Table 3. As it can be seen reporting on Laslett et al. (2005), none of the tests exhibited a positive predictive value of more than 60% whereas the negative predictive values were higher. Laslett et al. (2005) and Slipman et al. (1998) showed that a history and clinical examination with at least three positive tests has a positive predictive value of 60%, whereas Laslett et al. (2005) found that having two out of four SI joint tests positive led to positive and negative predictive values of 67% and 93% respectively. What is interesting in the study of Laslett et al. (2005) is that they found a gradual decrease in positive and negative predictive values with an increasing use of positive SI joint tests (more than three).

**Table 3. Predictive values of individual tests (all values converted in percentages)**

<table>
<thead>
<tr>
<th>First author, Year</th>
<th>Test</th>
<th>Positive Predictive value</th>
<th>Negative Predictive value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laslett, 2005</td>
<td>Compression</td>
<td>52%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>Distraction</td>
<td>60%</td>
<td>81%</td>
</tr>
<tr>
<td></td>
<td>Gaenslen’s (R) test</td>
<td>47%</td>
<td>76%</td>
</tr>
<tr>
<td></td>
<td>Gaenslen’s (L) test</td>
<td>50%</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>Sacral thrust test</td>
<td>56%</td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td>Thigh thrust - 4P test</td>
<td>58%</td>
<td>92%</td>
</tr>
</tbody>
</table>
Sensitivity

Table 4 shows the sensitivity of individual tests reported in the studies chosen for this review. The results of several tests can be compared from one study to another (Figure 1). When the test was mentioned by more than one study then the mean value and standard deviation was calculated.

As seen from Table 4 and Figure 1 the most sensitive test was found to be the palpation of the long dorsal ligament (89%), with the compression and the 4P test following with 69% of sensitivity. The test scoring the least in sensitivity was the Gillet test with 47%. The differences in the reported values within the same test could be due to different pain relief standards each study used required for the injection to be positive. For example, Broadhurst used >70% pain relief standard, Laslett >80%, while Dreyfuss 90%.

Specificity

Table 4 depicts the specificity of the individual tests evaluated in the various studies. Again the results of several tests can be compared from one study to another (Figure 1). The most specific test for SIJ dysfunction was found to be the gapping test (81%), with the 4P and the sacral sulcus test to follow (80% and 75% respectively). In order for the reader to have an overview of which tests score higher in both sensitivity and specificity we calculated the mean of those two components and then standard deviation (Table 5). The standard deviation enables the reader to see which tests had the most congruent results in both factors, i.e., their results did not differ much for both sensitivity and specificity. It deserves to be mentioned that 4P test was the only test scoring high in both sensitivity and specificity factors (M = 74,50%, SD = 7,77) with the distraction test following, although is standard deviation is higher (M = 70,50%, SD = 14,84) (see Figure 2 & Table 5). The compression, sacral thrust and Patrick test follow with identical values almost, but compression test has no difference in the percentage means of sensitivity and specificity. On the other hand, one could argue that the Patrick test was rated in two papers, which therefore could explain much of its final standard deviation. A closer look though between the studies testing this particular test shows that already the Patrick test has high scores of standard deviation within each factor (sensitivity-
specificity). Concerning sacral thrust the deviation of it makes it a less better candidate than compression test.

