Technical and measurement report

The reliability of selected motion- and pain provocation tests for the sacroiliac joint

Hilde Stendal Robinsona,*, Jens Ivar Broxb, Roar Robinsonc, Elisabeth Bjellandc, Stein Solemc, Trym Telje c

aSection for Health Science, University of Oslo, P.O. Box 1153, Blindern, N-0318 Oslo, Norway
bThe National Hospital, Orthopaedic Department, N-0027 Oslo, Norway
cHans & Olaf Physiotherapy Clinic, Oslo, Norway

Received 17 March 2004; received in revised form 3 June 2005; accepted 20 September 2005

Abstract

The objective of the study was to assess inter-rater reliability of one palpation and six pain provocation tests for pain of sacroiliac origin. The sacroiliac joint (SIJ) is a potential source of low back and pelvic girdle pain. Diagnosis is made primarily by physical examination using palpation and pain provocation tests. Previous studies on the reliability of such tests have reported inconclusive and conflicting results. Fifty-six women and five men aged 18–50 years old were included in the study. Fifteen patients had ankylosing spondylitis; 30 women had post partum pelvic girdle pain for more than 6 weeks; and 16 people had no low back or pelvic girdle pain. All participants were examined twice on the same day by experienced manual therapists. Percentage agreement and kappa statistic were used to evaluate the tests reliability. Results showed percentage agreement and kappa values ranged from 67% to 97% and 0.43 to 0.84 for the pain provocation tests. For the palpation test the percent agreement was 48% and the kappa value was 0.06. Clusters of pain provocation tests were found to have good percentage agreement, and kappa values ranged from 0.51 to 0.75. In conclusion this study has shown the reliability of the pain provocation tests employed were moderate to good, and for the palpation test, reliability was poor. Clusters out of three and five pain provocation tests were found to be reliable. The cluster of tests should now be validated for assessment of diagnostic power.

Keywords: Sacroiliac joint; Pelvic girdle pain; Clinical tests; Reliability

1. Introduction

The sacroiliac joint (SIJ) as a source of back pain is a recurrent subject of controversy (Walker, 1992), but several authors state that the SIJ is a potential source for pain in the lumbar spine and buttock area (Potter and Rothstein, 1985; Shaw, 1992; Schwarzer et al., 1995; Mooney, 1997).

The prevalence for SIJ dysfunction as a primary source of low back pain is reported from 0.4% (Cyriax, 1978) to 35% (Schwarzer et al., 1995) to 98% (Shaw, 1992). This disparity is partly explained by the lack of valid criteria that prevalence can be judged by van der Wurff et al. (2000a). Although the SIJ is accepted as a source of pain, there is no general agreement concerning the different diagnostic tests and their reliability and validity. The primary diagnosis of sacroiliac pain is made by clinical history and physical examination. A wide variety of SIJ tests are available for detecting dysfunction, but no test seems to be superior to another. Many of the tests are influenced by various structures in the lower back, the hip joint and the soft tissues overlying the SIJ and consequently the tests lose their precision (Maigne et al., 1996). Furthermore, assessment and interpretation of the tests are often not
standardized. However, it is necessary for test results to be both valid and reliable, since reliability alone is not sufficient to support the quality of a diagnostic test (van der Wurff et al., 2000b).

Several studies have assessed inter-examiner reliability of tests for SIJ pain and dysfunction. These tests are divided into those that assess movement or position by palpation (palpation tests) and those that stress the structure to reproduce the patient’s symptoms (pain provocation tests) (Laslett and Williams, 1994).

Previous studies have reported that reliability is low for palpation tests and from poor to excellent for pain provocation tests (Potter and Rothstein, 1985; Laslett and Williams, 1994; Strender et al., 1997).

A single test might not be sufficient for diagnosing SIJ pain and some authors have suggested the use of a cluster of tests (Haas, 1991; Cibulka and Koldehoff, 1999; Kokmeyer et al., 2002; Riddle and Freburger, 2002). Others doubt this will be better because it will be possible for two observers to disagree on single tests, yet be in agreement on the final conclusion (van der Wurff et al., 2000a; Freburger and Riddle, 2001). There must be an assumption that all the tests used in a cluster should have acceptable reliability.

