Reliability and Validity of selected Pain Provocation Tests at the Sacroiliac joint.

INVESTIGATORS:

Dr. Brian Nicholls D.O., M.A.
PHONE: (03) 9248 1150
EMAIL: Brian.Nicholls@vu.edu.au

Joseph Thompson B.Sc. (Clinical Science)
PHONE: (03) 9585 3170
EMAIL: joeyt334@hotmail.com

CAMPUSS:

Victoria University of Technology, City Flinders Campus.
Faculty of Human Development, School of Health Sciences. Melbourne, Australia.
Reliability and Validity of selected pain provocation tests at the Sacroiliac joint.
Dr. Brian Nicholls D.O., M.A. & Joseph Thompson B.Sc. (Clinical Science)
Victoria University of Technology, Health Sciences. Melbourne, Australia.
Abstract
The sacroiliac joint (SIJ) is a true source of pain commonly tested in osteopathic practice. This study investigated the inter examiner reliability and validity of pain provocation tests at the SIJ. Two examiners tested fifty participants, mean age 23. Sixteen participants were symptomatic for sacroiliac joint dysfunction. Examiners were blinded to participant's inclusion criteria and examination findings. Cohen's kappa and percentage agreement were used to evaluate inter examiner reliability. Validity was measured using sensitivity, specificity, positive and negative predictive values. Percentage agreement between examiners was 76% and Cohen's kappa findings included a kappa = 0.475 overall. Total validity findings included; sensitivity 0.34, specificity 0.93, positive predictive value 0.70 and negative predictive value 0.75. Investigation identified a poor level of validity of pain provocation tests at the SIJ and of inter examiner agreement using compression and gapping tests. A good level of inter examiner agreement occurred using the thigh thrust and Faber test (k=0.674 and 0.611 respectively).

Key Words: Sacroiliac joint, diagnostic test, pain, reliability, and validity.
Introduction

Sacroiliac joint dysfunction (SIJD) describes pain in or around the joint with associated hypomobility throughout a portion of the joint's range of motion\textsuperscript{1}. This may be accompanied by a variation in the structural relationship between the sacrum and ilium\textsuperscript{2,3}. Despite being easily described, the sacroiliac joint has been a difficult area of diagnosis for manual therapists. The ability of the sacroiliac joint to undergo movement has also been controversial and although studies now demonstrate that small but significant movements occur at the SIJ\textsuperscript{4}, the arthrokinematics of the joint remain controversial. It is generally accepted within the medical community, however that the SIJ can be a source of pain\textsuperscript{1,2,4,5,6,7,8,9}. This has been supported by demonstrating patient pain relief following anesthetic injection into the joint, indicating that the SIJ is a true source of pain\textsuperscript{9,10}.

Pelvic levels, motion palpation tests and pain provocation tests are the three main categories for clinical tests at the SIJ. Few reliable outcomes have been demonstrated for mobility tests\textsuperscript{5,7} or pelvic position tests at the SIJ \textsuperscript{5,11}, some pain provocation tests however have been found to be reliable\textsuperscript{5,8,10,12,13,14}. These tests may produce higher levels of reliability due to the use of patient feedback during testing. Reliability of a test procedure is defined as the ability of examiners to agree when using the same test. Validity of a test procedure is defined as the ability of a test to determine the presence or absence of a pathology or condition.
Existing evidence of reliability and validity from pain provocation investigations at the SIJ.

Laslett and Williams\textsuperscript{12} tested 51 patients with low back pain. They found agreement between 86-91\% for the Gaenslen, thigh thrust, compression and distraction tests, kappa in the same study was found to be k=0.69-0.82 for iliac compression, iliac gapping, thigh thrusts and Gaenslen tests which showed a high level of agreement. This study had an adequate sample size and statistical analysis for reliability. Low back pain was used as an inclusion criteria however and this may have limited the validity of results.