**Table 4.** Sensitivity and specificity of individual tests (all values converted in percentages)

<table>
<thead>
<tr>
<th>Test</th>
<th>First author</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaenslen test</td>
<td>Dreyfuss</td>
<td>68%</td>
<td>28%</td>
</tr>
<tr>
<td></td>
<td>Laslett</td>
<td>53% (R) 50% (L)</td>
<td>71% (R) 70% (L)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(M = 57%, SD = 9.64)</td>
<td>(M = 56%, SD = 24.54)</td>
</tr>
<tr>
<td>Gillet test</td>
<td></td>
<td>47%</td>
<td>63%</td>
</tr>
<tr>
<td>Compression – Approximation test</td>
<td>Laslett</td>
<td>69%</td>
<td>69%</td>
</tr>
<tr>
<td>Distraction test - Gapping</td>
<td>Laslett</td>
<td>60%</td>
<td>81%</td>
</tr>
<tr>
<td>Patrick – Faber test</td>
<td>Dreyfuss</td>
<td>64%</td>
<td>19%</td>
</tr>
<tr>
<td></td>
<td>Broadhurst</td>
<td>77%, 50%</td>
<td>100, 100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(M = 64%, SD = 13.50)</td>
<td>(M = 73%, SD = 46.76)</td>
</tr>
<tr>
<td>Palpation of the long dorsal ligament test</td>
<td>Dreyfuss</td>
<td>89%</td>
<td>14%</td>
</tr>
<tr>
<td>4P / thigh thrust test</td>
<td>Dreyfuss</td>
<td>39%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Broadhurst</td>
<td>80%, 69%</td>
<td>100%, 100%</td>
</tr>
<tr>
<td></td>
<td>Laslett</td>
<td>88%</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(M = 69%, SD = 21.46)</td>
<td>(M = 80%, SD = 24.63)</td>
</tr>
<tr>
<td>Sacral thrust</td>
<td>Dreyfuss</td>
<td>63%</td>
<td>75%</td>
</tr>
</tbody>
</table>
**Figure 1.** Percentage scores for sensitivity and specificity concerning the eight tests mentioned in the selected papers.

**Table 5.** Mean validity scores for each test mentioned in the selected literature

<table>
<thead>
<tr>
<th>Test</th>
<th>Mean</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaenslen test</td>
<td>56.50</td>
<td>.70</td>
</tr>
<tr>
<td>Gillet test</td>
<td>55.00</td>
<td>11.31</td>
</tr>
<tr>
<td>Compression – Approximation test</td>
<td>69.00</td>
<td>.00</td>
</tr>
<tr>
<td>Distraction test - Gapping</td>
<td>70.50</td>
<td>14.84</td>
</tr>
<tr>
<td>Patrick – Faber test</td>
<td>68.50</td>
<td>6.36</td>
</tr>
<tr>
<td>Palpation of the long dorsal ligament test</td>
<td>51.50</td>
<td>53.03</td>
</tr>
<tr>
<td>4P / thigh thrust test</td>
<td>74.50</td>
<td>7.77</td>
</tr>
<tr>
<td>Sacral thrust</td>
<td>69.00</td>
<td>8.48</td>
</tr>
</tbody>
</table>
**Search results for LJ**

An initial search yielded almost no results dealing with evidence based assessment of Lumbar joint dysfunction (including literature concerning systematic review papers on this area). The majority of the literature was referring to studies concerning the reliability of testing some of the lumbar assessment tools and the extension McKenzie techniques, while two articles were found concerning the validity of the PosteriorAnterior manipulation (PA). Most articles found were dealing with pregnant women and some other neurological pathologies. The active straight leg raising (ASLR) test is often discussed, but in relation again to pregnancy, which was an exclusion criterion for this study. Nevertheless, if used to diagnose posterior pelvic pain since pregnancy, the ASLR had a sensitivity of 0.87 and a specificity of 0.94 (Mens et al., 2001). According to Mens et al. (2001), the test is a suitable diagnostic instrument, it is easy to perform; reliability, sensitivity, and specificity were also high.

For the modified-modified Schober Test, Tousignant et al. (2005) found that it has a moderate validity ($r=0.67$).
4.1 Discussion of the SIJ findings

In general, as Riddle et al. (2002) report, it is quite obvious that for a precise diagnosis of a possible SIJ region dysfunction, there has to be a set of appropriate tools indicated for this diagnosis. Failure of this attempt will lead to incorrect delivery of therapeutic interventions. In this case the individual’s problem theoretically could be exacerbated.

From the results of this review, it is clear that there are some tests appearing to have higher sensitivity and specificity than others. It would seem prudent to include tests with higher sensitivity and specificity in an examination where SI joint dysfunction is among the differential diagnoses. As van der Wurf et al. (2000) noted a test should have higher values of both in order to be of use, in addition to being reliable. In this case from this review it can be concluded that the best suited tests that scored high in both sensitivity and specificity are the 4P, with the distraction test, the compression and sacral thrust following.

