

An annotated bibliography of spinal motion palpation reliability studies

Michael Haneline, DC, MPH
 Robert Cooperstein, MA, DC
 Morgan Young, DC
 Kristopher Birkeland, BA

Background: *Several literature reviews have addressed the reliability of spinal and sacroiliac (SI) motion palpation (MP), finding that, in general, interexaminer reliability is slight and intraexaminer reliability is moderate.*

Methods: *We performed a literature search of four biomedical databases to locate articles that dealt with MP of the spine or SI joints. The abstracts of the retrieved citations were independently screened for inclusion by two of the authors. The full-text of potentially includable articles was examined by the same two authors to assess whether they met all of the inclusion criteria. The validity of the included studies was evaluated using a 6-point scale.*

Results: *The initial searches netted 415 citations; another 30 were harvested from the secondary search. Fifty-nine articles were removed as duplicates and 305 failed to meet the inclusion criteria. Another 33 were excluded because they did not adequately describe the method of analysis, used a combination of tests, were not actually MP studies, or were not reliability studies.*

Annotated bibliography: *Summaries of 48 articles that dealt with the reliability of spinal and SI MP are presented. Where appropriate, we have commented on some of the methodological deficiencies that were discovered.*

(JCCA 2009; 53(1):40-58)

Contexte : *Plusieurs analyses documentaires ont abordé la question de la fiabilité de la palpation dynamique (MP) rachidienne et sacro-iliaque (SI) pour en conclure que, de manière générale, la fidélité inter-examineurs est légère et la fidélité intra-examineurs est modérée.*

Méthode : *Nous avons effectué un dépouillement d'ouvrages spécialisés dans quatre bases de données biomédicales pour repérer des articles portant sur la palpation dynamique (MP) rachidienne ou de l'articulation sacro-iliaque (SI). Les résumés des citations récupérées ont fait l'objet d'une présélection indépendante par deux des auteurs en vue d'une utilisation ultérieure. Le texte complet des articles pouvant être inclus ont été révisés par lesdits auteurs dans le but de vérifier s'ils répondaient aux critères d'inclusion. La validité des études retenues a été évaluée sur une échelle comptant 6 points.*

Résultats : *La recherche initiale a généré 415 citations, plus 30 additionnelles au cours d'un second dépouillement. On a retranché 59 articles, qui n'étaient que des reproductions tandis que 305 ne répondaient pas aux critères d'inclusion. Trente-trois ont été exclus faute de décrire de manière adéquate leur méthode d'analyse, d'utiliser une conjugaison de tests, de ne pas porter vraiment sur la palpation dynamique (MP) ou de n'être tout simplement pas fiables.*

Bibliographie annotée : *Des résumés de 48 articles portant sur la fiabilité de la palpation dynamique rachidienne et sacro-iliaque (SI) sont présentés. Selon les*

cas, nous avons commenté certaines lacunes en ce qui a trait à leur méthodologie.
(JACC 2009; 53(1):40–58)

KEY WORDS: motion palpation, spine, sacroiliac, reliability

MOTS CLÉS : palpation dynamique, rachidien, sacro-iliaque, fiabilité

Background

Several literature reviews have addressed the reliability of spinal and sacroiliac motion palpation (MP),^{1–4} finding that, in general, interexaminer reliability is slight and intraexaminer reliability is moderate. Two other reviews covering a variety of manual examination procedures included MP studies as well, one by Seffinger et al,⁵ who reported on the reliability of studies that involved general spinal palpation, and another by Stochkendahl et al,⁶ who reported on manual examination of the spine. In the process of conducting a literature search for another review article on the reliability of MP, which investigated differences between studies that utilized the end-feel method versus those that used the excursion method of MP, we located a number of relevant studies that were not included in these previous literature reviews. Accordingly, we present herein a more comprehensive annotated bibliography of studies that have considered the intra- and interexaminer reliability of MP.

In the article summaries, where appropriate, we have commented on some of the methodological deficiencies that were present. For instance, some of the studies only provided percent agreement or correlation statistics, instead of more accepted indices of agreement like Kappa⁷ and the intraclass correlation coefficient (ICC).⁸ The use of percent agreement alone in reliability studies may overestimate the true amount of agreement as it does not correct for agreement observed due to chance alone.⁹ Also, it is possible for examiners' findings to be highly correlated, yet at the same time be in disagreement. This phenomenon occurs when one examiner consistently scores subjects higher or lower than the other examiner.¹⁰ Other methodological concerns include the failure to randomize the order of examiners, the inclusion of only healthy subjects, and inadequate blinding of examiners and patients.

Methods

We performed a literature search of MEDLINE-PubMed, Manual Alternative and Natural Therapy System (MANTIS), the Index to Chiropractic Literature (ICL), and the Cumulative Index to Nursing and Allied Health Literature (CINAHL) databases to locate articles that dealt with MP of the spine or sacroiliac (SI) joints. The search spanned the years from 1965 through January 2007. Search terms included “motion palpation,” “spine,” and “sacroiliac.” A secondary search was conducted using the references cited in the first group of papers retrieved.

The inclusion criteria were as follows: the articles were in the English language and investigated the intra- and/or interexaminer reliability of manual MP of the spine or SI regions; the studies involved humans, were published in a refereed journal, and were published between 1965 and January 2007. Articles were excluded if they were not consistent with the inclusion criteria or if they were letters, commentaries, or editorial articles. Articles were also excluded if their methods or data presentation were unclear, or if the results of the MP evaluation were combined with other tests (e.g., pain provocation).

The abstracts of the initially retrieved citations were independently screened for inclusion by two of the authors (MH and RC) for consistency with the inclusion and exclusion criteria. The full-text of the articles that seemed includable were obtained and examined by the same two authors to more closely evaluate whether they met all of the inclusion criteria. Any disagreements about which articles should be included were resolved by consensus.

The included studies were evaluated for the presence of methodological deficiencies by two of the authors (MH and MY) using a 6-point scale developed by Stochkendahl et al.⁶ specifically to assess the quality of reproducibility studies.

The degree of examiner agreement presented in the studies was characterized using the following interpretation of kappa values: 0, None; 0–0.2, Slight; 0.2–0.4, Fair; 0.4–0.6, Moderate; 0.6–0.8, Substantial; 0.8–1.0, Almost perfect.¹¹ A kappa value of 0.4 is commonly chosen as the lower limit of acceptable reliability.¹² When agreement was reported using the intraclass correlation coefficient (ICC), the following interpretation scale was used: >0.75, good reliability; 0.40 to 0.75, fair to good reliability; <0.40, poor reliability.¹³

In writing our review, we attempted to take advantage of the strengths of both systematic and narrative reviews. Systematic reviews, because they use explicit methods to methodically search, critically appraise and synthesize the available literature on a specific issue, are easily reproducible by others. We incorporated this strength of the systematic review by being very transparent on our inclusion criteria for papers selected. We also took advantage of the narrative review format, because our goal was to provide the chiropractic clinicians and educators easy access to virtually all the published literature on the reliability of motion palpation, with the stipulation that some of this literature is marked with obvious flaws (some of which we point out). We wanted our review to be largely non-technical, and free of biostatistical jargon so that it may be easily assimilated by readers; while at the same time be free of the bias (e.g., selection bias and reporting bias) that can taint overly non-rigorous reviews. In short, we have attempted to strengthen our annotated narrative review of motion palpation studies with an infusion of systematic review methods, hoping to capitalize on the benefits of each approach. We did not provide a synthesis of the included studies, however; making a systematic review is an important next step in adequately covering this topic.

Results

The initial searches netted 415 citations; plus, another 30 were harvested from the secondary search. Fifty-nine of the 445 total articles were removed as duplicates and 305 failed on closer examination to meet the inclusion criteria. After reading the full-text versions of the remaining 81 articles, another 33 were excluded because they did not adequately describe the method of analysis, they had used a combination of tests rather than merely MP, were not actually MP studies, or were not actually reliability

studies. Forty-eight articles were ultimately selected for inclusion. Tables 1 and 2 provide summary information of the inter- and intraexaminer reliability studies respectively. Those that used both types of assessment are listed in both tables.

Annotated bibliography

1. Bergstrom E, Courtis G. An inter- and intra-examiner reliability study of motion palpation of the lumbar spine in lateral flexion in the seated position. *Eur J Chiropractic* 1986; 34: 121–41.