Similarly McCombe et al\textsuperscript{14} used low back pain as the basis for their study. Findings included a kappa value of k=0.36-0.16 for the compression and gapping tests. These scores may be attributed to the use of weaker pain provocation tests as shown by other investigations. The findings from studies that utilise low back pain as an inclusion criteria however may still be valid. Schwarzer et al\textsuperscript{6} concluded after investigation that SIJD is a significant source of low back pain, and 15\% of low back pain is due to SIJD\textsuperscript{15}.

Potter and Rothstein\textsuperscript{5} found percentage agreement of 76\% for iliac compression and 94\% for iliac gapping at the SIJ. The investigation had a small sample size of 17, this may have inflated results. This study also used only chi-square tests as statistical analysis, which is a weak form of reliability. This level of statistical analysis limits the value of the results found.

Dreyfuss et al\textsuperscript{10} had a large sample size (85) and undertook adequate statistical analysis for both reliability and validity. These inclusions strengthen the results
found. Agreement level was calculated to be between 82-85% for the Gaenslen, thigh thrust and Patrick’s tests. Moderate kappa values were identified ranging from 0.61-64 for Gaenslen, thigh thrust and Patrick’s tests. Positive predictive value ranging between 0.72-0.96, specificity 0.16-0.50 and sensitivity of 0.36-0.71 when using the thigh thrust, Gaenslen and Faber tests.

Broadhurst and Bond\textsuperscript{16} used anesthetic joint blocks and a control (saline) on patients with a suitable pain referral pattern who had a positive Patrick’s sign and thigh thrust. A 70% decrease in pain from the anesthetic was set as the threshold for dysfunction within the SIJ. Findings included a positive predictive of 0.70-0.75, specificity 1.0 and sensitivity 0.77-0.8. This study was weakened by the exclusion of negative predictive values from the statistical analysis.

Slipman et al\textsuperscript{17} also used SIJ blocks in their investigation into the validity of pain provocation tests at the SIJ. A positive predictive value of 0.6 was calculated where 60% of patients who were found to have an SIJD had it confirmed by SIJ block. This study provided a positive predictive value but failed to provide negative predictive value, sensitivity and specificity weakening results.

Maigne et al\textsuperscript{9} undertook double anesthetic blocks on 54 patients. Ten participants felt 75% of pain relief. None of the compression, distraction, Gaensien, or Patrick’s tests were found to be useful predictor of pain provocation. The results found may have been influenced by the use of only the chi square test as statistical analysis. This study however had a good sample size and attempted to rule out low back dysfunction involvement by utilising lumbar X-ray and computer tomography. Patients were only accepted when pain
and tenderness was found over the posterior aspect of the SIJ. This may have increased the chance that the SIJ is the true source of pain in participants.

The level of pain relief is another thing to consider when interpreting the results from the investigations by Maigne et al\textsuperscript{9}, Dreyfuss et al\textsuperscript{10}, Broadhurst and Bond\textsuperscript{16} and Slipman et al\textsuperscript{17} that involved SIJ blocks. Each investigation set a different level of pain relief as criteria for inclusion respectively; 75%, 90%, 70% and 80%. An acceptable level for SIJD would most likely be 70-75% or below, because the SIJ may be just one aspect of a lumbo-pelvic dysfunction clinically.

These findings from investigations indicate the validity and reliability of the pain provocation tests at the SIJ is still contentious. Part of the reason for this is that some of results cannot be presumed significant because of small sample size inflating results or insufficient statistical analysis presented. Statistical analysis should include measures of reliability (kappa, percentage agreement) and measures of validity (sensitivity, specificity, positive and negative predictive values.)