Although we found high validity scores concerning the compression and separation tests others do not and they report low sensitivity and overall efficacy (Albert et al., 2000). McCombe (1989) recommended using these tests with caution, while Albert (2000) goes even further by recommending discarding these tests in favour of more reliable tests. On the other hand, Kokmeyer et al. (2002) choose compression and separation tests to be the most reliable ones. This discrepancy in opinions can be due to the fact that for example, Albert et al. (2000) studied the specific tests on pregnant women. This can potentially lead to some confounding factors that this review tried to avoid by not including such ‘special’ populations. On the other hand, Kokmeyer et al. (2002) investigated the SIJ dysfunction in normal male population (construction workers).

Only five tests had both sensitivity and specificity greater than 60% in at least one study (a value of 50% or lower for a test has been deemed not acceptable by some authors (van der Wurf et al., 2000)): compression, distraction, Patrick, 4P and sacral thrust. Most of these tests were supported by studies with very high methodology scores (Broadhurst et al., 1998, Laslett et al., 2005). Unfortunately the positive predictive values for the above mentioned tests were not extremely high; their negative predictive values were at least 80% in the study by Laslett et al. (2005). From the results of this review it appears that employing a group of tests in order to diagnose SI joint seems to be desirable, as Laslett...
et al. (2005), demonstrate that employing this approach results in adequate sensitivity, specificity and predictive values. Slipman et al. (1998) demonstrated an acceptable positive predictive value when three out of five tests were positive. Cibulka et al. (1999) also states that a combination of tests improves specificity reducing thus, the number of false positive results.

Another point that has to be considered is the ‘golden standard’ criterion that was chosen for the evaluation of the studies. As stated in Kokmeyer et al. (2002) it is difficult to make a statement regarding the validity of the SIJ tests with a golden standard that fails to include all the structures that these tests are intended to evaluate. In other words, it is difficult to find a diagnostic tool that can be set as a golden standard, simply because the SI joint comprises of many structures and the injection block cannot at once have an effect on all of them (for a comprehensive discussion on this topic see Berthelot et al. (2006)). Until today this golden standard for the SIJ is missing, thus it leaves the diagnostic injection blocks as the most valid way of diagnosing pathological SIJ conditions.

Shortcomings in the literature

Aside from low methodological quality of some of the included studies, one of the shortcomings from most of the studies was the lack of single or double blinding, confirmatory blocks, and randomized controlled trials. Inconsistency in the definition of a positive response to the joint block (>70% to >90%) was also a detriment. Another noticeable issue was that of selecting the clinical tests employed in each study. The protocol of Dreyfuss et al. (1996) was the most reasonable since they conducted a pre-screening of the tests to find the most reliable ones. This limits the bias of authors selecting the tests.

It is important to note that Dreyfuss et al. (1996) and Laslett et al. (2005) have found much different results concerning the validity of various tests. This fact can be attributed to different force being exerted by the examiner. Cibulka et al. (1998) gives another dimension into this fact by stating that individual SI joint tests can be unreliable due to anatomic variation in bony prominence, excessive adipose tissue and due to the different skills of each examiner. This topic though ends up concerning inter ratter reliability, which is not goes beyond this review’s ambitions.
Conclusions

There is not enough evidence to support the use of most clinical SI joint tests, yet the studies reported in this review can be considered of the highest quality due to their standardized procedure and the use of the same ‘golden standard’ measure. It appears that the best practise to date is the combination of a number of provocative tests with a minimum number of positive tests (two out of four or three out of five etc.) needed for the diagnosis of SIJ dysfunction. The 4P, the sacral thrust, the compression and distraction test maybe likely candidates to include in such an examination since they have relatively high scores of validity (above 60% in at least one study).

In general there is need for more research on the SIJ topic. The future studies have to be improved and used strict criteria including large scale placebo randomised trial, including also a control group. Another important issue is that apart from the validity, the study protocol must also include the assessment of the reliability of the selected tests.