Van der Wurff et al. (2000a,b) published a review article on reliability studies. They found that the majority of studies (six out of 11) reported that the reliability of SIJ tests for mobility and pain provocation was low. There was a tendency for ‘positive’ conclusions to be inversely proportional to the methodological score (van der Wurff et al., 2000a). Their recommendation for future studies was that the population should include a control group without SIJ dysfunction together with a group of patients with presumed SIJ dysfunction (van der Wurff et al., 2000b).

Because of the inconclusive and conflicting results in previous studies of tests for SIJ pain, our study aimed to assess the inter-examiner reliability for seven commonly used clinical tests. We also included tests not previously evaluated. We included only one palpation test because previous studies have reported their poor reliability. To avoid confusion and discrepancies in the testing methods, we used standardized operational definitions and the tests were performed by experienced physiotherapists who are specialized in manual therapy (MT).

The first group of participants were recruited among patients with ankylosing spondylitis (AS). They could have no ankylosis in the pelvic area and be without obvious kyphosis; the second group of patients were recruited among women with post partum pelvic girdle pain for more than 6 weeks; and the control group were healthy subjects with no low back pain, pelvic girdle pain or hip pain for the previous 3 months. The first and second groups were selected to increase the chance that the origin of their pain could be the SIJ.

The following data was collected for each participant: age, gender, diagnosis, number of childbirths (groups 2 and 3), duration of symptoms, former back pain, pain location using a pain drawing and pain intensity by using a visual analogue scale (VAS).

The physiotherapists in this study, specialized in MT, had an average of 5.8 years of work experience after completing their MT education.

2.1. Testing procedure

Every participant was examined by two physiotherapists, with a break of 1 h between examinations. One of the therapists examined all the participants; the second therapist was randomly selected from a pool of three therapists based on their availability for a given participant. The second assessor examined the participant first or second according to randomisation. Both assessors examined each participant and both performed all the included tests. The therapists were blinded for the participants’ diagnosis, their pain history and pain drawing. The two assessors were also blinded to the results of the other. In clinical practice it is important to reproduce the patients’ pain in an examination in order to determine the pain origin. In this study we have tried to take this into account, by asking the participant not only if they felt pain during testing but also if they could recognize the pain. During the examination we did not ask about pain localisation.

For the pain provocation tests, the results were classified using a specially designed examination form recording pain in three categories. They were: concordant pain (reproduction of the patients’ pain), discordant pain (pain different from patients’ pain) and no pain. The results for the mobility test were also classified in three categories: (1) movement in right SIJ greater than movement in left SIJ; (2) movement in right SIJ less than movement in left SIJ and; (3) movement in right SIJ equal to movement in left SIJ. After each examination the form was completed, placed in an envelope and sealed to avoid bias.

2.2. Tests

Tests 1–5 are passive pain provocation tests. Test 6 is a palpation test. Test 7 is an active pain provocation
test. The latter means that the provocation procedure is generated by the participants themselves:

(1) **Compression test**: The subject lies on their side and the examiner stands in front of the subject. Pressure is applied into the pelvis by the examiner leaning his chest against the uppermost iliac crest. The test is assumed to stretch the posterior sacroiliac ligaments or compress the anterior part of the SIJ (Fig. 1).

(2) **Distraction or “gapping” test**: The subject lies supine close to the side of the table. The examiner stands at the side of the table facing the subject. The examiner applies cross-armed pressure to the anterior superior iliac spines (ASIS) directed laterally. This procedure is assumed to stretch the anterior sacroiliac ligaments.

(3) **Posterior pelvic pain provocation test (P4)**: The subject lies supine on the table. Standing at the side of the table, the examiner flexes the ipsilateral leg to approximately 90° hip flexion while knee remains relaxed. The examiner applies a graded force through the long axis of femur, thereby causing a posterior shearing stress in the SIJ. Excessive adduction of the hip is avoided, because the combination of flexion and adduction normally is uncomfortable or painful to the patient.

(4) **Patrick–Faber test**: The subject lies supine on the table, and the examiner stands next to the subject. The examiner brings the subjects ipsilateral hip and knee into flexion and puts the heel against the knee of the other leg and then fixates the contralateral ASIS to ensure the lower back stays in a neutral position. The ipsilateral leg is then lowered against the table and the examiner applies a light over-pressure to the subject’s knee. It is assumed that both the anterior sacroiliac ligament and the hip joint are stressed.

(5) **Bilateral internal rotation of the hip and unilateral internal rotation of the hip** (Fig. 2): The subject lies prone on the table. In both tests the examiner internally rotates the femur at the hip joint and when reaching the end of the movement it is assumed that the SIJ is stressed.