Two experienced chiropractors (DCs), who were pre-trained in the study procedures, examined the lumbar spines (L1-L5) of 100 asymptomatic subjects. The examiners assessed the end-feel of these joints in lateral flexion. The examiners were blindfolded during the palpation procedure and one of the study controllers placed the examiners' thumbs over each subject's L5 spinous process. The lumbar spinous processes were marked by a study controller prior to the motion palpation (MP) examination. Interexaminer reliability in this study was conveyed only by mean percent agreement, which was 81.8% for the evaluations based on both level and direction of fixations, and 74.4% for evaluations based on levels alone. Intraexaminer reliability was also reported by mean percent agreement, which was 95.4% for both examiners. The authors suggested that their findings pointed to a high degree of examiner agreement; though, they did not report kappa or ICC values, which would be necessary to correctly represent the degree of agreement that occurred beyond chance. Furthermore, only asymptomatic subjects were included in this study, which is considered to be a methodological flaw in reliability studies. This is because there should be a mixture of subjects with and without the condition or there will be an increased chance that rates of agreement will be biased.

2. Binkley J, Stratford PW, Gill C. Interrater reliability of lumbar accessory motion mobility testing. *Phys Ther* 1995; 75:786–92.

Six physical therapists (PTs) evaluated the posterior-anterior (P-A) accessory motion of the lumbar regions (L1-S1) of 18 low back pain patients. The PTs were described as average experienced clinicians who

worked in an outpatient setting. The article was not clear whether the examiners were blinded from each other's findings. A nine-point scale was used to rank the degree of mobility, ranging from 9 = "no motion" to 1 = "severe excess motion." Pain provocation found at any lumbar level was also recorded. The overall objective of the examiners was to locate treatable spinal levels, based on the findings of the motion and pain evaluation. The reliability portion of the study involved the assessment of only one spinous process, which was arbitrarily marked by one of the investigators prior to palpation. Interexaminer reliability was judged as being poor, with ICC = 0.25 (95% confidence interval [CI] 0.00–0.44) and kappa (κ) = 0.09. The authors concluded that PTs should exercise caution when making clinical decisions related to motion assessment of specific spinal levels using P-A accessory motion testing.

3. Bolin P, Keating J, Brist J, Denver G. Interexaminer reliability of palpatory evaluations of the lumbar spine. American Journal of Chiropractic Medicine 1988; 1:5–11.

Two DCs (one senior intern and one recent graduate), who were blinded from each other's findings, conducted MP examinations of the lumbar spines (T12-S1) of 50 subjects. Twenty-three subjects were symptomatic low back pain patients and 27 were asymptomatic. The examiners rehearsed the study's palpation procedures for approximately 20 hours prior to its inception. Segments were considered fixated if a "hard end-feel" was detected on MP. The interexaminer reliability marginally exceeded chance levels, with kappa values that ranged from –0.05 to 0.33 and percent agreement from 60% to 90%. There was no difference in reliability when the symptomatic subjects were analyzed separately from the asymptomatic subjects.

4. Brismée JM, Gipson D, Ivie D, et al. Interrater reliability of a passive physiological intervertebral motion test in the mid-thoracic spine. J Manipulative Physiol Ther 2006; 29:368–73.

Three PTs who were experienced, certified manual therapy instructors independently examined the mid-thoracic regions (T5-T7) of 41 asymptomatic subjects to assess passive intervertebral motion. For the exami-

nation, each subject was seated with the examiners standing by their side. Subjects were passively moved by the examiners into thoracic extension, side-bending and segmental rotation during palpation. Each examiner contacted the putative T7 spinous process using the thumb to detect relative changes in the position of the T6 spinous processes compared to that of T7. The order of examiners was randomized and they were blinded from each other's findings. Interexaminer reliability of a pre-selected thoracic segment was determined by having the examiners track the position of the spinous processes by palpation and then specify the direction of lateral flexion that produced the most segmental rotation. The authors reported fair to substantial reproducibility between examiners with percent agreement ranging from 63.4% to 82.5% and kappa values ranging from 0.27 to 0.65.

5. Carmichael JP. Inter- and intra-examiner reliability of palpation for sacroiliac joint dysfunction. J Manipulative Physiol Ther 1987; 10:164–71.

Three chiropractic students, blinded from each other's findings, examined the SI joints of 54 asymptomatic subjects using the Gillet step test. The examiners received nine sessions of training in the study protocols. The authors reported that interexaminer reliability was slight, with kappa values ranging from –0.65 to 0.19 (mean 0.023) and percent agreement from 66% to 100% (mean 85.3%). Intraexaminer reliability was reported to be fair, with pair-wise kappa values ranging from –0.02 to 0.69 (mean 0.314) and percent agreement from 75.5% to 100% (mean 89.2%). The authors concluded that the test is clinically useful for a single examiner in assessing SI joint mobility, but that its interexaminer reliability was only weakly supported.

6. Christensen HW, Vach W, Vach K, et al. Palpation of the upper thoracic spine: an observer reliability study. J Manipulative Physiol Ther 2002; 25:285–92.

Intra- and interexaminer reliability of upper thoracic region (T1-T8) was assessed by having two experienced DCs examine 56 subjects (29 patients with known or suspected stable angina pectoris and 27 control subjects without chest pain) using sitting and prone MP to evaluate biomechanical dysfunction

(end-play restriction or joint play). The examiners were blinded from each other's findings. One experienced DC examined 14 chest pain patients and 15 control subjects in the intraexaminer portion of the study. The first examiner marked the spinous process of T1 on each subject prior to palpating. Examiners were considered to be in agreement when their calls were within one spinal segment of each other. Intraexaminer reliability was found to be good ($K = 0.59$ to 0.77), while interexaminer reliability was low ($K = 0.24$ for sitting and 0.22 for prone MP). On the other hand, both intra- and interexaminer reliability was poor ($K < 0.50$) when a stricter requirement for agreement was used, wherein the examiners were required to agree on the precise level of involvement.

7. **Comeaux Z, Eland D, Chila A, Pheley A, Tate M. Measurement challenges in physical diagnosis: refining inter-rater palpation, perception and communication. Journal of Bodywork and Movement Therapies 2001; 5:245–253.**

Three osteopaths (DOs), each with more than 10 years of experience, motion palpated what they described as the lower cervical and upper thoracic regions, which ranged from C2 through T8 of 54 asymptomatic subjects. Blinding of the examiners was not apparent. The MP procedure involved rotation or side-bending of each segment to discern if there was an “inversion in direction of resistance versus ease” as compared with the adjacent segments. Subjects were seated during the examination and the initial examiner marked the T1 and T8 levels with adhesive dots. Each examiner then identified what they thought was the most restricted segment in the test zone. Interexaminer agreement in this study was considered to be poor to good ($K = 0.12$ and 0.56).

8. **Deboer KF, Harmon R, Jr., Tuttle CD, Wallace H. Reliability study of detection of somatic dysfunctions in the cervical spine. J Manipulative Physiol Ther 1985; 8:9–16.**

Three experienced DCs, blinded from each other's findings, palpated the cervical regions (C1-C7) of 40 asymptomatic subjects to assess the test's interexaminer and intraexaminer reliability. Subjects were seated in front of the examiners, who assessed the

segments for the presence of fixations on side bending, flexion, extension, and rotation. The authors reported that their results were variable, but that moderate degrees of consistency were apparent among examiners. They also noted that findings were dissimilar between the upper, middle and lower cervical regions, with virtually no agreement above chance in the middle cervical area (mean $K_w = 0.03$ for inter- and $K_w = 0.07$ for intraexaminer reliability) and moderate agreement in the lower cervical spine (mean $K_w = 0.42$ for inter- and $K_w = 0.40$ for intraexaminer reliability). In the upper cervical region, the mean weighted kappa was 0.10 for inter- and 0.48 for intraexaminer agreement.

9. **Degenhardt BF, Snider KT, Snider EJ, Johnson JC. Interobserver reliability of osteopathic palpatory diagnostic tests of the lumbar spine: improvements from consensus training. J Am Osteopath Assoc 2005; 105:465–73.**

This study involved an initial interexaminer reliability trial in which three DOs with specialized training in neuromusculoskeletal medicine assessed intersegmental motion in the lumbar spine (L1-L4) in 42 asymptomatic subjects in a blinded fashion. The palpation procedure involved P-A springing of the lumbar spinous processes to assess the quality of motion. The second phase of this study involved a four month training period on the palpation procedures in an effort to promote consensus between the examiners. After the training period, the examiners reassessed the interexaminer reliability of lumbar spine motion in another group of 15 to 33 asymptomatic subjects. Kappa scores for the procedure improved from $K = 0.10$ before training to $K = 0.20$ ($P = 0.04$) after training. The authors concluded that consensus training improved the reproducibility of the test, although reliability was still not considered to be acceptable.