4.2 Discussion of the Lumbar Joint findings

Much to the researchers’ surprise not enough evidence based literature can be obtained for such an important topic in everyday practice, such as the low back pain. Our results agree with Magee (2008). In the new edition of Magee no results concerning the tests being mentioned in our module book were included apart from Schober and SLR. It must be mentioned though that Magee reports the validity for other lumbar tests mentioned in his book and students are encouraged to use the ones in Magee with the highest sensitivity and validity ratio. Moreover, concerning validity of lumbar tests the guidelines report no such studies as well. In accordance, we develop in the next paragraph some possible problems concerning clinimetrics in the lumbar joint and potential limitations a researcher has to keep in mind.

Evaluation of the studies and problems in the attempt to develop clinimetrics in lumbar spine dysfunction

To assure the validity and reliability of tests described in a study, many factors have to be taken into consideration. Due to the above mentioned parameters of classification, it is
very difficult to categorize patients into groups with homogeneity; nevertheless it is necessary to compare intervention programs.

Malliou et al (2006) suggest the following implications in order to provide a methodical study:

- The samples have to be defined as acute, subacute or chronic LBP patients, and then select the participants according to the pain characteristics

- To assure the homogeneity of the groups, the clinical diagnosis needs to be established clearly

- The selection criteria need to be considered in relation to the specific aim of the study, concerning the age variation, gender, inclusion and exclusion criteria and so on.

Frymoyer et al (1992) stated that one of the major barriers in LBP research is that experimental results between investigative groups are not comparable, as important data and results are not collected in a systematized model, and make it therefore hard to interpret the published data.

Malliou et al (2006) suggest in order to achieve a complete evaluation of LBP patient’s capability, both self-report (such as interview and questionnaires) and actual functional performance measurements is recommended to be used. Functional tests are reliable and inexpensive, and if the appropriate tests are chosen according to specific parameters (such as age, gender and physical activities) the data will be reliable, valid and useful for the clinician. Another problematic area in the research of lumbar spine dysfunction is the establishment of one specific golden standard. The golden standard for LBP varies depending on if it is specific or non-specific LBP. Concerning pain assessment, the self-report is considered by authorities to be the golden standard of pain measurements (due to its consistency with the definition of pain). Nevertheless, pain is subjective, which causes a problem regarding its use. (Malliou et al., 2006)

With ankylosing spondylitis, the golden standard is believed to be X-rays, while Kennedy et al. (1995) state that the radiology scores, although termed as golden standard, are unreliable. In their opinion, the Bath Ankylosing Spondylitis Metrology Index (BASMI) may be judged to be more important in assessing ankylosing spondylitis and become a golden standard. For the Modified-Modified Schober Test, Tousignant et al. (2005) state as the golden standard X-rays. Furthermore, as a general statement, these authors declare
radiography (X-rays) to be the best golden standard to establish the criterion-related validity of goniometric measurements. The tests for malingering have as a golden standard MRI testing.

To conclude with, testing in lumbar spine is complicated, and not much evidence was found to underlay the tests with proper objective results concerning their validity, reliability, specificity and sensitivity. Further research is needed to analyze and discuss the clinimetrics of the tests used for lumbar spine diagnostics.

4.3 Discussion - Differential diagnosis of the Lumbosacral area: Is it a myth or is it feasible?

From the literature being reviewed it seemed that the topic of the SIJ has risen in interest the last few decades, maybe due to the fact that it can be a significant cause of low back pain. In the expense of this other possible causes of lumbar dysfunctions which several tests try to address (for example, instability) are not very well documented (for an excellent study concerning descriptors of lumbar instability c.f., Cook et al., 2006). Thus, the potential question of whether one could differentiate through testing which areas are involved in a low lumbosacral area pain it cannot be answered by the present literature. One interesting algorithm is proposed by Laslett et al. (2005) concerning SIJ provocation is proposed (Figure 3). Thus, anamnesis starts becoming a very important tool in a physiotherapist’s toolbox. Of course, one can argue that once a physiotherapist applies the SIJ evidence based indicated tests and they prove to be negative, then logically the pain can be attributed to the lumbar area. This way of thinking though includes a whole mark since the lumbar involvement in pain can not still be ruled in (since the pain can be from a referring source for example).

