SPINE Volume 29, Number 19, pp E413–E425
©2004, Lippincott Williams & Wilkins, Inc.

Reliability of Spinal Palpation for Diagnosis of Back and Neck Pain

A Systematic Review of the Literature

Michael A. Seffinger, DO,* Wadie I. Najm, MD,† Shiraz I. Mishra, MD, PhD,‡ Alan Adams, DC, MS,§ Vivian M. Dickerson, MD,‖ Linda S. Murphy, MLIS,¶ and Sibylle Reinsch, PhD**

Study Design. A systematic review.

Objectives. To determine the quality of the research and assess the interexaminer and intraexaminer reliability of spinal palpatory diagnostic procedures.

Summary of Background Data. Conflicting data have been reported over the past 35 years regarding the reliability of spinal palpatory tests.

Methods. The authors used 13 electronic databases and manually searched the literature from January 1, 1966 to October 1, 2001. Forty-nine (6%) of 797 primary research articles met the inclusion criteria. Two blinded, independent reviewers scored each article. Consensus or a content expert reconciled discrepancies.

Results. The quality scores ranged from 25 to 79/100. Subject description, study design, and presentation of results were the weakest areas. The 12 highest quality articles found pain provocation, motion, and landmark location tests to have acceptable reliability (K = 0.40 or greater), but they were not always reproducible by other examiners under similar conditions. In those that used kappa statistics, a higher percentage of the pain provocation studies (64%) demonstrated acceptable reliability, followed by motion studies (58%), landmark (33%), and soft tissue studies (0%). Regional range of motion is more reliable than segmental range of motion, and intraexaminer reliability is better than interexaminer reliability. Overall, examiners' discipline, experience level, consensus on procedure used, training just before the study, or use of symptomatic subjects do not improve reliability.

Conclusion. The quality of the research on interreliability and intrareliability of spinal palpatory diagnostic procedures needs to be improved. Pain provocation tests are most reliable. Soft tissue paraspinal palpatory diagnostic tests are not reliable.

Key words: reproducibility of results, palpation, observer variation, neck pain, low back pain, systematic review, diagnostic tests. **Spine 2004;29:E413–E425**

Health care professionals examine and diagnose patients with cervical, thoracic, and lumbar back pain on a daily basis. Back pain, indeed, is rated among the most important factors affecting the health status in old age and is part of a more general syndrome of poor health.[1] In one study, the prevalence of back pain, work related and non work related, was 18%, and the prevalence of lost workdays due to back pain was approximately 5%.[2] For most patients, the symptoms are nonspecific. Nonspecific or idiopathic (musculoligamentous) pain accounts for at least 70% of etiologies of low back pain.[3] Approximately 85% of neck pain is attributed to chronic musculoligamentous stresses and strains or acute or repetitive neck injuries, of which acceleration-deceleration ("whiplash") is the most common.[4]

History, physical examination and eventually diagnostic imaging and laboratory tests are used to appraise the etiology of the problem and to make sure that underlying serious pathology is not missed.[5] However, despite the fact that the presenting problem or complaint might be the same, the diagnostic evaluation often depends on the individual health care provider's specialty and training.[6] Many health care disciplines have developed their own tests, diagnostic evaluations, and language to describe and communicate their findings and management protocols.[7] Common among all is that the physical evaluation of patients presenting with a complaint of back pain often consists of several important elements, such as general observation, assessment of joint range of motion, palpation of back structures, and neurovascular examination.

The national low back pain evaluation guidelines in several countries recommend spinal palpatory diagnosis and treatment options include manipulation in the initial weeks of an acute mechanical back pain episode.[8] Spinal palpation tests used to determine if manipulative treatments are indicated and/or to evaluate the effectiveness of the intervention essentially involve assessments of symmetry of bony landmarks, quantity and quality of