The collated information previously investigated indicates there is still debate as to whether reliability and validity of pain provocation tests at the SIJ exists. Pain provocation tests are key tools in detecting sacroiliac joint dysfunction (SIJD) in daily practice. Of 186 therapists surveyed 75% use screening procedures believed to test SIJ function\textsuperscript{1}. Hence, furthering this research is important to osteopaths and all other manual therapists who rely on pain provocation as part of their diagnosis and treatment plan. If the tests these practitioners are using are not reliable or valid, manual therapists may be using an incorrect treatment due
to an incorrect diagnosis. If increased reliability and validity of these tests were to be found, it would instill confidence in therapists to rely on these tests and increase the efficacy of their treatment. Therefore the general aim of the study was to further investigate the reliability and validity of pain provocation tests at the SIJ.
Method

The Victoria University faculty of Human Development Ethics Committee granted ethics approval. Fifty participants (21 male, 29 female) with a mean age of 23 from junior year levels within the osteopathic course at Victoria University (VU) were tested. All participants gave informed consent (Appendix 3 & 4). Inclusion criteria for the study were; age 18-50: Symptomatic patients (16) had pain over the referral area from the SIJ; ipsilateral groin or buttock and posterior superior thigh\textsuperscript{15,18}, with either, 1) a current SIJ problem as diagnosed by a manual therapist or 2) positive thigh thrust or Patrick’s sign. The remaining asymptomatic participants (34) had never been diagnosed with SIJD or had any history of low back pain or pain over the SIJ referral area. Exclusion criteria included treatment undertaken on the SIJ within the past week.

The examiners were two fifth year VU osteopathic students who also gave informed consent. The selected examiners were given an explanation of the examinations (Appendix 2). Two training sessions were undertaken by both examiners under the guidance of a fellow final year osteopathic student, all procedures were demonstrated and practiced by each examiner to ensure the tests were performed accurately. Tests were performed as according to Magee\textsuperscript{18}, Greenman\textsuperscript{19} and Evans\textsuperscript{20}. The pressure application for each test was advised using terms including “build up to firm pressure” or “light steady pressure”, as indicated by orthopedic assessment texts\textsuperscript{18,19,20}. The testing examiners were equally experienced and trained in all the procedures to prevent any variability of results.
Examiners and participants were blinded from the inclusion status. The investigation was undertaken in the Victoria University Osteopathic Clinic. Examiners and participants were not permitted to discuss findings with each other. Examiners were separated. Examiner (A) tested participants followed by examiner (B). A record sheet informed the practitioner of the side that was to be tested (Appendix 1). If the participant was asymptomatic a random side was selected, similarly for a participant with indistinguishable pain from right to left, and the most painful side for symptomatic participants was tested. Participants were informed when the test was beginning and when the test had ended. They were asked by the examiner to report any occurrence of pain over the SIJ on the tested side, which indicated a positive finding. The examiner noted a negative when no pain was felt. The tests chosen had the highest validity and reliability as seen in previous studies$^{12,13}$. Tests used: Gaenslen’s test, gapping test, iliac compression test, thigh thrust and Faber (Patrick’s sign)$^{13,18,19,20}$ (Appendix 2). The data was collated on Microsoft Excel, and statistical analysis was performed with SPSS. Percentage agreement and Cohen’s kappa were calculated. Cohen’s kappa was set at minimum level $k>0.6^{21,22}$. Validity was measured by calculating sensitivity, specificity, positive predictive value and negative predictive values (Table-1)$^{8,15,23}$. Acceptable values were set at $x\geq0.80^{10}$. 
### TABLE 1 – Calculation of Validity\textsuperscript{8,16,23.}

<table>
<thead>
<tr>
<th></th>
<th>Participant with disease</th>
<th>Participant without disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Positive test</strong></td>
<td>A: True positive</td>
<td>B: False positive</td>
<td>A + B</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Negative test</strong></td>
<td>C: False negative</td>
<td>D: True negative</td>
<td>C + D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>A + C</td>
<td>B + D</td>
<td>A + B + C + D</td>
</tr>
</tbody>
</table>

**Sensitivity:** How often a test shows pathology when it is present.

**Specificity:** How often a test is normal when no pathology is present.

**Positive predictive value:** Indicates the likelihood of the patient having a dysfunction when the test is positive.

**Negative predictive value:** Indicates the likelihood of the patient not having a dysfunction when the test is negative.