10. **Downey B, Taylor N, Niere K. Can Manipulative Physiotherapists Agree on which Lumbar Level to Treat Based on Palpation? Physiotherapy 2003; 89:74–81.**

Six PTs with advanced training in manipulative physiotherapy and at least 7 years of clinical experience examined the lumbar regions (L1-L5) of 20 low back

pain patients. Examiners used P-A pressure on the vertebrae in an effort to determine the spinal level thought to be causing the patient's symptoms. The PTs were blinded from each other's findings, but were permitted to communicate with the patient about the presence of pain during palpation. The findings of the study pointed to fair agreement ($K = 0.37$) between examiners for palpation, although there was poor agreement about naming which vertebral segment was involved ($K = 0.09$). The authors therefore concluded that the palpation technique used in this study may not be reliable and clinicians should not base treatment decisions on these findings alone.

11. Fjellner A, Bexander C, Faleij R, Strender LE. Interexaminer reliability in physical examination of the cervical spine. J Manipulative Physiol Ther 1999; 22:511–6.

The cervical and upper thoracic regions (C0-T5) of 48 subjects were examined by two PTs who had at least six years of clinical experience and were blinded from each other's findings. Eleven of the subjects reported neck symptoms during the study, even though they did not consider themselves to be neck pain patients, and 35 of the subjects were asymptomatic. One subject was excluded because of increased pain experienced during the examination. Intersegmental joint excursion was tested throughout the cervical and upper thoracic regions, while joint end-feel was only tested with the two upper cervical vertebrae and the occiput. Subjects were seated with the examiner standing by their side. Reported reliability coefficients for the joint excursion tests ranged from $K_w = -0.016$ to 0.49 and percent agreement from 41% to 92%. The upper cervical end-feel tests resulted in weighted kappa values that ranged from 0.01 to 0.18 and percent agreement from 60% to 87%. The authors concluded that the tests of passive intersegmental movement evaluated in this study had poor reliability. However, when comparing the results of tests performed on subjects with and without symptoms, more tests were reported to have acceptable kappa values among the symptomatic subjects as compared with the asymptomatic subjects. Accordingly, further research on this issue was recommended.

12. Flynn T, Fritz J, Whitman J, et al. A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. Spine 2002; 27:2835–43.

This was a prospective cohort study involving low back pain patients, which included a component that evaluated the interexaminer reliability of SI motion palpation. In order to determine the test's interexaminer reliability, 55 low back pain patients were evaluated by two examiners who were blinded from each other's results. Although several tests of SI function were evaluated, the Gillet step test (which assesses the joint's excursion) was the only motion test to demonstrate moderate agreement ($K = 0.59$). The standing and sitting flexion tests, which are also designed to detect SI excursion, resulted in poor and fair reliability coefficients respectively ($K = -0.08$ and 0.25).

13. Gonella C, Paris SV, Kutner M. Reliability in evaluating passive intervertebral motion. Phys Ther 1982; 62:436–44.

Five PTs with at least 3 years of experience examined the lumbar regions (T12-S1) of 5 asymptomatic subjects to evaluate the intra- and interexaminer reliability of passive intervertebral motion. Examinations were performed under "blinded" and "normal" conditions. The blinding involved having the examiners wear blackened scuba masks so they would not be able to recall the results of each subjects' first evaluation, which could lead to mistakenly high intraexaminer reliability. However, it was not clear whether the examiners were blinded from each other's findings. The lumbar segments from T11 to S1 were assessed with subjects lying on their sides with their hips and knees flexed. The examiners stood in front of the subjects, reaching across the table to perform the palpation. The authors argued that their results pointed to reasonably good or acceptable intraexaminer reliability, while interexaminer reliability was unacceptable. These conclusions are unconvincing, however, since they were based on visual inspection and comparison of the raw data rather than being based on statistical analyses that are typically utilized to represent agreement between examiners.

14. Haas M, Raphael R, Panzer D, Peterson D. Reliability of manual end-play palpation of the thoracic spine. *Chiropr Tech* 1995; 7:120–4.

Seventy-three volunteer subjects, 66% of whom were symptomatic with pain and/or stiffness (49% with pain and 56% with stiffness), had their thoracic spines examined by 2 blinded DCs with at least 15 years of teaching and clinical experience. The T3 through L1 spinous processes were marked prior to the examination. Subjects were seated in front of the examiner, who assessed the rotational end-play of each segment from T3–4 to T12-L1 by contacting the spinous processes. Reported kappa values for interexaminer reliability ranged from 0.08 to 0.22 (overall 0.14), which the authors suggested represented poor test reproducibility. Overall intraexaminer reliability was $K = 0.50$ ($K = 0.55$ for one of the examiners and $K = 0.43$ for the other), indicating moderate concordance.

15. Hanten WP, Olson SL, Ludwig G. Reliability of Manual Mobility Testing of the Upper Cervical Spine in Subjects with Cervicogenic Headache. *J Man Manip Ther* 2002; 10:76–82.

Two experienced PTs examined 2 groups of 20 subjects who were diagnosed with cervicogenic headache. The examiners independently evaluated one of the groups in a blinded fashion to determine the degree of interexaminer reliability. The other group of 20 subjects was evaluated by one of the examiners on 2 consecutive days to assess intraexaminer reliability. A total of 15 mobility tests were performed in random order on each subject. The mobility portion of the examination consisted of P-A and lateral pressures applied to the prone patients at various points of the vertebrae from C1 to C3. Reported kappa values for interexaminer reliability were quite variable, ranging from -0.71 to 0.86 and percent agreement from 70% to 95%. Intraexaminer agreement fared better, with kappa values ranging from $K = 0.21$ to 0.80 and percent agreement from 60% to 90%. The authors concluded that mobility testing of the cervical spine was a reliable tool to use in the identification of subjects with symptomatic cervicogenic headaches.

16. Herzog W, Read LJ, Conway PJ, Shaw LD, McEwen MC. Reliability of motion palpation procedures to detect sacroiliac joint fixations. *J Manipulative Physiol Ther* 1989; 12:86–92.

Ten DCs with 1 to 10 years of experience, who were blinded from each other's findings, examined 10 symptomatic low back pain patients diagnosed with SI joint dysfunction and one asymptomatic control subject using the Gillet step test. Fixations were graded as mild, moderate or complete. Intraexaminer reliability scores were reported as percent agreement, 68% for positive findings, 79% for negative findings, and 72% for agreeing about the side of involvement. Chi-square testing showed that examiners' first session scores were statistically significantly different from their second session scores. Interexaminer reliability percent agreement scores ranged from 54% to 78% for the detection of positive findings and 60% to 64% for the detection of the correct side. The drawbacks of using percent agreement alone in reliability studies have previously been discussed. In contrast with the findings of other studies, more experienced examiners had lower intraexaminer agreement scores than those with less experience.

17. Hicks GE, Fritz JM, Delitto A, Mishock J. Interrater reliability of clinical examination measures for identification of lumbar segmental instability. *Arch Phys Med Rehabil* 2003; 84:1858–64.

Four PTs with four to eight years of experience examined the lumbar regions (L1-L5) of 63 consecutive low back pain patients to evaluate the interexaminer reliability of several clinical tests that are commonly used to identify lumbar segmental instability. Examinations were done in a single blinded fashion, in which the examiners were not aware of the other examiners' findings. In the MP module, the examiner applied P-A pressure over the spinous process of the tested segment to assess the quality of motion (i.e., hypermobility, normal mobility or hypomobility). The weighted kappa (K_w) for the tested segments ranged from -0.02 to 0.26 and percent agreement ranged from 52% to 69%. The authors concluded that segmental mobility testing was not reliable in the lumbar spine.

- 18. Humphreys BK, Delahaye M, Peterson CK. An investigation into the validity of cervical spine motion palpation using subjects with congenital block vertebrae as a 'gold standard'. BMC Musculoskelet Disord 2004; 5:19.**

While the primary objective of this study was to assess the utility of the congenital block vertebra as a "gold standard" to assess the validity of motion palpation, it also included an interexaminer reliability component. The examiners were 20 fourth year chiropractic students, blinded from each other's findings, who motion palpated the cervical spines of three seated subjects with single level congenital block vertebrae. The examiners were unaware of the presence of the congenital block vertebrae, and were simply asked to identify the most hypomobile segment. The skin overlying the cervical spinal levels was marked prior to examination to facilitate the identification of the correct level of involvement. Reliability coefficients pointed to substantial agreement for the examiners to recognize the most hypomobile segment ($K = 0.65$). There was also substantial agreement ($K = 0.76$) regarding the identification of the fused C2-C3 segments and moderate ($K = 0.46$) agreement regarding C5-C6.