From the *Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine of the Pacific, Western University of Health Sciences, Pomona, CA; Departments of †Family Medicine, ‖Obstetrics & Gynecology, and **Physical Medicine & Rehabilitation, University of California, Irvine, Medical Center, Orange, CA; ‡Department of Epidemiology and Preventive Medicine, School of Medicine, University of Maryland, Baltimore, MD; §Office for Academic Affairs and Office of the Provost, Florida State University, Tallahassee, FL; and ¶Science Library Reference Department, University of California, Irvine, CA.
Acknowledgment date: October 13, 2003. First revision date: December 16, 2003. Acceptance date: December 22, 2003.
Supported by the 41st Trust Grant, the Susan Samueli Center for Complementary and Alternative Medicine, University of California at Irvine, and the Osteopathic Manipulative Medicine Department, College of Osteopathic Medicine of the Pacific at Western University of Health Sciences.
The manuscript submitted does not contain information about medical device(s)/drug(s).
Institutional/foundation funds were received in support of this work. No benefits in any form have been or will be received from a commercial party related directly or indirectly to the subject of this manuscript.
Address correspondence and reprint requests to Michael A. Seffinger, DO, Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine of the Pacific, Western University of Health Sciences, 309 E. 2nd Street, Pomona, CA; E-mail: mseffinger@westernu.edu

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

regional and segmental motion, paraspinal soft tissue abnormalities, and tenderness on provocation. The ability to arrive at an accurate palpatory assessment depends mainly on the validity and reliability of the palpatory tests used.

Although validity and reliability are often used interchangeably in the literature, they are not synonymous. Validity is the accuracy of a measurement of the true state of a phenomenon.[9] Reliability measures the concordance, consistency, or repeatability of outcomes.[10]

Over the past 30 years, scientists with diverse professional training have investigated the validity and/or reliability of spinal palpatory tests used to diagnose nonspecific back pain.[11–13] Several narrative reviews of the literature on spinal palpatory diagnostic procedures have been published.[14–17] However, only two systematic reviews of reliability studies of spinal palpatory tests have been published. One is a limited review of chiropractic literature on palpatory diagnostic procedures for the lumbar-pelvic spine[18]; the other[19] focused on the reliability of sacroiliac joint palpatory tests. The reliability of spinal palpatory diagnostic procedures for neck and back problems remains unclear. There is no comprehensive systematic review of the literature on the reliability of cervical, thoracic, and lumbar spinal palpatory diagnostic procedures.

The authors performed a systematic review of original research articles, from all disciplines, published in peer-reviewed journals in order to assess the quality of the literature and answer the clinical question: "What is the intra- and interexaminer reliability of spinal palpatory diagnostic procedures?"

■ Materials and Methods

A multidisciplinary team conducted the systematic review at the Susan Samueli Center for Complementary and Alternative Medicine (University of California, Irvine), between October 2001 and December 2002. The research team included expertise in database searches, clinical research, evidence-based medicine, research design, and statistics methodology. The clinicians represented content area experts in osteopathic, chiropractic, and family medicine/primary care.

A comprehensive strategy, including the exploration of 13 online databases and a manual search of appropriate literature, guided the search for pertinent articles that addressed the study question. Articles were limited to human studies published in peer-reviewed journals or dissertations published between January 1, 1966 and October 1, 2001. All databases were searched using a basic search template. When appropriate, minor modifications to the basic search template were made to optimize the search strategy in individual databases. The 13 databases included: PubMed MEDLINE, MANTIS, MD Consult, Web of Science, EMBASE, CINAHL, BIOSIS Preview, Index to Chiropractic Literature, OSTMED, OCLC FirstSearch, Digital Dissertation, PEDro, and Cochrane. Selection of these databases was determined by the availability of online resources accessible from our institution and affiliated institution libraries, as well as potential inclusion of articles from osteopathic medicine, allopathic medicine, chiropractic medicine, manual medicine, and physical therapy. The manual search included gleaning references cited in studies selected from the online search, and consulting experts and researchers in the fields of chiropractic and osteopathic medicine. A detailed document of the search strategy and outcome are described in detail in another article.[20]

The inclusion/exclusion criteria were adapted, modified, and developed, after review and discussion of guidelines published by leaders in the field of systematic reviews[21] and meta-analysis.[22,23] Inclusion criteria were: articles in any language that pertained to manual spinal palpation procedures to any and all regions of the human spine (excluding the sacral region); included measurement for the intra- and/or interexaminer reliability of manual spinal palpation; published between January 1, 1966 and October 1, 2001 in a peer-reviewed journal article, monograph, or dissertation. Exclusion criteria were: articles inconsistent with the inclusion criteria; anecdotal, speculative or editorial in nature; included a whole regimen of tests or methods, without separate data for each test and/or the data for spinal palpatory procedures could not be ascertained.