These values were calculated without the use of SIJ anesthetic blocks, the gold standard in detecting SIJD. The assumption was made that the pain referral patterns from the SIJ quoted in texts are reliable\textsuperscript{9,15,18,24.}
Results

Fifty participants were involved in the study: 29 women and 21 men ranging in age from 19-42 (mean age=23). Sixteen of the participants were symptomatic for SIJD, the remaining thirty-four were asymptomatic for SIJD. Of the symptomatic group only 34% of tests were positive overall, with the thigh thrust being most valid at 56% true positive. Results from the asymptomatic group showed a 7% false positive rate overall, with the thigh thrust having the highest false positive rate at 19%. Percentage agreement between examiners was 76% overall. Chart one represents the true/false positive and percentage agreement results.

Chart 1 – True and False positives Vs Inter Examiner Agreement

For the tests combined sensitivity was found to be 0.34, specificity 0.93, positive predictive value 0.70 and negative predictive value 0.75. Individual standouts were the positive predictive value and the specificity for the Gaenslen’s test, found to be 1.0. Cohen’s kappa findings included $k = 0.475$ overall (Appendix 7). Individual test findings for predictive values and kappa are shown in Table 2.
Table 2 – Measures of validity and reliability for individual/overall tests.

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive predictive value</th>
<th>Negative predictive value</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaenslen’s</td>
<td>0.31</td>
<td>1.0</td>
<td>1.0</td>
<td>0.76</td>
<td>0.556</td>
</tr>
<tr>
<td>Gapping</td>
<td>0.22</td>
<td>0.94</td>
<td>0.64</td>
<td>0.72</td>
<td>0.291</td>
</tr>
<tr>
<td>Compression</td>
<td>0.25</td>
<td>0.94</td>
<td>0.67</td>
<td>0.73</td>
<td>0.110</td>
</tr>
<tr>
<td>Thigh thrust</td>
<td>0.56</td>
<td>0.81</td>
<td>0.58</td>
<td>0.80</td>
<td>0.674</td>
</tr>
<tr>
<td>Patrick’s sign (Faber)</td>
<td>0.38</td>
<td>0.96</td>
<td>0.8</td>
<td>0.77</td>
<td>0.611</td>
</tr>
<tr>
<td>Overall</td>
<td>0.34</td>
<td>0.93</td>
<td>0.70</td>
<td>0.75</td>
<td>0.475</td>
</tr>
</tbody>
</table>
Discussion

Overall results shown in chart 1 and table 2 indicate a moderate degree of interexaminer agreement between the two practitioners with percentage agreement = 76% and $k=0.475$, which is significant. Poor validity of the tests is demonstrated with an overall sensitivity = 0.34.

The most significant interexaminer reliability test results were the thigh thrust ($k=0.674$) and the Faber test ($k=0.611$) which showed a high level of agreement. Closely followed by the Gaenslen test $k=0.566$. These combined results demonstrated an overall moderate level of reliability of the selected pain provocation tests at the SIJ. The only exception would be the compression test which demonstrated poor agreement between the practitioners with a $k=0.110$.

Interexaminer reliability of the tests may have been supported but the validity could not be accepted at an overall sensitivity of 0.34. The positive predictive value at 0.70 and negative predictive value at 0.75 were very close to the acceptable level of 0.80, while the specificity found can be accepted at a level of 0.93. The Gaenslen, thigh thrust and Faber tests had excellent levels of specificity between 0.81-1.0. These findings indicate the tests can consistently find true negative results in an asymptomatic population, but that they were unable to accurately identify the presence of SIJD in symptomatic patients consistently. This may be have been due to the small sample size of symptomatic patients (16) and the utilisation of unsubstantiated inclusion criteria for the symptomatic subjects. Inclusion criteria that may have affected results were the use of pain referral patterns over the area from the SIJ and current SIJD
diagnosed by a manual therapist. These may have influenced results as many tissues other than the SIJ can refer over the groin, buttock and posterior superior thigh. Secondly, the diagnosis of an SIJD by a manual therapist may have involved unsubstantiated pain provocation testing to make a diagnosis, resulting in inclusion of patients who do not have a true SIJD.