(6) **Test of joint-play (palpation test)**: The subject lies prone. When testing the right SIJ the examiner stands by the left side of the table facing the subject. The examiner’s right hand lifts the right ilium (the hand under anterior spina iliaca superior) and the left index finger palpates the movement between the sacrum and ilium with the heel of that hand stabilising the sacrum. This test is assumed to test the passive range of motion (ROM) between the ilium and the sacrum (Magee, 1997).

(7) **Drop-test**: This is an active pain provocation test. The subject performs a standardized motion which
may provoke the pain. Standing on one foot, the
subject lifts the heel from the floor and drops down
on the heel again (Fig. 3).

Before the study started, the physiotherapists had
training sessions to make sure that the tests were
assessed and interpreted in the same manner.

2.3. Statistical analysis

Statistical analysis of the data was performed using
the SPSS statistical package version 11.0. For each SIJ
test on each side, the percentage agreement between
the two therapists and the kappa agreement coefficient \(k\) with 95% confidence interval were calculated. The
kappa coefficient effectively discounts the proportion
of agreement that is expected by chance. \(K\) ranges in
values between \(-1\) and \(+1\). Positive values signify
agreement better than chance; 0 denotes agreement no
better than chance; and negative value signify agreement
worse than chance (Altman, 1991).

We also looked at two clusters including the pain
provocation tests with the best kappa values, and
calculated percentage agreement and kappa for the
clusters for each side separately. The following five tests
were included in cluster 1: the distraction test, P4,
Patrick–Faber, bilateral internal rotation and one-sided
internal rotation. We calculated a sum score for
examination 1 and for examination 2 for each side
separately. Maximum score was 5 on each side (five
positive tests). Agreement on a case definition was based
on at least three/four positive tests out of five. Cluster 2
included 3 tests (P4, Patrick–Faber and internal rota-
tion) and case definition was two positive tests out of
three. The sum score was calculated as for cluster 1, with
a maximum score 3.

3. Results

Sixty-one people (56 women and five men) with a
mean age 31.6 years participated in the study. Fifteen
patients had AS and low back pain. They had no
ankylosis in the pelvic area and were without obvious
kyphosis; 30 women had post partum pelvic girdle pain
for more than 6 weeks; and 16 subjects (the control
group) did not report low back pain, pelvic girdle pain
or hip pain for the previous 3 months. Table 1 presents
the demographic data for the participants. The patients
in the AS group (15) and post partum group (30) had
experienced their symptoms for a median period of 82
and 11 months, respectively. All the patients had
reported pain in the pelvic and low back region on pain
drawings. All participants were examined twice, and
there were no dropouts. One participant was excluded
due to language problems.

The percentage agreement for each pain provocation
test was between 67% and 97% and the kappa
coefficient from moderate to good with \(k=0.43–0.84\).
The percentage agreement was 48% and the kappa was \(-0.06\) for the joint-play test.

When we merged discordant pain and no pain, and repeated the analysis, the percentage agreement increased to between 74% and 97% and kappa also increased for all tests to \(k = 0.48–0.88\). The exceptions were compression test right side and the drop test (Table 2).

For the case definition (agreement on at least three out of five tests) in cluster 1, it was 76–90% agreement and kappa was between 0.51 and 0.75. For cluster 2 it was 89–91% agreement and kappa between 0.69 and 0.75 (Table 3). The healthy controls were correctly identified by both clusters.

4. Discussion

The reliability of the pain provocation tests in this study is acceptable. Reliability can be influenced by three factors: the participants, the therapists and the clinical tests. The participants were recruited from patients seen by doctors, physiotherapists and midwives, and can be seen as a consecutive sample. The observed difference concerning sex and age in the groups is considered of little importance. We attempted to have only women in the AS group, but we were not able to recruit the desired number of women with AS. The overall prevalence of AS is reported from 0.1% to 1.4%. The ratio between men and women is reported 2–3:1 (Gran and Husby, 1993, 1998). According to a recent study there is no important clinical or radiological gender difference reported (Ryall and Helliwell, 1998). The control group has a matched number of male participants. The AS patients selected had no ankylosis in the pelvic area and they had no obvious kyphosis to disturb the blinding of the therapists.