- 19. Inscoe E, Witt P, Gross M, Mitchell R. Reliability in Evaluating Passive Intervertebral Motion of the Lumbar Spine. Journal of manual & manipulative therapy 1995; 3:135-43.**

Two PTs with at least 4 years of experience and who were blinded from each other's findings examined the passive intervertebral motion of the lumbar regions (T12-S1) of six low back pain patients. The spinous process of the L5 vertebra was marked on each subject prior to palpation. Only forward bending of segments was assessed, which was graded as being normal, hypomobile or hypermobile. Subjects were side-lying with the examiners reaching across the table during the examination. The data analysis generated percent agreement, as well as Scott's pi, which, similar to the kappa statistic, provides information about agreement that occurs above chance levels. Intraexaminer reliability was determined to be acceptable, with percent agreement 66.7% and 75.0% for examiners A and B respectively. Above

chance agreement was 41.9% and 61.3%. For interexaminer reliability, percent agreement was 48.6%, which was only 18.4% above chance. Consequently, there was only weak support for interexaminer reliability in this study.

- 20. Jull G, Bullock M. A motion profile of the lumbar spine in an aging population assessed by manual examination. Physiotherapy Practice 1987; 3:70-81.**

Intra- and interexaminer reliability of lumbar (T12-S1) MP was evaluated as part of a larger study that investigated joint motion associated with ageing. The intraexaminer portion of the study involved one experienced PT, trained in manipulative therapy, who evaluated 20 asymptomatic subjects. Motion was graded on a five-point scale that ranged from 1 = slightly hypermobile to 5 = markedly hypomobile. Pearson's correlation was calculated between the findings of the initial and second examinations, with r values ranging from 0.81 to 0.98, which was considered to represent a good to high degree of reliability. The interexaminer reliability portion of the study included another experienced PT and 10 asymptomatic subjects. Percent agreement was calculated in addition to Pearson's correlation, which resulted in complete agreement 86% of the time and disagreement no greater than one grade on the five-point motion rating scale that was used (where 1 = slightly hypermobile and 5 = markedly hypomobile) 14% of the time. Correlation ranged from $r = 0.82$ to 0.94. The authors thought these findings corresponded to a good to high degree of consistency between examiners.

- 21. Keating JC, Jr., Bergmann TF, Jacobs GE, Finer BA, Larson K. Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality. J Manipulative Physiol Ther 1990; 13:463-70.**

Three licensed DCs, who initially practiced their palpation methods on 18 volunteers, examined 21 symptomatic low back pain patients and 25 asymptomatic subjects to determine the interexaminer reliability of eight segmental evaluative procedures. The examiners were blinded from each other's findings and their order of conducting the tests was rand-

omized. Lumbar spinal segments were examined from the level of T11-T12 through L5-S1. Patients were seated with the examiners behind them for the MP tests. Both active and passive MP was evaluated, which showed little significant agreement between examiners. Reported kappa values for MP ranged from -0.18 to 0.31.

22. Leboeuf C. Chiropractic examination procedures: a reliability and consistency study. J Aust Chiropr Assoc 1989; 19:101-4.

A group of 45 chronic LBP patients were examined by four fifth-year DC students to assess the inter- and intraexaminer agreement of six chiropractic tests. It was not apparent whether the examiners were blinded from each other's findings. The only test that resulted in significantly better intraexaminer agreement when compared with interexaminer agreement values was MP. The authors noted that there was a high rate of agreement per segment, but most of the consensus involved negative findings. No actual statistics were provided in the article, only graphs that compared the findings of the various tests. Only percent agreement was calculated and the authors suggested that tests resulting in more than 70% agreement were "clinically satisfactory." From the graphs, it appears that both intra- and interexaminer for MP was in excess of 90% agreement. The limitations of using percent agreement in examiner reliability studies were previously mentioned in the Bergstrom and Courtis summary.

23. Lindsay D, Meeuwisse W, Mooney M, Summersides J. Interrater Reliability of Manual Therapy Assessment Techniques. Phys Ther Can 1995; 47:173-80.

Two PTs, with at least six years of experience, assessed the interexaminer reliability of 37 tests commonly used to evaluate lumbosacral dysfunction, including passive accessory glide of the SI joints and P-A glide of the lumbar segments (L1-S1). Subjects were eight members of a cross-country ski team who exhibited a variety of conditions ranging from normal to mildly pathological. For the SI tests, the subjects were in a supine position with the standing examiners applying various pressures to the pelvis while assessing motion in the SI joints. The examiners

were blinded from each other's findings and the order of assessment was randomized. The SI tests were generally more reliable than those for the lumbar spine, but MP in this region was judged to be generally unreliable, with reported kappa values ranging from 0.0 to 0.6 and percent agreement from 50% to 100%. P-A glide in the lumbar region was particularly unreliable, with kappa values ranging from -0.3 to 0.0 and percent agreement from 14% to 50%.

24. Love RM, Brodeur RR. Inter- and intra-examiner reliability of motion palpation for the thoracolumbar spine. J Manipulative Physiol Ther 1987; 10:1-4.

To assess inter- and intraexaminer reliability of thoracolumbar MP, eight senior chiropractic students, who were blinded from each other's findings, examined 32 asymptomatic subjects. Each of the examiners had at least 1 year of experience with the spinal scanning procedure that was used. Subjects were seated in front of the examiner, who scanned for the presence of fixation from T1 to L5 and called only one segment as being the most hypomobile. The degree of reliability was represented by Pearson's correlation coefficient. The results of the intraexaminer reliability portion of the study were variable, with r values ranging from -0.07 to 0.65 and averaging 0.32. The results of the interexaminer reliability tests were also quite variable, with r values ranging from -0.57 to 0.49 and averaging 0.02 for trial 1 and 0.09 for trial 2. The authors suggested that intraexaminer reliability in this study was "statistically significant," while interexaminer reliability was not. However, due to the previously mentioned limitations associated with assessing examiner concordance using Pearson's correlation, as well as the low correlation coefficients, intraexaminer reliability is not supported by the study's results.

25. Lundberg G, Gerdle B. The relationships between spinal sagittal configuration, joint mobility, general low back mobility and segmental mobility in female homecare personnel. Scand J Rehabil Med 1999; 31:197-206.

The interexaminer reliability of passive lumbar (T10-

S1) segmental mobility testing was estimated among three experienced PTs. They rated the degree of segmental mobility on a five-point scale that ranged from extreme hypermobility to extreme hypomobility. The examiners were not aware of each others' findings. There were 150 asymptomatic subjects included in the study who were randomly selected from a larger sample consisting of 607 female home-care personnel. Flexion, extension, rotation, and translatory joint play of the interspaces from T10/11 to L5/S1 were assessed with subjects lying on their sides with their hips and knees flexed. The examiners stood in front of the subjects and reached across to contact the lumbar region to perform the palpation. Interexaminer agreement was reported to be good ($K_w = 0.59$ to 0.75), especially in the lowest lumbar segments.

26. Maher C, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. Phys Ther 1994; 74:801-9; discussion 809-11.

This study assessed the interexaminer reliability of P-A central pressure testing in the lumbar spine (L1-L5). Ninety low back pain patients were evaluated by six paired PTs within three separate therapy practices. Each of the PTs had at least five years of experience and was trained in manipulative physiotherapy. The first examiner marked the skin over the targeted spinous processes so the second examiner could rate the same vertebra. Otherwise, they were blinded from each other's finding. The PTs used their own preferred technique to examine P-A central pressure. Stiffness was graded on a scale that ranged from $-5 =$ "markedly decreased stiffness" to $5 =$ "markedly increased stiffness" and $0 =$ "normal stiffness." Resulting intraclass correlation coefficient (ICC) values ranged from 0.03 to 0.37 and percent agreement from 21% to 29% , which the authors thought represented poor reliability.

27. Maher C, Latimer J, Adams R. An investigation of the reliability and validity of posteroanterior spinal stiffness judgments made using a reference-based protocol. Phys Ther 1998; 78:829-837.