Initially, 59 of 797 articles were identified by our search. On further review and discussion, 11 of these articles were excluded for the following reasons: no separate data analysis for each procedure[12,15,24–30]; no spinal palpatory diagnostic test used[31]; or data displayed only as graphics.[32] One article was added at a later date following a hand search of references found in a clinical review article.[33] Forty-nine articles met the inclusion criteria and were included in this review. Two articles were in German and one in French and reviewed by authors and/or a content expert fluent in the language.

After review and discussion of published guidelines,[21,34–38] including the Centre for Reviews and Dissemination recommendations,[34] and prior research,[39,40] the authors developed an instrument to assess the quality of the articles. The quality assessment instrument scored studies primarily on constructs pertinent to internal validity (*i.e.*, scientific rigor) and reproducibility of research. It was operational in five primary categories: study subjects, examiners, study conditions, data analysis, and results. By consensus among the authors, a weighting scheme gave more importance to certain elements within the five primary categories. For instance, a description of the palpatory procedure was weighted 8 as opposed to a description of the study conditions (*i.e.*, facilities), which was weighted as 1, indicating a higher value for the former information.

To standardize the review and scoring procedures between reviewers, the authors developed and pilot tested a brief but clear definition and coding instructions protocol. Six reviewers independently reviewed and scored all the articles selected for the study. The reviewers were blinded to the articles' authors, title, and journal. Each article was randomly assigned to two reviewers. After reviewing all the assigned articles, scores were tabulated for each category and matched. When the reviewers' scores differed by more than 10% variance (*i.e.*, ratio of standard deviation/mean), it denoted a disagreement between the paired reviewers. When disagreement was identified, reviewers met to discuss and reconcile differences in their scores on each of five primary categories (*i.e.*, study subjects, examiners, study conditions, data analysis, and results). If reviewers were unable to reconcile differences in their quality scores, the article was reviewed by two content experts and scored by consensus.

■ Results

Forty-nine articles met our inclusion-exclusion criteria and were included in this systematic review. Four of these 49 articles reported on two distinct interexaminer

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 1. Characteristics of Studies Reviewed

Characteristic	N	Percentage*
Study type		
Interrater reliability	39	74
Intrarater and interrater reliability	14	26
Study subjects		
Human	53	100
Examiner background		
Physical Therapist (P.T.), practitioner and/or student	19	36
Doctor of Chiropractic (D.C.), practitioner and/or student	15	28
Doctor of Osteopathic Medicine (D.O.) practitioner and/or student	9	17
Medical Doctor (M.D.),	6	11
Combination (P.T. and M.D. or D.C, D.O., and M.D.)	3	6
Diplomate of Osteopathy (D.O.- Australia),	1	2
Spinal location		
Cervical	14	26
Thoracic	4	8
Lumbar	24	45
Combination (cervicothoracic, thoracolumbar, full spine)	11	21
No. of studies using which types of palpatory procedures†		
Motion tests	36	68
Pain provocation tests	21	40
Paraspinal soft tissue palpation tests	12	23
Landmark position assessment tests	5	9
Consensus on palpatory procedures used		
Yes	42	79
No	7	13
Not stated	4	8
Examiners trained on palpatory procedures used		
Yes	23	43
No	22	42
Not stated	6	11
Both trained and untrained	2	4
Sample size of study subjects		
<21	16	30
21–40	9	17
41–60	15	28
>60	13	25
Sample size of examiners		
<3	23	43
3–5	18	34
>5	12	23
Study design		
Correlational or cross-sectional	36	68
Repeated measure	16	30
Other	1	2
Random selection of subjects		
Yes	4	8
No	46	87
Unclear, not known	3	6
Subjects' clinical presentation		
Symptomatic	14	26
Asymptomatic	16	30
Symptomatic and asymptomatic	9	17
Unclear, not known	14	26
Examiners blinded to subjects' medical condition		
Yes	21	40
No	7	13
Not stated	25	47
Subjects blinded to examination findings		
Yes	5	9
No	2	4
Not specified	46	87
Examiners blinded to each other's findings		
Yes	28	53
No	6	11
Not stated	19	36
Measure of association statistics used‡		
kappa (or weighted kappa)	37	70
Percent agreement	24	45
Intraclass correlation coefficient	5	9
χ^2 (observed vs. expected)	2	4
Percent disagreement	1	2
Pearson R	1	2
Other (level of agreement, F test, Scott's pi ratio, Bartlett's test)	4	8
Articles weighted mean quality scores, quartiles§		
1st quartile (67.5–79, 75.1–100%)	12	24
2nd quartile (60–67, 52.2–75.0%)	13	27
3rd quartile (48–59, 25.1–52.1%)	11	22
4th quartile (0–47, 0–25.0%)	13	27
Article publication date		
Pre-1980	1	2
1980–1984	6	1
1985–1989	12	24
1990–1994	9	18
1995–1999	15	31
2000–2001	8	16