Other authors’ results from similar reliability investigations showed mixed results (Table 3). The percentage agreement data presented by Laslett and Williams\textsuperscript{12}, Potter and Rothstein\textsuperscript{5} and Dreyfuss et al\textsuperscript{10} was similar to our findings for the Gaenslen, thigh thrust and Faber tests (Patrick’s sign). The Kappa data presented by Laslett and Williams\textsuperscript{12}, and Dreyfuss et al\textsuperscript{10} was also similar to our findings for the Gaenslen, thigh thrust and Faber tests (Patrick’s sign). This demonstrated a good level of inter examiner reliability. McCombe et al\textsuperscript{14} found less convincing Kappa scores for the Gapping and Compression tests. Our findings also included kappa scores that indicated that these tests were the least reliable of the cluster regime.

Dreyfuss et al\textsuperscript{10} and Broadhurst and Bond\textsuperscript{16} demonstrated mixed validity findings similar to ours (Table 4). The majority of our results showed no true agreement with the existing data. An exception included a similar pattern in positive predictive values.
Table 3 – Existing reliability data.

<table>
<thead>
<tr>
<th>Test Percentage agreement</th>
<th>Authors</th>
<th>Dreyfuss et al\textsuperscript{10}</th>
<th>Potter and Rothstein\textsuperscript{6}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distraction (Gapping)</td>
<td>Laslett and Williams\textsuperscript{12} 91%</td>
<td>-</td>
<td>94%</td>
</tr>
<tr>
<td>Compression</td>
<td>91%</td>
<td>-</td>
<td>76%</td>
</tr>
<tr>
<td>Gaenslen’s</td>
<td>86%</td>
<td>82%</td>
<td>-</td>
</tr>
<tr>
<td>Thigh Thrust</td>
<td>91%</td>
<td>82%</td>
<td>-</td>
</tr>
<tr>
<td>Patrick’s sign (Faber)</td>
<td>-</td>
<td>85%</td>
<td>-</td>
</tr>
</tbody>
</table>

| Kappa                     | Laslett and Williams\textsuperscript{12} 0.69 | Dreyfuss et al\textsuperscript{10} - | McCombe et al\textsuperscript{14} 0.36 |
| Distraction (Gapping)     | 0.73    | -                                 | 0.16                                   |
| Compression               |         |                                   |                                        |
| Gaenslen’s                | 0.72    | 0.61                              | -                                     |
| Thigh thrust              | 0.82    | 0.64                              | -                                     |
| Patrick’s sign (Faber)    | -       | 0.62                              | -                                     |

Table 4 – Existing Validity data.

<table>
<thead>
<tr>
<th>Predictive values</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive predictive value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test \ Author</td>
<td>Dreyfuss et al\textsuperscript{10}</td>
<td>Broadhurst &amp; Bond\textsuperscript{16}</td>
<td>Dreyfuss et al\textsuperscript{10}</td>
</tr>
<tr>
<td>Thigh thrust</td>
<td>0.36</td>
<td>0.8</td>
<td>0.50</td>
</tr>
<tr>
<td>Patrick’s sign (Faber)</td>
<td>0.69</td>
<td>0.77</td>
<td>0.16</td>
</tr>
<tr>
<td>Gaenslen’s</td>
<td>0.71</td>
<td>-</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Our research was limited by the inability to use the gold standard\textsuperscript{9,10,16,17} for intra-articular SIJD diagnosis, anesthetic blocks to the SIJ, to find symptomatic patients. Therefore one inclusion criterion was based on clinical diagnosis of the SIJD, which was the study's basis. It has been hypothesised that pain provocation tests stress the lumbar spine, hips and soft tissues overlying the SIJ as well as the SIJ\textsuperscript{9}, this may have been a factor influencing results. Another downfall with using pain provocation tests was a lack of accuracy due to variability of forces\textsuperscript{25}. For these reasons as well as joint mechanics and location, investigation into pain provocation at the SIJ remains a challenge.