Five PTs with at least five years of experience exam-

ined the P-A stiffness of the third lumbar vertebra in 40 asymptomatic subjects. Prior to examination, the first rater marked the subject's lumbar spinous process to facilitate the PTs identifying the intended segment. The examiners were blinded from each other's findings and they used their own preferred method of palpation to assess P-A stiffness. A grading scale that ranged from $-5 =$ "markedly decreased stiffness" to $5 =$ "markedly increased stiffness" (with a midpoint $0 =$ "normal stiffness") was used to rate the degree of stiffness that was present. The results of the reliability portion of the study pointed to generally high reliability, with ICC values ranging from 0.50 to 0.77 and standard error of measurement (SEM) ranging from 1.58 to as low as 0.72 points. However, the authors warned against generalizing their results because of an incorporated validity study in which the reliability of the P-A stiffness tests did not correlate very well with a mechanical device criterion.

28. Marcotte J, Normand MC, Black P. The kinematics of motion palpation and its effect on the reliability for cervical spine rotation. J Manipulative Physiol Ther 2002; 25:E7.

The primary objective of this study was to evaluate the extent to which controlling the examiners' body kinematics during cervical spine rotational MP would have an impact on the reliability of the test. Thus, the reliability of MP when kinematics were carefully controlled and reproducible could be compared with its reliability when kinematics were not controlled very well. In order for the examiners' kinematics to have been considered successfully controlled, the forward inclination of the head and neck (i.e., flexion) of the supine subjects had to be consistently greater than 6 degrees. Nine subjects were selected on the basis that they had histories of cervical spine mechanical disorders. All the subjects were included in a training phase, but only three were retained for the experiment. The examiners were 24 chiropractic students, who initially received 12 hours of supervised training in the study protocols, and an experienced DC. Each examiner palpated from C0 to C7 on one of the three subjects during eight sessions. A margin of error of 2 contiguous vertebrae was ac-

cepted as being in agreement (e.g., identifying C3 as fixated would correlate with identification of C2 through C4). The authors concluded that when kinematics were successfully controlled, the reliability of cervical MP improved from kappa values of 0.34 and 0.35 to 0.68. It is not entirely clear how, in the absence of the measuring device used in this study, a clinician could maintain the kinematics recommended under typical chiropractic field conditions.

29. Marcotte J, Normand MC, Black P. Measurement of the pressure applied during motion palpation and reliability for cervical spine rotation. J Manipulative Physiol Ther 2005; 28:591–6.

The authors learned from a previous study in 2002 that the interexaminer reliability of cervical MP improved when examiners' kinematics were carefully controlled. As a follow-up, one of the primary purposes of the current study was to determine if the pressure applied during cervical spine rotational MP affected interexaminer reliability findings. The examiners were 23 chiropractic students and one experienced DC who received 12 hours of training on the study procedures prior to the experiment and were blinded from each other's findings. Examiners were fitted with flexible pressure sensors over their palpating finger and subjects were only palpated from the left side. Three asymptomatic subjects were examined by only one-third of the examiners. The amount of pressure used by the examiners did not have an effect on reliability, but the reliability was found to be high ($K = 0.7$ to 0.75) when the kinematics were carefully controlled.

30. McPartland JM, Goodridge JP. Counterstrain and traditional osteopathic examination of the cervical spine compared. Journal of Bodywork and Movement Therapies 1997; 1:173–178.

Two DOs with at least 10 years of experience, blinded from each other's findings, examined the upper cervical region of 18 subjects (7 symptomatic neck pain patients and 11 asymptomatic). The examiners evaluated lateral translation of CO on C1 and C2 on C3, as well as rotation of C1 on C2 with the neck in maximum flexion. The position of subjects and examiners during the palpatory procedure was not

clear, but the "... examiners palpated the left side of the subject's neck with their left hand, and the right side with their right hand." Four other tests of cervical spine function were evaluated at the same time. Motion restriction was defined as "abnormal quality of resistance" and "abnormal end-feel" and was graded on a 0–10 scale, with 0 = "no abnormality" and 10 = "maximum dysfunction." Interexaminer agreement was reported to be only fair in this study ($K = 0.34$ and percent agreement 66.7%).

31. Meijne W, van Neerbos K, Aufdemkampe G, van der Wurff P. Intraexaminer and interexaminer reliability of the Gillet test. J Manipulative Physiol Ther 1999; 22:4–9.

This study evaluated the intraexaminer and interexaminer reliability of the Gillet step test, which is a test used to assess SI joint motion. Two PT students in their last year of training, blinded from each other's findings, examined 38 male subjects, of which nine were symptomatic and 29 were asymptomatic. Only reasonably slender subjects were selected because of difficulties associated with palpating the required contact points of heavier individuals. The examination procedure was repeated after a minimum of four days to assess intraexaminer reliability. The results of the study showed kappa values ranging from -0.30 to 0.75 for interexaminer reliability and percent agreement 48% to 100%. The intraexaminer results ranged from $K = -0.39$ to 0.65 and percent agreement 44% to 100%. However, mean kappa values were low. The authors therefore concluded that the Gillet step test does not appear to be reliable.

32. Mior S, King R, McGregor M, Bernard M. Intra and interexaminer reliability of motion palpation in the cervical spine. J Can Chiropr Assoc 1985; 29:195–9.

Two chiropractic student examiners, who were in their final year of clinical training, received three months of specialized instruction on the study palpation procedures prior to its inception. They subsequently examined the joint play in the upper cervical regions (C0–C2) of 59 asymptomatic subjects. Examiners were blinded from each other's findings, as well as any confounding information. Subjects were

supine as the examiners palpated using their finger tips. The resulting kappa value for interexaminer reliability was 0.15 and percent agreement was 61%. For intraexaminer reliability, kappa was 0.37 for examiner one and 0.52 for examiner two and percent agreement was 71% and 79% respectively. The authors also calculated the percentage of agreement that occurred beyond chance, which was only 15% for inter- and 37% and 52% for intraexaminer reliability.

33. Mior SA, McGregor M, Schut B. The role of experience in clinical accuracy. J Manipulative Physiol Ther 1990; 13:68–71.

This study assessed the interexaminer reliability of 74 chiropractic student interns palpating the SI joints of approximately 20 subjects, some with SI fusion. The results of the students' palpation skills were compared in order to contrast before and after 1 year of clinical experience. The study also compared the intra- and interexaminer reliability of experienced clinicians with more than five-years of clinical experience. It was not apparent that the order of examiners was randomized or that they were blinded from each other's findings. Reported kappa values for the students' interexaminer reliability ranged from 0.00 to 0.30, while the range was from 0.00 to 0.17 for the experienced DCs. Intraexaminer reliability for the DCs ranged from 0.15 to 1.00, which was interpreted as being moderate to almost perfect. There were no significant differences between the students' initial scores and those obtained after 1 year of clinical experience. Accordingly, the authors suggested that these findings bring into question the role of experience in improving the reliability of MP in SI region.

34. Mootz RD, Keating JC, Jr., Kontz HP, Milus TB, Jacobs GE. Intra- and interobserver reliability of passive motion palpation of the lumbar spine. J Manipulative Physiol Ther 1989; 12:440–5.

Two DCs with at least seven years of experience examined the lumbar regions (L1-S1) of 60 asymptomatic student volunteers to assess the intra- and interexaminer reliability of passive MP for the presence of "hard end-feel." Subjects were seated during the examination, with examiners seated behind them.

Examiners contacted the targeted segments with one hand while moving the subject's torso with the other. The order of examiners was randomized, but it was not apparent whether they were blinded from each other's findings. The authors noted that intraexaminer agreement beyond chance was moderate at the L1-L2 level ($K = 0.39$) and minimal at L4-L5 ($K = 0.21$), but there were no significant agreements in the mid-lumbar region. Interexaminer agreement was judged poor at all assessed levels, with kappa values ranging from -0.17 to 0.17 . The spinal segments were grouped regionally and then re-analyzed, which netted slight improvements in intraexaminer agreement, but not interexaminer agreement.

35. Nansel DD, Peneff AL, Jansen RD, Cooperstein R. Interexaminer concordance in detecting joint-play asymmetries in the cervical spines of otherwise asymptomatic subjects. J Manipulative Physiol Ther 1989; 12:428–33.

Two pairs of chiropractic practitioners (consisting of two college faculty, one clinic staff doctor, and one student intern) examined 270 asymptomatic subjects for the presence of increased "joint play resistance" at end-range of cervical lateral flexion. Only the middle and lower cervical regions were examined and subjects were screened for the presence of joint-play asymmetry on lateral flexion in order to be included in the study. The authors emphasized that "positive palpatory findings" were comprised of a difference between joint-play at end-range comparing one side with the other, rather than the end-range restrictions of one segment as compared with the other segments. The first pair of examiners palpated subjects in a seated position, whereas the second pair examined them while supine. All examiners were blinded from each other's findings. The first examiner, after finding a "positive" segment, indicated the side of greatest fixation and the degree of asymmetry. An observer placed their fingers over the fingers of the first palpator, which were subsequently taken away, to mark the contact for the second palpator. The second palpator then examined only the specified segment and indicated the side of greatest fixation. The reported kappa value of pooled data was 0.01, which was judged to be extremely low.