Moreover, even though the SIJ seems to have been studied thorough and being well documented with standard procedures it might not be so. For example, standards used in recent studies of SIJ pain diagnosis are diverse. The International Association for the Study of Pain (IASP) has proposed criteria for making the diagnosis of symptomatic SIJ and are: (1) pain is present in the region of the SIJ, (2) stressing the SIJ by clinical tests that are selective for the joint reproduces the patient’s pain, (3) selectively infiltrating the putatively symptomatic joint with local anaesthetic completely relieves the patient of pain (Merskey and Bogduk, 1994). It is noted that publication (Bogduk and McGuirk, 2002, p.
double comparative blocks are recommended for confirmation of the diagnosis. Although false positive rates for SIJ injections have not been previously reported, a rate of 7.7% may be calculated from data presented from one study (Schwarzer et al., 1995) and 20.5% from another (Maigne et al., 1996).

Another procedure that also can be criticized is the injection block. This is the only available criterion standard against which clinical tests can reasonably be evaluated for validity; it is acknowledged that false negative and false positive responses to injection are possible. For example, as Laslett reports, that when there is a defect in the articular capsule, leakage of anaesthetic into adjacent areas may occur (Fortin et al., 1994a; Schwarzer et al., 1995), and pain relief may be a reflection of an anaesthetic affect of these structures rather than the SIJ structures. This possibility and the unknown effect of psychosocial influences on pain responses to invasive diagnostic procedures may contribute to the false positive and negative rates. In addition, intra-articular injection of anaesthetic has the potential to ablate SIJ pain when originating within the joint cavity, but is unlikely to have an anaesthetic effect on SIJ structures external to the joint (Grieve et al., 1988, Maigne et al., 1996; Laslett et al., 2003). Where SIJ structures external to the joint cavity are actual pain generators, an intra-articular injection of local anaesthetic into and confined to the joint space will produce a false negative diagnostic result, whereas the clinical examination may possibly correctly identify the periarticular and unanaesthetized SIJ structures as pain generators.

5.1 Points of interest for the ESP students and teachers

Conclusions
From this literature review a few interesting points that students from the ESP program should keep in mind are the following:

General conclusions:
- The evidence based practice in physiotherapy is a new concept. The student should be eager though not to accept each research study with the same critical view. It depends on the journal that one study is published (and its impact factor that indicates how much of a scientific impact this journal has on the world). Moreover, as it could be seen sometimes in humanity sciences a choice of the best standardized research method (i.e., injection blocks) might not be the optimal and
in the best case there is one. In other cases this standard might be difficult to define and set.

- Evidence based practice is good and fashionable as long as it does not become the only goal. New suggestions and theories are more valuable and sometimes the experience of a teacher can be as valid. For example, the three phases testing while trying to differentiate between SIJ and lumbar joint dysfunction suggested by one of the teachers teaching assessment (Pim Ranzijn) is an interesting concept and practice, but it was not found in any literature being searched.

- In general there is more need for evidence to support the SIJ tests, which should include strict criteria such as a large scale placebo randomized trials, control group, and next to validity, the protocol should also include the assessment of the reliability of the tests. Nevertheless, the studies reported in this review can be considered of good quality due to their standardized procedure and the use of the same golden standard. The best practice for the diagnosis of SIJ is to date the combination of a number of tests, including the ones recommended by this review.

Specific SIJ and LJ conclusions:

How could this systematic review change the assessment practice?