*Numbers do not always add up to 100 due to rounding.
†The number of studies adds to more than 53 since many studies tested more than one palpatory procedure.
‡The number of studies adds to more than 53 since many studies used more than one statistical test.
§Range of weighted mean quality score and percentage are included in the parentheses.

reliability studies. Thus, the total number of studies included in the 49 articles is 53. Descriptions of the characteristics of the studies are summarized in Table 1.

Paired reviewers initially disagreed on the quality score of 16 (33%) of the 49 articles. Quality scores of the 49 articles ranged from 25 to 79 of 100. The authors compared quality scores of articles in the top quartile (67.5–79) to those in the bottom quartile (25–47). No correlation between quality score and year of publication, examiners' disciplines (clinical degree or specialty training), or procedure evaluated was found. All studies were lacking in description of subjects. Study design, description of study conditions and examiners' professional training, data analysis, and presentation of results were the weakest areas in the lower quality studies.

Interestingly, symptomatic (back or neck pain) subjects were recruited only in 14 (26%) of the 53 studies, and both symptomatic and asymptomatic subjects were recruited in only 9 of 53 (17%). Additionally, two studies assessed the effect of hypertensive subjects on the reliability of palpatory findings.[41,42]

The authors synthesized the data only from the higher quality articles (quality score 67.5 of 100 or greater). Most (two thirds) of the higher quality articles used the more rigorous kappa or weighted kappa measure of association to determine degree of reliability.[43] Results and characteristics of all of the studies are reported in Tables 2 through 5. These tables are organized per palpatory test used under the categories of: motion tests, pain provocation tests, soft tissue tests, and landmark tests. Articles that reported on the reliability of a variety of palpatory tests appear in more than one evidence table.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 2. Quality Scores, Study Characteristics, and Intraexaminer and Interexaminer Reliability for Motion Palpation Tests