36. Olson KA, Paris SV, Spohr C, Gorniak G. Radiographic Assessment and Reliability Study of the Craniovertebral Sidebending. J Manual Manipulative Ther 1998; 6:87-96.

Six PTs, with at least four and one-half years of experience, examined the upper cervical regions (C0-C2) of 10 asymptomatic subjects. The study's primary purpose was to assess the effect of patient positioning on the reliability of craniovertebral passive motion testing and end-feel assessment. The intra- and interexaminer reliability of MP in the upper cervical region was also evaluated. The order of examiners was randomized, but it was not clear whether they were blinded from each other's findings. Subjects were in a supine position with the examiner palpating from the head of the table. Both end-feel and passive motion of each segment was tested. Intraexaminer reliability for passive motion assessment resulted in kappa scores that ranged from -0.02 to 0.14, while interexaminer reliability scores ranged from -0.03 to 0.18. Kappa scores for end-feel assessment ranged from 0.01 to 0.31 for intraexaminer reliability and from -0.04 to 0.12 for interexaminer reliability. The highest intraexaminer reliability was obtained with subjects in the physiological neutral position. The authors suggested that all of the test positions showed poor interexaminer reliability.

37. Paydar D, Thiel H, Gemmell H. Intra- and Interexaminer Reliability of Certain Pelvic Palpatory Procedures and the Sitting Flexion Test for Sacroiliac Joint Mobility and Dysfunction. J Neuromusculoskel Sys 1994; 2:65-9.

Thirty-two asymptomatic subjects were evaluated for the presence of SI joint restriction by two chiropractic student intern examiners who had at least one year of clinical experience using the procedure. Examiners were blinded from each other's findings and the order of subject allocation was randomized. The SI test involved the examiners placing their thumbs bilaterally over the posterior superior iliac spines and having the seated subject bend forward at the waist. The side of greatest SI motion restriction was signified by the side that moved more superiorward. The resulting kappa scores were 0.09 for interexaminer reliability and 0.29 for intraexaminer reliability. Per-

cent agreement scores were 34.4% and 54.1% respectively. The authors suggested that their results pointed to poor to fair inter- and intraexaminer concordance.

38. Phillips DR, Twomey LT. A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. Man Ther 1996; 1:82-7.

This study was designed to determine the accuracy of manual examination of the lumbar spine (L1-L5) alone and when accompanied by a verbal subject response. The diagnosis of the symptomatic lumbar segment was compared to a segmental diagnosis that was obtained by spinal anesthetic blocks. For the intra- and interexaminer reliability portion of the study, passive physiological intervertebral movements (joint motion) and passive accessory intervertebral movements (end-feel) were tested. Two PTs, with unknown experience levels, examined the lumbar regions of 72 subjects (63 were low back pain patients and 9 were asymptomatic). The examiners were blinded from each other's findings. For the tests of joint motion, percent agreement ranged from 55% to 99% and weighted kappa values ranged from -0.11 to 0.32. For the end feel tests, percent agreement ranged from 74% to 100% and weighted kappa values ranged from -0.16 to 0.28, -0.15 to 0.24.

39. Potter L, McCarthy C, Oldham J. Intraexaminer reliability of identifying a dysfunctional segment in the thoracic and lumbar spine. J Manipulative Physiol Ther 2006; 29:203-7.

The SI regions of seventeen patients with a chief complaint of unilateral buttock pain were examined by eight PTs who had at least two years of clinical experience and specialized in orthopedic physical therapy. The examiners were blinded from each other's findings and subjects were randomly assigned to each examiner. A total of 13 SI joint tests were evaluated in this study, including three that assessed mobility, the Gillet step test, as well as the standing and sitting flexion tests. Overall percent agreement for the joint mobility tests was 46.7%, 43.8%, and 50% respectively, which was considered by the authors to be poor. The only tests that exceeded the established

70% threshold for reliability required subjective responses from the patients and thus were not truly MP tests as we have defined them.

40. Rhudy T, Sandefur M, Burk J. Interexaminer/in-tertechnique reliability in spinal subluxation assessment: a multifactorial approach. American Journal of Chiropractic Medicine 1988; 1: 111–4.

Using their preferred examination method and subject placement, three experienced DCs evaluated the full spines (C1-L5) of 17 patients who presented for chiropractic care with a variety of conditions. Several examination parameters were evaluated, including MP. However, the MP procedure was not described. Examiners were in separate rooms during the evaluations, but it was not clear if they were prohibited from discussing their findings with each other. Kappa values were calculated in this study, although the calculated values were not provided. Instead, only the interpretations of the kappa scores were provided (i.e., poor, fair, etc.). The findings of the kappa analysis ranged from “none” to “substantial” agreement. The T7-T9 region showed the best agreement, while the C3-C4 and T10-L2 areas represented the worst.

41. Robinson HS, Brox JI, Robinson R, Bjelland E, Solem S, Telje T. The reliability of selected motion- and pain provocation tests for the sacroiliac joint. Man Ther 2007; 12:72–9.

Sixty-one subjects (15 with ankylosing spondylitis, 30 with post partum pelvic girdle pain, and 16 asymptomatic) were examined by two PTs, twice in one day. The PTs were manual therapy specialists who had an average of 5.8 years of experience. Examiners were blinded from each other’s findings and the order of examiners was randomized. The subjects’ SI joints were tested using one MP test and six pain provocation tests. In the MP test, the examiner lifted the ilium of the prone patient using the anterior superior iliac spine and palpated movement of the tested SI joint with the index finger of the opposite hand. The kappa value was -0.06 and percent agreement was 48% for the MP test, which the authors thought represented poor interexaminer agreement.

42. Sebastian D, Chovvath R. Reliability of palpation

assessment in non-neutral dysfunctions of the lumbar spine. Orthopaedic Physical Therapy Practice 2004; 16:23–6.

The interexaminer reliability of two PTs with approximately 13 years of experience was assessed for their ability to detect dissimilar motion patterns at the L5 level. It was not clear whether the palpations were carried out independently, although the order of examination was randomly assigned. Subjects included 31 low-back pain patients, with or without leg pain. Patients were standing during the examination and were asked to bend forward, backward and to a neutral position while the examiners palpated the bilateral transverse processes of L5. The examiners’ objective was to determine if a prominence appeared over one of the transverse processes, which purportedly points to a unilateral facet fixation (i.e., “stuck” in flexion or extension). The authors reported “substantial” agreement, with an overall kappa value of 0.688.

43. Smedmark V, Wallin M, Arvidsson I. Inter-examiner reliability in assessing passive intervertebral motion of the cervical spine. Man Ther 2000; 5:97–101.

Two PTs, each having more than 25 years of experience, independently assessed the cervical regions (C1-T1) of 61 neck pain patients to determine if three tests of cervical and first rib intersegmental mobility could identify segments as being stiff versus not stiff. The order of examiners was randomized, although it was not obvious whether they were blinded from each other’s findings. The results were analyzed using percent agreement and the kappa coefficient. Interexaminer reliability was reported to be fair to moderate, with percent agreement ranging from 70% to 87% and kappa values from 0.28 and 0.43. The authors made a concerted effort to standardize patient and examiner starting positions. Also, the examiners had worked together for 17 years and had developed standards of assessment together. In spite of these efforts, the results of this study showed lower than expected levels of agreement.

44. Strender LE, Lundin M, Nell K. Interexaminer reliability in physical examination of the neck. J Manipulative Physiol Ther 1997; 20:516–20.

The interexaminer reliability of seven commonly used clinical tests was evaluated, including three tests that assessed mobility in the upper cervical region (C0-C3). Two PTs, each having at least 21 years of experience, independently examined 50 subjects, 25 of whom were symptomatic neck pain patients and 25 of whom were asymptomatic. The examiners were blinded from each other's findings and the order of examinations was randomized. Reported kappa values for the mobility tests ranged from 0.06 to 0.15 and percent agreement ranged from 26% to 44%. The values were thought to suggest poor interexaminer reliability. The authors indicated that only two of the 10 clinical tests that were evaluated had an acceptable level of reliability, the foraminal compression test and palpation for pain in the occipital muscles.