Recommendations for further research would include the use of spinal anesthetic blocks for inclusion of symptomatic participants. It may be that instruction sessions for involved practitioners should be undertaken by a qualified practitioner experienced in using the specific tests and also a greater number of practice sessions might need to be undertaken to prevent variability in force and direction. A larger number of examiners should be utilised to increase statistical relevance. Although a greater number of examiners may increase the rate of testing-induced false positives with a high number of repeated tests.

It is important to remember that pain provocation tests are part of a broader picture in regards to SIJD diagnosis. Diagnosis is not only made from pain provocation test but also the utilisation of a history of presenting complaint, range of motion palpation, pelvic point palpation and muscle dysfunction tests.
Conclusion

This investigation has identified a poor level of validity of pain provocation tests at the SIJ, and a poor level of inter examiner agreement using compression and gapping tests. It is difficult to know whether the methodological alteration suggested will change results, as the problem appears to lie within the sensitivity of the tests themselves. A positive finding was the good level of inter examiner agreement when using the Gaenslen test, thigh thrust and Faber test. These tests also had an excellent level of specificity between 0.81-1.0. However the tests investigated show limited value within osteopathic practice when used independently.
References:


APPENDIX 1.

Examiner number __________

Participant code: ________________

Sex - M F

HILIGHTED SIDE IS SIDE TO BE TESTED:
Please mark the finding within appropriate box

<table>
<thead>
<tr>
<th>Joint</th>
<th>Left hand side</th>
<th>Right hand side</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAENSLEN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAPPING</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ILIAC COMPRESSION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>THIGH THRUST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FABER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVERALL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX 2.
The tests are adapted from Magee\textsuperscript{18}, Greenman\textsuperscript{19} and Evans\textsuperscript{20}.

Testing protocol and diagnosis for examiners of the pain provocation procedure.

**Gaenslen's test (pelvic torsion)**

Patient position: side lying position with the test leg superior. The test leg is hyperextended at the hip and the patient holds the lower leg flexed against the chest.

Examiner: Stabilizes the pelvis and extends the superior leg at the hip.

Pain in the SIJ indicates a positive test.

Gaenslen's test is sometimes done with the patient supine, but this position may limit the amount of hyperextension available.

Pressure = light steady pressure applied for the duration of the test.

**Gapping test (distraction)**

Patient positioning: supine with the affected side close to the side of the table.

Examiner: Stands on the affected side, the examiner applies cross arm pressure to the ASIS. The pressure is exerted in a posterior and lateral direction with the arms.

Discomfort over the contact point (skin and ASIS) is something to keep in mind.

A positive test is indicated by pain felt at the SIJ.

Pressure = a build up of pressure.

**Iliac compression test (Approximation)**

Patient positioning: side lying with the affected side up. Hips are flexed to approximately 45°, and knees 90°.

Examiner: stands behind the patient and places folded hands over the anterior edge of the iliac crest and applies a downward pressure.

Pain is felt in the ipsilateral SIJ in a positive result.

Pressure = a build up of pressure.

**Thigh thrust (Femoral shear test)**
Patient position: Supine with the contralateral leg extended.

Examiner: Standing at the affected side, the examiner flexes the ipsilateral leg to approximately 90° of hip flexion while the knee remains relaxed. The examiner then slightly adducts the femur and encloses the knee with folded hands. The examiner applies a graded force through the long axis of the femur, which causes anterior-posterior shear at the sacroiliac joint on the same side.

Pain is felt in the ipsilateral SIJ in a positive result.

Pressure = a build up of pressure.

**Faber (Patrick’s sign)**

Patient position: supine

Examiner: stands next to the patient on the affected side. The subject brings the ipsilateral knee into flexion with the medial side of the heel against the knee of the other leg. The examiner fixates the contralateral ASIS to ensure that the lower back remains in a neutral position. The subject then lowers the ipsilateral leg towards the table as far as possible while the foot remains in contact with the table. The examiner applies light overpressure to the patient’s knee.

Pressure = light steady pressure applied for the duration of the test.