Study	Quality Score	Examiners, Subjects	Type of Reliability, Spinal Motion Tests, and Results	Interpretation*
Strender *et al*[48]	79.0	2 PT; 25 Sx, 25 ASx subjects	InterEx, cervical segmental K = 0.09–0.15; 26–44% agreement	Low reliability
Schops *et al*[49]	77.5	5 Physicians; 20 Sx subjects	InterEx, cervical and thoracic segmental K = 0.6–0.8 for 1st 2 examiners; 0.2–0.4 for all 5	Low to high reliability, examiner dependent
Fjellner *et al*[44]	74.0	2 PT; 47 (11 Sx and 35 ASx, 1 UMS) subjects	InterEx, cervical and thoracic, regional and segmental Regional ROM: Kw > 0.4 in 6 of 8 tests except for rotation; Regional end-feel motion tests: Kw > 0.4 in 3 of 8 tests Passive segmental tests: Kw > 0.4 in 5 of 58 exams	Regional ROM, except for rotation, some end-feel and some segmental motion tests: medium reliability; most end-feel and segmental exams had low reliability
Love *et al*[45]	72.0	8 DC students; 32 ASx subjects	IntraEx and InterEx, thoracic and lumbar segmental IntraEx: Pearson's r = 0.302–0.6856 InterEx: Index of Association statistic (R) = 0.023–0.0852	IntraEx more reliable than InterEx
Johnston *et al*[42]	71.0	3 DO; 307 (153 hypertensive) subjects	InterEx, cervical and thoracic segmental Higher level of InterEx agreement in subsample with more hypertensives ($\chi^2 = 27.75$, $df = 1$, $P < 0.001$)	More reliable in hypertensive subjects
Lundberg *et al*[52]	68.0	2 PT; 150 UMS subjects	InterEx, thoracic and lumbar segmental K (w) = 0.42–0.75	Medium to high reliability
Keating *et al*[46]	67.5	3 DC; 46 (21 Sx and 25 ASx) subjects	InterEx, thoracic and lumbar segmental Active motion palpation mean K = 0.00–0.25; Passive motion palpation mean K = −0.03–0.23	Low reliability; no significant differences between Sx and ASx subjects
Johnston *et al*[41]	67.0	3 DO (2 students); 132 Asx (some hypertensive) subjects	InterEx, cervical and thoracic segmental 39.5% observed vs. 26.0% expected agreement, $P < 0.05$	More reliable in hypertensive subjects
Maher *et al*[66]	66.0	6 PT; 90 Sx subjects	InterEx, lumbar segmental 13–43% agreement ICC = −0.4 −0.73	Low reliability
Grant *et al*[67]	65.5	4 DC students; 60 UMS subjects	IntraEx and InterEx, lumbar segmental IntraEx: 85–90% agreement InterEx: 66.7% agreement	IntraEx more reliable than InterEx
Haas *et al*[68]	64.5	2 DC; 73 (48 Sx and 25 ASx) subjects	IntraEx and InterEx, thoracic segmental IntraEx: K = 0.43–0.55 InterEx: K = 0.14 (segmental level) and K = 0.19 (segmental restriction)	IntraEx: medium reliability; InterEx: low reliability; no difference between Sx and ASx subjects
Deboer *et al*[69]	64.5	3 DC; 40 Asx subjects	IntraEx and InterEx, cervical segmental IntraEx: 45–75% agreement; K (w) = 0.01–0.76 InterEx: 21–58% agreement; K (w) = −0.03–0.45	IntraEx: low reliability, except one value was high at C1–C2; InterEx: low to medium reliability, more reliable at C6–C7 than C1–C5
Phillips *et al*[70]	63.0	2 PT; 72 (63 Sx and 9 ASx) subjects	InterEx, lumbar segmental 55–100% agreement K (w) = −0.15–0.32	Low reliability; includes quality of motion and end-feel or tissue response during motion testing
Strender *et al*[53]	62.5	2 PT; 50 Sx subjects	InterEx, lumbar regional and segmental Regional ROM: 87–94% agreement; K = 0.43–0.74 Segmental: 72–88% agreement; K = 0.38–0.75	Regional ROM–extension and lateral bend: medium reliability Segmental: medium to high reliability at lumbosacral joint and 'one segment above it'
Strender *et al*[53]	62.5	2 MD; 21 Sx subjects	InterEx, lumbar regional and segmental Regional ROM: 83–86% agreement; K = 0.11–0.35 Segmental: 48–86% agreement; K = −0.08–0.24	Regional ROM–extension and lateral bend: low reliability Segmental: low reliability
Mastriani *et al*[71]	61.5	3 PT; 16 Sx subjects	InterEx, lumbar segmental L3–L4: 70–73% agreement; All segments combined: 62–66% agreement	Low reliability; more reliable at L3–L4
Boline *et al*[72]	60.0	2 DC (1 student); 50 (23 Sx and 27 ASx) subjects	InterEx, lumbar segmental K = −0.05–0.31	Low reliability; no significant differences between Sx and ASx subjects
Inscoe *et al*[73]	59.0	2 PT; 6 Sx subjects	IntraEx and InterEx, lumbar segmental IntraEx: 66.67% and 75.00% agreement; Scott's pi = 41.89% and 61.29% InterEx: % = 48.61% agreement; Scott's pi = 18.35%	IntraEx more reliable than InterEx
Nansel *et al*[74]	58.5	4 DC (1 student); 270 Asx subjects	InterEx, cervical segmental K = 0.013	Low reliability
Marcotte *et al*[55]	58.0	3 DC; 12 Sx subjects	IntraEx (only 1 examiner) and InterEx, cervical regional IntraEx: 90.6% agreement; K = 0.78 (trained examiner), $P < 0.01$ InterEx: 82.3–93.2% agreement; K = 0.57–0.85, $P < 0.01$	Regional ROM (end-feel) IntraEx reliability: high reliability; InterEx (even if 1 examiner is untrained) medium to high reliability; kappa higher among the 2 trained examiners
Johnston *et al*[75]	56.5	5 DO (3 students); 70 UMS subjects	InterEx, cervical segmental Permutation testing (a measure of agreement) of the sum (D) of the absolute value of difference between the 2 examiners and each of the 3 students Student 1: D (mean) = 15.2, SD = 2.0, $P < 0.01$ Student 2: D (mean) = 13.2, SD = 3.5, $P < 0.15$ Student 3: D (mean) = 15.6, SD = 3.5, $P < 0.35$	Significant InterEx reliability for 1 of the 3 student examiners when compared with the 2 osteopathic physicians
Bergstrom *et al*[76]	55.5	2 DC students; 100 UMS subjects	IntraEx and InterEx, lumbar segmental IntraEx for segmental level and direction: 95.4% agreement for both examiners; InterEx for both level and direction: 81.8% agreement; for level only: 74.8% agreement	Medium reliability; IntraEx more reliable than InterEx
Mior *et al*[13]	55.5	2 DC; 59 Asx subjects	IntraEx and InterEx, cervical segmental IntraEx: K = 0.37 and 0.52 InterEx: K = 0.15	IntraEx: low to medium reliability InterEx: low reliability