45. **Streder LE, Sjoblom A, Sundell K, Ludwig R, Taube A. Interexaminer reliability in physical examination of patients with low back pain. Spine 1997; 22:814–20.**

This study evaluated the interexaminer reliability of several clinical tests that are commonly used to evaluate low back pain patients. Seventy-one low back pain patients were examined by two PTs, who examined 50 patients, and two MDs, who examined 21 patients. All four examiners had “long clinical experience” with low back pain patients. The examinations were conducted independently and the order of examiners was randomized. The MP portion of the examination involved the assessment of intersegmental mobility of the segments between L4-L5 and L5-S1. Patients were lying on their side with hips and knees flexed and the examiner standing to their side. The patients were passively moved with one hand, while the opposite hand palpated the targeted segments. Reported kappa values for the dichotomized results ranged from -0.08 to 0.75 and percent agreement from 48% to 88%. However, the PTs' scores were much higher ($K_w = 0.75$ at L5-S1 and 0.66 at L4-L5) than those of the MDs, which were judged to be poor.

46. **Tong HC, Heyman OG, Lado DA, Isser MM. Interexaminer reliability of three methods of combining test results to determine side of sacral**

restriction, sacral base position, and innominate bone position. J Am Osteopath Assoc 2006; 106:464–8.

The SI regions of 24 low back pain patients were examined by 1 MD and 3 DOs, two examiners at a time. No information was provided as to their experience levels. They were blinded from each other's findings, but only the first examiner was blinded to the patients' histories. The study's objective was to compare the interexaminer reliability of three palpation methods used to determine: 1) the side of SI joint dysfunction, 2) sacral base position, and 3) innominate bone position. Three tests of SI joint dysfunction were carried out, including the seated flexion test, standing flexion test, and standing stork test (i.e., Gillet step test). Kappa values ranged from 0.06 to 0.11 for the seated flexion test, from 0.14 to 0.30 for the standing flexion test, and from 0.27 to 0.50 for the standing stork test.

47. **Vincent-Smith B, Gibbons P. Inter-examiner and intra-examiner reliability of the standing flexion test. Man Ther 1999; 4:87–93.**

Nine DOs examined the SI regions of nine asymptomatic subjects using the standing flexion test, which was performed by having the standing subject bend forward while the examiner palpated the level of the posterior superior iliac spines (PSISs) with their thumbs. A positive test occurred when one PSIS moved in a more cephalad or ventral direction, pointing to hypomobility on the opposite side. The examiners were blinded from each other's findings and their order of conducting the tests was randomized. Each of the examiners had at least four years of experience using the standing flexion test. The kappa value for intraexaminer reliability was 0.46 and percent agreement was 68%, which was considered to be moderate. The kappa value for interexaminer reliability 0.05 and percent agreement was 42%, which was thought to represent statistically insignificant reliability. The authors concluded that the reliability of the standing flexion test was questionable.

48. **Wiles M. Reproducibility and interexaminer correlation of motion palpation findings of the sacroiliac joints. J Can Chiropr Assoc 1980; 24:56–69.**

The SI regions of 46 asymptomatic subjects were independently examined by six pairs of DC examiners (eight total) to assess joint motion. Examiners had an average of 2.75 years of experience. Mobility assessments were graded as grade 1, “normal mobility”; grade 2, “moderate or transient restriction”; or grade 3, “severe or complete restriction.” The inferior and superior portions of the SI joints were evaluated separately, and a test was performed as well for bilateral SI motion. Interexaminer reliability was represented by Pearson’s correlation and percent agreement, which ranged from $r = 0.13$ to 0.43 and 47% to 64% respectively. The authors suggested that some of the results of the inferior and bilateral joints reached a level of statistical significance.

Funding sources and conflicts of interest
None.

Acknowledgements

We thank the Palmer College of Chiropractic West library supervisor, Wendy Kubow, for her assistance in locating and obtaining full-text articles.

References

1. Huijbregts PA. Spinal motion palpation: a review of reliability studies. *J Man Manip Ther.* 2002; 10:24–39.

2. Breen A. The reliability of palpation and other diagnostic methods. *J Manipulative Physiol Ther.* 1992; 15:54–6.
 3. Panzer DM. The reliability of lumbar motion palpation. *J Manipulative Physiol Ther.* 1992; 15:518–24.
 4. Dishman RW. Static and dynamic components of the chiropractic subluxation complex: a literature review. *J Manipulative Physiol Ther.* 1988; 11:98–107.
 5. Seffinger MA, Najm WI, Mishra SI, et al. Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature. *Spine.* 2004; 29:E413–25.
 6. Stochkendahl MJ, Christensen HW, Hartvigsen J, et al. Manual examination of the spine: a systematic critical literature review of reproducibility. *J Manipulative Physiol Ther.* 2006; 29:475–85, 485 e1–10.
 7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; 33:159–74.
 8. Shrout PE, Fleiss JL. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin.* 1979; 2:420–8.
 9. Haas M. Statistical methodology for reliability studies. *J Manipulative Physiol Ther.* 1991; 14:119–32.
 10. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2000:65–66.
 11. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol.* 1987; 126:161–9.
 12. Rosner B. Fundamentals of biostatistics. 4th ed. Belmont, Calif.: Duxbury Press, 1995.
 13. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2000.

Table 1 *Motion palpation interexaminer reliability studies*

Author, Bibliography #	Region	Examiners, experience	Subjects	Quality Score	Findings	Degree of Reliability
Bergström & Courtis, 1	L1-L5	2 DC, Pre-trained	100 Asx	67%	% = 65 to 88	Inconclusive
Binkley et al, 2	L1-S1	6 PT, at least 6 yrs	18 Sx	50%	K = 0.09 ICC = 0.25 (CI, 0–0.39)	Slight
Boline et al, 3	T12-S1	1 DC (<1 yr), 1 St	50 (23 Sx, 27 Asx)	83%	K = –0.05 to 0.33 % = 60 to 90	None to fair
Brismée et al, 4	T5-T7	3 PT, ≥12 yrs	41 Asx	50%	K = 0.27 to 0.65 % = 63 to 83	Fair to substantial
Carmichael, 5	SI	10 DC St	54 Asx	50%	K = –0.07 to 0.19 % = 66 to 100	None to slight
Christensen et al, 6	T1-T8	2 DC, Exp	107 (51 angina, 56 Asx)	100%	K = 0.22 to 0.24	Fair

Comeaux et al, 7	C2-T8	3 DO, >10 yrs	54 Asx	67%	K = 0.12 to 0.56	Slight to moderate
Deboer et al, 8	C1-C7	3 DC, Exp	40 Asx	50%	$K_w = 0.03$ to 0.42	Slight to moderate
Degenhardt et al, 9	L1-L4	3 DO, <10 yrs	15 Asx	50%	K = 0.20 % = 66	Slight
Downey et al, 10	Lumbar	6 PT, 7 to 15 yrs	30 Sx	33%	K = 0.23 to 0.54	Fair to moderate
Fjellner et al, 11	C0-C2	2 PT, 6 & 8 yrs	48 (11 Sx, 36 Asx)	67%	$K_w = 0.01$ to 0.18 % = 60 to 87	Slight
Fjellner et al, 11	C0-T5	2 PT, 6 & 8 yrs	48 (11 Sx, 36 Asx)	67%	$K_w = -0.16$ to 0.49 % = 41 to 92	None to moderate
Flynn et al, 12	S1	8 PT, Exp	55 Sx	33%	K = -0.08 to 0.59	None to moderate
Gonella et al, 13	T12-S1	5 PT, ≥ 3 yrs	5 Asx	17%	Visual inspection of raw data	Inconclusive
Haas et al, 14	T3-L1	2 DC, 15 yrs	73, 49% Sx	67%	K = 0.08 to 0.22	Slight to fair
Hanten et al, 15	C1-C3	2 PT, Exp	40 Sx	50%	K = -0.71 to 0.86 % = 70 to 95	None to almost perfect
Herzog et al, 16	S1	10 DC, >1 yrs	11(10 Sx, 1 Asx)	33%	% = 54 to 78	Inconclusive
Hicks et al, 17	L1-L5	3 PT, 1 DC/PT, 4 to 8 yrs	63 Sx	33%	$K_w = -0.02$ to 0.26 % = 52 to 69	None to slight
Humphreys et al, 18	C1-C7	20 DC St, 4 th year	3 with congenital block vertebrae	50%	K = 0.46 to 0.76	Moderate to substantial
Inscocoe et al, 19	T12-S1	2 PT, ≥ 4 yrs	6 Sx	17%	Scott's Pi = 18.4% % = 33.3 to 58.3	Not acceptable
Jull & Bullock, 20	T12-S1	2 PT, Exp	10 Asx	0%	r = 0.82 to 0.94 % = 86	Inconclusive
Keating et al, 21	T12-S1	3DC, >2.5 yrs	46 (21 Sx, 25 Asx)	67%	K = -0.18 to 0.31	None to fair
Leboeuf, 22	L1-S1	4 DC St	45 Sx	17%	% > 90	Inconclusive
Lindsay et al, 23	L1-S1	2 PT, ≥ 6 yrs	18 (Sx & Asx)	100%	$K_w = -0.03$ to 0.6 % = 14 to 100	None to moderate
Lindsay et al, 23	S1	2 PT, ≥ 6 yrs	18 (Sx & Asx)	100%	$K_w = 0.2$ to 0.6 % = 50 to 100	Slight to moderate
Love & Brodeur, 24	T1-L5	8 DC St	32 Asx	17%	r = 0.01 to 0.49	Inconclusive
Lundberg & Gerdle, 25	T10-S1	3 PT, Exp	150 Asx	50%	$K_w = 0.59$ to 0.75	Moderate to substantial
Maher & Adams, 26	L1-L5	6 PT, ≥ 5 yrs	90 Sx	67%	ICC = -0.4 to 0.73 % = 13 to 43	Poor to fair
Maher et al, 27	L3	5 PT, ≥ 5 yrs	40 Asx	33%	ICC = 0.50 to 0.77 SEM = 0.72 to 1.58	Fair to good