- First of all a good example of keeping always a critical mind in all evidence based literature is the Guidelines practice. Guidelines are referred and appreciated by each physiotherapist but they still always should be checked for updates and to which patient group they refer to. In our literature review case the guidelines for the SIJ were suggesting that palpation of the long dorsal ligament, the Gaenslen’s test, Trendelenburg and symphysis palpation should be included among the testing for SIJ dysfunction. This is not though congruent with our results. One possible explanation is that the guidelines incorporated also literature concerning pregnant women and in this field it is indeed the case that for example the palpation of the symphysis scores high in sensitivity (Albert et al., 2000). It must be noted that even the highest in quality papers of our review do not recommend or test the above mentioned tools for detection of SIJ dysfunction.
Concerning the lumbar joint guidelines it was a surprise that they fail to systematically produce a comprehensive algorithm for testing and excluding various possible sources of lumbar dysfunction. Testing and validity of it is completely missing from the lumbar guidelines.

Can we differentiate between lumbar and sacroiliac dysfunction?

• In the absence of evidence based literature for the specific lumbar tests we searched it would premature to recommend a possible algorithm for that. Thus, there is an interesting open question of whether the tests used in the module book concerning lumbar testing should be still kept in the module books. What a careful reader would recommend would be another literature review concerning testing that is mentioned in other books (such as Magee) and then maybe a proposal of the most valid ones can arise.

• Till then, the only good and evidence concrete algorithm we could find (instead of constructing due to the mentioned problems above) is the following by Laslett et al., 2005 (Figure 3). We quote him as follows:

Concerning the SIJ best tests and in which order to use them:

‘Because the thigh thrust and distraction tests have the highest individual sensitivity and specificity, respectively (see Table 5 in our case), performance of these tests first seems reasonable. If both tests provoke familiar pain, no further testing is indicated. If one test is positive, the compression test is applied and if positive, a painful SIJ is likely and no further testing is required. If compression is not painful the sacral thrust test is applied. If this is painful, SIJ pathology is likely, whereas if it is not painful, SIJ pain is unlikely. Not only does this rule avoid subjecting patients to unnecessary tests, but also would in most cases permit a diagnosis even if one or more tests were not completed presents a diagnostic algorithm for this reasoning process’.

Concerning pain not originating from the SIJ:

‘When severe pain occurs with all body movements (e.g. acute disc prolapse, fractures, etc.), pain is provoked by any test including the provocation SIJ tests. In these circumstances interpretation of the SIJ tests is inappropriate’. In Laslett’s opinion, where
another source of pain is known to be a major source of pain, the interpretation of the SIJ tests as evidence of a symptomatic SIJ should be avoided or entertained only with skepticism.

The most important thing that the study keeps in notice is that no single study can satisfy all criteria recommended by advisory groups (Deyo et al., 1994), and that their study (Laslett et al., 2005) is no exception.

This is something the physiotherapist student should always keep in mind, not in order to prevent him/her from creating some standard procedures in his/her testing, but in order to always be open minded and challenged.
Figure 3.
Diagnostic algorithm for SIJ pain using provocation SIJ tests: distraction, 4P/ thigh thrust, compression and sacral thrust. Laslett et al. (2007) developed this diagnostic algorithm which can be used in diagnosis. The algorithm starts with the distraction and thigh thrust tests, as according to the researchers those tests have the highest individual sensitivity and specificity. From these results on, a systematic reasoning process can be induced. With this help, the physiotherapist and the patient would avoid unnecessary tests (saving time and unnecessary strain in the area), and a diagnosis can be permitted even if one or more tests were not completed.
5.2 Manual of tests recommended from this systematic review

see Appendix 1

Acknowledgements

We express our gratitude to our coach Bob van de Berg for his sensible comments, his patience and understanding towards us. We are especially thankful to our client Pim Ranzijn that gave us the necessary inspiration on this topic and his support and trust in us. Special thanks also to our external quality advisor Konstantinos Kotoglou and our English teacher who corrected the use of the English language, Zisis Folias.

This report claims no interests.

References


PEDro Scale: http://www.pedro.fhs.usyd.edu.au/


Appendix 1

**Manual of tests recommended from this systematic review**

**Sacroiliac Joint (SIJ)**

**Compression / Distraction Test**

The patient is supine. Pressure is applied first in a posterior and lateral direction (compression) on the ASIS simultaneously. Pressure is then applied in an anterior and medial direction on the ASIS (distraction). If pain is reproduced in the SI joint region with either maneuver, the test is positive.