While preparing the study we discussed how to interpret the participants’ responses to the tests. Did the participant recognize the pain (concordant), or was it another pain (discordant)? This is similar to what we are doing in clinical practice. We are attempting to reproduce the patients’ actual pain on examination. Still the subjective nature of pain may confound the

<table>
<thead>
<tr>
<th>Demographic data</th>
<th>Ankylosing spondylitis (n = 15)</th>
<th>Post partum pelvic pain (n = 30)</th>
<th>No back pain (n = 16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (range)</td>
<td>34.2 (25–50)</td>
<td>31.2 (25–43)</td>
<td>29.6 (24–34)</td>
</tr>
<tr>
<td>Women (%)</td>
<td>12 (80)</td>
<td>30 (100)</td>
<td>14 (87)</td>
</tr>
<tr>
<td>Parity</td>
<td>1</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Age in months, youngest child, mean (range)</td>
<td>6.3 (1.5–20)</td>
<td>5.7 (2.5–12)</td>
<td>0.8 (0.0–1.0)</td>
</tr>
<tr>
<td>Symptoms in earlier pregnancies (%)</td>
<td>6 (20)</td>
<td>1 (6.25)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Previous back pain (%)</td>
<td>14 (93)</td>
<td>15 (50)</td>
<td>6 (37.5)</td>
</tr>
<tr>
<td>Pain drawing in SIJ area (%)</td>
<td>12 (80)</td>
<td>22 (73)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Low back/pelvic area on pain drawing (%)</td>
<td>15 (100)</td>
<td>30 (100)</td>
<td>1 (6.25)</td>
</tr>
<tr>
<td>Period of pain in months, mean (range)</td>
<td>82 (6–252)</td>
<td>11 (4–24)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Mean of VAS in mm (SD)</td>
<td>47 (20.1)</td>
<td>36 (21.4)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Number of areas on pain drawing, mean (range)</td>
<td>14.6 (6–38)</td>
<td>7 (1–18)</td>
<td>1.5 (0–9)</td>
</tr>
</tbody>
</table>
interpretation of the test results (Saal, 2002). The research setting, with its restrictions, made it difficult to communicate about the patients’ pain. This may also have influenced the interpretation and thereby the reliability.

The order of the therapists on examination was randomized in the present study and therefore did not influence the results. None of the pairings of therapists did better than the others, but one pair had inferior results on two tests. They disagreed more on the Patrick–Faber test and on the compression test right side. These disagreements most likely contributed to a lower kappa on these tests compared with the others. During the training sessions prior to the study, there was no disagreement on performance and interpretation of the tests. The most obvious reason for this difference is that this couple examined fewer participants than the other pairings.

When discordant pain and no pain were merged into no pain, the percentage agreement and kappa increased markedly for every test except for the compression test right side and the drop test. For the drop test the number of patients reporting pain was small compared with the other tests. For the compression test on right side none of the participants reported discordant pain. This most likely explains the difference for the two tests.

Our findings are in agreement with other studies reporting acceptable reliability for the P4 test (Laslett and Williams, 1994; Dreyfuss et al., 1996) and we confirm the high reliability reported by Ostgaard et al. (1994). We also found acceptable reliability for the compression and distraction test. Previous studies have reported divergent results ranging from poor to acceptable for these tests. McCombe et al. (1989) reported poor reliability for both compression and distraction test, most likely because of differences in examination technique (McCombe et al., 1989). The side differences and the wide confidence intervals reported for the compression test and drop test in the present study suggest that the estimated kappa values of the best side for these tests are not good indicators of reliability. Therefore, an alternative interpretation of the variation reported is that these tests do not have acceptable reliability. Strender et al. (1997) conclude that the compression test is not reliable (Strender et al., 1997), while Laslett and Williams (1994) report good reliability for both tests, with an agreement of 88% and $k$ between 0.69 and 0.73 (Laslett and Williams, 1994).

In former studies most researchers have used a population of back patients when evaluating SIJ tests (Potter and Rothstein, 1985; Herzog et al., 1989; Dreyfuss et al., 1994; Laslett and Williams, 1994; Schwarzer et al., 1995; Maigre et al., 1996; Laslett, 1997; Broadhurst and Bond, 1998; Slipman et al., 1998; Cibulka and Koldehoff, 1999; Freburger and Riddle, 1999; Levangie, 1999; Meijne et al., 1999). Some researchers have studied the tests in pregnant or postpartum women (Ostgaard et al., 1994; Wormslev et al., 1994; Kristiansson and Svardsudd, 1996; Albert et al., 2000; Mens et al., 2001, 2002). In the present study we included two patient groups (patients with AS and women with post partum pelvic girdle pain) together with a healthy control group, to increase the chance that SIJ might be the origin for pain. This is in accordance with former recommendations for studies on SIJ tests (van der Wurff et al., 2000b).