Marcotte et al, 28	C0-C7	25 DC (1 Exp, 24 St)	3 Asx	33%	K = 0.6 to 0.8	Moderate to substantial
Marcotte et al, 29	C0-C7	24 DC (1 Exp, 23 St)	3 Asx	33%	K = 0.7 to 0.75	Moderate
McPartland & Goodridge, 30	C0-C3	2 DO, ≥10 yrs	18 (7 Sx, 11 Asx)	83%	K = 0.34 % = 66.7	Fair
Meijne, 31	SI	2 PT St	38 (9 Sx, 29 Asx)	83%	K = -0.30 to 0.75 % = 48 to 100	None to substantial
Mior et al, 32	C0-C2	2 DC St, 3 months training	59 Asx	50%	K = 0.15 % = 61	Slight
Mior et al, 33	SI	3 DC, >5 yrs, 74 St	15 Asx	33%	K = 0.00 to 0.30	None to fair
Mootz et al, 34	L1-S1	2 DC, ≥7	60 Asx	33%	K = -0.17 to 0.17	None to slight
Nansel et al, 35	Mid & lower C	4 DC (3 Exp, 1 St)	270 Asx	50%	K = 0.01 % = 45.6 to 54.3	Almost none
Olson et al, 36	C0-C2	6 PT, ≥4.5 yrs	10 Asx	33%	K = -0.04 to 0.12	None to slight
Paydar et al, 37	SI	2 DC St	32 Asx	50%	K = 0.09 % = 34.4	Slight
Phillips & Twomey, 38	L1-L5	2 PT, NI	72 (63 Sx, 9 Asx)	67%	K _w = -0.15 to 0.32 % = 55 to 99	None to fair
Rhudy et al, 40	C1-L5	3 DC, Exp	17 Sx	50%	K values not presented	Inconclusive
Robinson et al, 41	SI	2 PT, Ave 5.8 yrs	61 (45 Sx, 16 Asx)	83%	K = -0.06 % = 48	None
Smedmark et al, 43	C1-T1	2 PT, >25 yrs	61 Sx	67%	K = 0.28 to 0.43 % = 70 to 87	Fair to moderate
Strender et al, 44	C0-C3	2 PT, ≥21 yrs	50 (25 Sx, 25 Asx)	83%	K = 0.06 to 0.15 % = 26 to 44	None to slight
Strender et al, 45	L5-S1	2 MD, 2 PT, Exp	71 Sx	67%	K = -0.08 to 0.75 % = 48 to 88	None to substantial
Tong et al, 46	SI	4 DO, NI	24 Sx	33%	Stork test K = 0.27 to 0.50 Flexion tests K = 0.06 to 0.30	Fair to moderate None to fair
Vincent-Smith & Gibbons, 47	SI	9 DO, ≥4 yrs	9 Asx	50%	K = 0.013 to 0.09 % = 34 to 50	Slight
Wiles, 48	SI	8 DC, 2.75 yrs Exp average	46 Asx	17%	r = 0.13 to 0.43 % = 47 to 64	Inconclusive

MP = motion palpation; C = cervical; T = thoracic; L = lumbar; S = sacral; SI = sacroiliac; Sx = symptomatic; Asx = asymptomatic; Inter = interexaminer reliability; Intra = intraexaminer reliability; K = Kappa; r = Pearson's correlation coefficient; % = percent agreement; CI = 95% confidence interval; SEM = standard error of measurement; DC = doctor of chiropractic; MD = doctor of medicine; DO = doctor of osteopathic medicine; PT = physical therapist; MT = manipulative therapist; St = student; Exp = experienced.

Table 2 Motion palpation intraexaminer reliability studies

Author, Bibliography #	Region	Examiners, experience	Subjects	Quality Score	Findings	Degree of Reliability
Bergström & Courtis ⁶²	L1-L5	2 DC, Pre-trained	100 Asx	50%	% = 91 to 100	Inconclusive
Carmichael ⁴⁸	SI	10 DC St	54 Asx	50%	K = -0.02 to 0.69 % = 75.5 to 100	None to fair
Christensen et al ⁶⁶	T1-T8	2 DC, Exp	107 (51 angina, 56 Asx)	100%	K = 0.59 to 0.64	Moderate to substantial
Deboer et al, 8	C1-C7	3 DC, Exp	40 Asx	25%	K _w = 0.07 to 0.40	Slight to moderate
Gonella et al, 13	T12-S1	5 PT, ≥3 yrs	5 Asx	0%	Visual inspection of raw data	Inconclusive
Hanten et al, 15	C1-C3	1 PT, Exp	20 Sx	25%	K = 0.21 to 0.80 % = 60 to 90	Fair to almost perfect
Herzog et al, 16	SI	10 DC, >1 yrs	11(10 Sx, 1 Asx)	25%	% = 68 to 79	Inconclusive
Inscoe et al, 19	T12-S1	2 PT, ≥4 yrs	6 Sx	0%	Scott's Pi = 41.9% to 61.3% % = 66.7 to 75.00	Not acceptable
Jull & Bullock, 20	T12-S1	1 PT, Exp	20 Asx	0%	r = 0.81 to 0.98 % = 87.5	Inconclusive
Love & Brodeur, 24	T1-L5	8 DC St	32 Asx	0%	r = 0.02 to 0.65	Inconclusive
Meijne, 31	SI	2 PT St	38 (9 Sx, 29 Asx)	100%	K = -0.39 to 0.65 % = 44 to 100	None to substantial
Mior et al, 32	C0-C2	2 DC St, 3 months training	59 Asx	50%	K = 0.37 to 0.52 % = 71 to 79	Fair to moderate
Mior et al, 33	SI	3 DC, >5 yrs, 74 St	20 (15 Asx, 5 with fused SI)	50%	K = 0.15 to 1.00	Slight to almost perfect
Mootz et al, 34	L1-S1	2 DC, ≥7	60 Asx	25%	K = -0.09 to 0.48	None to moderate
Olson et al, 36	C0-C2	6 PT, ≥4.5 yrs	10 Asx	25%	K = -0.02 to 0.31	None to fair
Paydar et al, 37	SI	2 DC St	32 Asx	25%	K = 0.29 % = 58.1	Fair
Potter et al, 39	SI	8 PT, >2 yrs	17 Sx	33%	% = 44 to 50	Inconclusive
Vincent-Smith & Gibbons, 47	SI	9 DO, ≥4 yrs	9 Asx	25%	K = 0.16 to 0.72 % = 44 to 88	Slight to substantial

MP = motion palpation; C = cervical; T = thoracic; L = lumbar; S = sacral; SI = sacroiliac; Sx = symptomatic; Asx = asymptomatic; Inter = interexaminer reliability; Intra = intraexaminer reliability; K = Kappa; r = Pearson's correlation coefficient; % = percent agreement; CI = 95% confidence interval; SEM = standard error of measurement; DC = doctor of chiropractic; MD = doctor of medicine; DO = doctor of osteopathic medicine; PT = physical therapist; MT = manipulative therapist; St = student; Exp = experienced.