(*Table continues*)

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 2. Continued

Study	Quality Score	Examiners, Subjects	Type of Reliability, Spinal Motion Tests, and Results	Interpretation*
Mootz *et al*[77]	55.0	2 DC; 60 UMS subjects	IntraEx and InterEx, lumbar segmental IntraEx: K = −0.11–0.48 and 0.05–0.46 InterEx: K = −0.19–0.17	IntraEx: low to medium reliability InterEx: low reliability
Johnston *et al*[78]	54.0	3 DO (2 students); 161 UMS subjects	InterEx, cervical regional Rotation: observed agreement = 18, expected agreement = 8.3, z = 3.64, α = 0.0005 Side-bending: observed agreement = 12, expected agreement = 5, z = 2.5, α = 0.03	Regional ROM: reliable (for rotation and side-bending)
Comeaux *et al*[79]	52.5	3 DO; 54 UMS subjects	InterEx, cervical and thoracic segmental	Low to medium reliability
Maher *et al*[80]	51.5	3 graduate PT students; 13 Asx subjects	K = 0.16–0.43 InterEx, lumbar segmental ICC = 0.50–0.62 ($P < 0.05$)	Posterior-Anterior pressure test at L3 (stiffness): low reliability
Maher *et al*[80]	51.5	2 PT; 27 ASx subjects	InterEx, lumbar segmental ICC = 0.77 ($P < 0.05$)	Posterior-anterior pressure test at L3 (stiffness): medium reliability; experience level, training, and consensus may have improved reliability
Binkley *et a*[81]	47.0	6 PT; 18 Sx subjects	InterEx, lumbar segmental For judgment on marked segmental level: K = 0.30, ICC = 0.69 For mobility rating on marked level: K = 0.09, ICC = 0.25	Posterior-anterior pressure test at L1–L5: low reliability
Smedmark *et al*[82]	42.0	2 PT; 61 Sx subjects	InterEx, cervical segmental 70–87% agreement; K = 0.28–0.43	Low to medium reliability
Richter *et al*[83]	40.0	5 MD; 61 Sx (26 IntraEx; 35 InterEx) subjects	IntraEx and InterEx, lumbar segmental IntraEx: K = 0.3–0.80 (tests combined and averaged) InterEx: left side-bending at L1–L2: K = 0.69–0.72 InterEx: for other motion tests at each lumbar level: K = 0.08–0.47	IntraEx: low to high reliability InterEx: low to medium reliability except for left side-bending at L1–L2 which was medium reliability
Olson *et al*[84]	37.5	6 PT; 10 ASx subjects	IntraEx and InterEx, cervical segmental IntraEx: K (for mobility) = −0.022–0.137 InterEx: K (for mobility) = −0.031–0.182 IntraEx: K (for end-feel) = 0.01–0.308 InterEx: K (for mobility) = −0.043–0.194	IntraEx and InterEx: low reliability
Lindsay *et al*[85]	35.0	2 PT; 8 UMS subjects	InterEx, lumbar segmental 8/20 tests had > 70% agreement; K = −0.5–0.30	Majority had low reliability, although 3 tests had 100% agreement (kappa not calculated with 100% agreement)
Rhudy *et al*[86]	34.0	3 DC; 14 Sx subjects	InterEx, full spine segmental Strength of agreement [(K score/sample size) x 100]: low = 35%, substantial = 11%, moderate = 12%, medium = 9%, almost perfect = 8%, not observed = 25%	Majority of tests had less than medium reliability
Van Suijlekom *et al*[87]	33.5	2 MD; 24 Sx subjects	InterEx, cervical segmental K = 0.27–0.46	Low to medium reliability
Johnston *et al*[11]	30.0	3 DO (2 students); 10 UMS subjects	InterEx, cervical and thoracic segmental 40–60% agreement before landmark marking; 54–75% agreement after landmark marking	Low reliability; improved reliability with landmark marking