Table 2
Inter-examiner agreement on SIJ tests when no pain and discordant pain is merged

<table>
<thead>
<tr>
<th>Test</th>
<th>% Agreement</th>
<th>Kappa</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compression test right side</td>
<td>82</td>
<td>0.48</td>
<td>0.18–0.78</td>
</tr>
<tr>
<td>Compression test left side</td>
<td>88</td>
<td>0.67</td>
<td>0.43–0.91</td>
</tr>
<tr>
<td>Distraction test</td>
<td>82</td>
<td>0.63</td>
<td>0.43–0.83</td>
</tr>
<tr>
<td>P4 right side</td>
<td>84</td>
<td>0.76</td>
<td>0.48–0.86</td>
</tr>
<tr>
<td>P4 left side</td>
<td>87</td>
<td>0.74</td>
<td>0.57–0.91</td>
</tr>
<tr>
<td>Patrick–Faber test right side</td>
<td>80</td>
<td>0.60</td>
<td>0.39–0.81</td>
</tr>
<tr>
<td>Patrick–Faber test left side</td>
<td>74</td>
<td>0.48</td>
<td>0.27–0.69</td>
</tr>
<tr>
<td>Bilateral internal rotation</td>
<td>79</td>
<td>0.56</td>
<td>0.33–0.79</td>
</tr>
<tr>
<td>Internal rotation right side</td>
<td>90</td>
<td>0.78</td>
<td>0.60–0.94</td>
</tr>
<tr>
<td>Internal rotation left side</td>
<td>89</td>
<td>0.88</td>
<td>0.75–1.01</td>
</tr>
<tr>
<td>Drop test right side</td>
<td>97</td>
<td>0.84</td>
<td>0.61–1.06</td>
</tr>
<tr>
<td>Drop test left side</td>
<td>88</td>
<td>0.47</td>
<td>0.11–0.83</td>
</tr>
</tbody>
</table>

Table 3
Cluster reliability results

<table>
<thead>
<tr>
<th>Case definition</th>
<th>n = 61</th>
<th>Kappa</th>
<th>95% CI</th>
<th>n = 45 (only patients)</th>
<th>Kappa</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1 right side</td>
<td>3 of 5</td>
<td>80</td>
<td>0.60</td>
<td>0.40–0.80</td>
<td>76</td>
<td>0.51</td>
</tr>
<tr>
<td>Cluster 1 left side</td>
<td>3 of 5</td>
<td>85</td>
<td>0.69</td>
<td>0.51–0.87</td>
<td>80</td>
<td>0.60</td>
</tr>
<tr>
<td>Cluster 1 right side</td>
<td>4 of 5</td>
<td>90</td>
<td>0.71</td>
<td>0.49–0.93</td>
<td>87</td>
<td>0.68</td>
</tr>
<tr>
<td>Cluster 1 left side</td>
<td>4 of 5</td>
<td>90</td>
<td>0.75</td>
<td>0.56–0.94</td>
<td>87</td>
<td>0.71</td>
</tr>
<tr>
<td>Cluster 2 right side</td>
<td>2 of 3</td>
<td>91</td>
<td>0.71</td>
<td>0.47–0.95</td>
<td>89</td>
<td>0.69</td>
</tr>
<tr>
<td>Cluster 2 left side</td>
<td>2 of 3</td>
<td>91</td>
<td>0.75</td>
<td>0.54–0.96</td>
<td>89</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Cluster 1: Distraction test, P4 left or right, Patrick–Faber left or right, bilateral internal rotation, internal rotation left or right. Cluster 2: P4, Patrick–Faber and internal rotation left or right, respectively.
We have chosen to present test results for the left and right side separately, in contrast to what other authors have done. We have used the statistically recommended method because the tests on the left and right sides are not independent in an individual patient (Altman, 1991).

Although some authors have suggested that motion and position palpation tests are reliable, (Herzog et al., 1989; Cibulka and Koldehoff, 1999) there is little evidence to support this view. In our study kappa was negative for the joint play test (−0.06). This is in agreement with previous studies from Potter and Rothstein (1985) and van Deursen et al. (1990) demonstrating poor reliability for palpation tests, although they did not use the same test (Potter and Rothstein, 1985; van Deursen et al., 1990).