The compression test can also be performed with the patient lying on the side, with the affected side up and the back towards the side of the treatment table. The examiner places his hands over the anterior crest of the iliac crest, and presses towards the floor, to cause forward pressure on the sacrum. The test is positive if there is an increased feeling of pressure in the SI joints, indicating a possible SI lesion or sprain of posterior SI ligaments, or both.

*The compression tests scores equally high on sensitivity and specificity (69%), whereas the distraction test has a higher specificity (81%) than sensitivity (60%) (Laslett et al., 2005)*
Posterior pelvic pain provoking test (4P)

The patient lies supine. One leg is flexed 90° at the hip- and knee joints. The examiner places his hands on the raised knee, and exerts pressure down the femur into the pelvis. *In this study, the 4P scored with a mean of 69% on sensitivity, and a mean of 80% on specificity.*

Patrick’s Faber Test

The patient lies supine. One leg is flexed, abducted, and rotated out in order that the heel rests on the opposite kneecap (therefore the test’s name: FABER = Flexion, Abduction, and External Rotation). The examiner then slowly lowers the knee of the test leg toward the examining table. The test is positive if the test leg’s knee remains above the opposite straight leg, and does not fall to the table. This positive test indicates pain on the medial side of the knee and femur, or in the inguinal region, being caused by an affected hip joint, iliopsoas spasm, or an affection of the SIJ. *The Patrick Faber test scored in this study only a mean of 64% on sensitivity, yet scoring 73% on specificity.*
Sacral Thrust Test

The patient is prone. The examiner delivers an anteriorly directed thrust directly over the sacrum. The test can be considered positive if pain is reproduced in the SI region. This test has according to Dreyfuss et al. (1996) a sensitivity of 63% and a specificity of 75%.

Lumbar Spine

Straight Leg Raising (Lasègue’s) Test

Although this test is mainly considered for the neurological tissue around the lumbar spine, the test may also be used to provoke a stress on the SIJ. The patient lies in a supine position while the examiner passively flexes the patient’s hip with the knee extended. Pain occurring after 70° flexion (or in hypermobile persons, this may only occur after 120° flexion), is usually indicative of joint pain. Next, the range of motion should be compared to the unaffected side. The examiner then does a passive bilateral straight leg raising (SLR) test in a similar fashion, pain occurring before 70° is usually indicative of SIJ problems. If, when doing SLR, the pain in the SIJ is unaltered or decreases, the examiner may suspect an anterior torsion. If the pain increases in the SIJ, a posterior torsion is possible. If pain increases on the opposite side, an anterior torsion on the opposite side should be suspected.

Modifications:
The examiner stabilizes and compresses the pelvis while the patient actively does the SLR providing form closure of the joints by squeezing the innominate bones together anteriorly. If the pain decreases or the SLR is easier to do, the test is considered positive for possible SIJ problems.
This modification tests force closure at the SIJ. The patient is asked to flex and rotate the trunk toward the side that the SLR is actively being performed. The trunk motion is resisted by the examiner. The two sides are compared for any difference. Force closure tests the ability of the muscles to stabilize the SIJ during movement.

A: Single leg, B: double leg

(Modified) Schober Test

This test is used to measure the flexion occurring in the lumbar spine. The examiner marks a point between the posterior spina iliaca superior, so at the level of S2, then points 5 cm below and 10 cm above that level are marked as well. The distance between these three points is measured, the patient is asked to flex forward, and the distance is measured again. The difference between the two measurements is an indication of the amount of flexion occurring in the lumbar spine.

For this test, there exist some modifications: the original Schober Test only marked one point, 10cm above the lumbosacral junction. The Modified Schober Test applies a point 5 cm below and 10 cm above the junction. The Modified-Modified Schober Test uses the PSIS as opposed to the lumbosacral junction, and uses a 15 cm landmark.