In recent years several authors have recommended a multi-test regimen for evaluation of SIJ pain (Haas, 1991; Cibulka and Koldehoff, 1999; Kokmeyer et al., 2002; Riddle and Freburger, 2002). Cibulka and Koldehoff (1999) used only palpation tests and classified the tests as positive and negative, regardless of what side the SIJ dysfunction was (Cibulka and Koldehoff, 1999). As previously discussed it is recommended to calculate kappa for each side (Altman, 1991). Riddle and Freburger (2002) used the same four tests as Cibulka, but concluded differently. They found poor reliability for the cluster, with a slightly higher kappa on single tests (Riddle and Freburger, 2002).

Palpation is considered an important tool in MT, but it seems difficult to gain reliable inter-examiner agreement. A valid diagnostic test procedure requires reliable tests. (Herzog et al., 1989; van der Wurff et al., 2000a, b). Other factors influence validity, but the aim of the present study was to evaluate reliability.

Kokmeyer et al. (2002) used a regimen of five SIJ pain provocation tests, and the threshold for a positive selection was set at three positive tests. Using this definition, only seven out of 78 subjects had positive tests identified by both examiners. Weighted kappa was 0.70 (Kokmeyer et al., 2002). They also calculated kappa for each individual test (distraction test, compression test, Gaenslen test, Patrick test and Thigh thrust (P4) test) and reported kappa between 0.46 and 0.67, which is in agreement with our results on single tests. Kokmeyer et al. (2002) used a cluster to make a diagnosis in each patient (Kokmeyer et al., 2002). In the present study the main purpose was to examine the reliability for single tests and no diagnoses were made by the examiners. When we calculated kappa for a cluster of tests, we looked at each side separately and the agreement on case definition was based on three positive tests out of five in cluster 1. This is in accordance with what clinicians do in their daily practice when diagnostic decisions are judged from the results of several tests. Kappa was good (0.60–0.75) (Altman, 1991), which is in accordance with the results from Kokmeyer et al. (2002). The good kappa results should be linked to the case definition and the number of cases. When three positive tests out of five were required we found agreement on 18 and 20 cases for the right and left sides, respectively. This is a higher number of cases than in former studies, probably because of the selected population. We also looked at a case definition of four positive tests out of five for cluster 1, and found kappa to be good also for this definition (0.68–0.75) (Altman, 1991). The difference of clusters 1 and 2 is that the latter consists of three one-sided tests only (P4, Patrick–Faber and internal rotation) and case definition was set on two positive tests out of three in cluster 2. Kappa was also good for this cluster (0.69–0.75) (Altman, 1991). We found agreement on case definition for 8 and 10 cases on right and left sides, respectively, when using cluster 2. This is fewer than when using cluster 1. The asymptomatic individuals were correctly identified by both clusters. The main difference when using the two clusters is on the number of cases identified.

The validity of several tests included some of the tests included in this present study have been reported previously (Dreyfuss et al., 1996; Slipman et al., 1998). There is still no agreement about the gold standard for SIJ pain. Anaesthetic blocks have been used, but seem to investigate intra-articular sources of pain, thus neglecting the structures surrounding the SIJ (van der Wurff et al., 2000b). The pain provocation tests claim to load the whole SIJ complex including the surrounding structures. If this is the case such tests should be classified as provocation tests for pelvic girdle pain, but a gold standard which takes this into consideration is still not available (van der Wurff et al., 2000b; Kokmeyer et al., 2002). Our aim was to assess reliability. Results were good for a few individual tests and for a medium and a small cluster of tests, but poor for the commonly used palpation tests. Future studies should assess the validity of individual tests and especially clusters of tests. In the present study the examiners were physiotherapists experienced in MT. They were working in the same setting, attempting to optimize agreement by practicing the skills of the required procedures. Thus, less agreement is expected in an ordinary clinical setting or between various medical specialists and procedures.

5. Conclusion

Among experienced therapists, reliability was moderate to good for all the pain provocation tests and poor for the palpation test (joint play). For the compression test and the drop test the confidence intervals were wider and the side difference larger than for the other tests. In clinical practice we usually make conclusions based on the results of several tests. The clusters of three and five pain provocation tests used in this study showed good
reliability, although further studies are needed to assess their validity.

Acknowledgment

The authors would like to thank Professor Nina Vollestad at Section for Health Science, University of Oslo for valuable advice on the statistics in this study.

References