PT = physical therapist; DO = doctor of osteopathic medicine; DC = doctor of chiropractic; MD = medical doctor; Sx = symptomatic; Asx = asymptomatic; UMS = undefined medical status; IntraEx = intraexaminer; InterEx = interexaminer; K = kappa; C = cervical; T = thoracic; L = lumbar.
*The examiners' reliability rating indicated as reliable or unreliable is based on measures of association such as kappa (K) or weighted kappa (K (w)), Pearson r, or Index of Association. The kappa value is the difference between observed and expected agreement (K = observed agreement-expected agreement/1 − expected agreement). kappa values range from −1 to 1, with 1 signifying complete agreement, 0 signifying agreement no better than by chance, and −1 signifying complete disagreement. Commonly accepted interpretations of the kappa statistic are 0.00–0.39 = poor or low reliability; 0.40–0.74 = fair to good, or medium reliability; 0.75–1.00 = excellent or high reliability. The authors determined a test to have acceptable reliability if the kappa value was 0.40 or greater. If kappa values were provided in addition to percent agreement, the more rigorous kappa value was used as the preferred statistic to determine level of reliability. For percent agreement and intraclass correlation coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analyses required a case-by-case analysis to make the determination of degree of reliability.

The majority of spinal palpatory diagnostic tests demonstrated low reliability. Data from the higher quality studies (quality score 67.5 of 100 or greater) showed acceptable reliability for the following spinal palpatory diagnostic procedures: 1) interexaminer regional range of motion of the cervical spine[44]; 2) intraexaminer thoracic and lumbar segmental vertebral motion tests[45]; 3) interexaminer pain provocation at a) L4–L5 and L5–S1,[46] b) lumbar paraspinal myofascial trigger points (between trained examiners only),[47] c) the cervical spine,[48,49] and d) at T1 and the sternocleidomastoid (SCM) muscle[49]; and 4) identification of a nominated lumbar vertebral spinous process.[50,51] One study found cervical and thoracic segmental motion tests to be more reliable in hypertensive subjects.[42]

There were mixed reliability results for interexaminer cervical, thoracic, and lumbar segmental vertebral motion tests. One study showed a medium to high degree of reliability in these procedures,[52] but others did not.[45,46,48] Two studies had mixed results depending on the examiners or the tests they used,[44,49] demonstrating that these palpatory procedures were not consistently reproducible by other examiners under similar study conditions.

Only one study compared the reliability of examiners from one discipline with the reliability of examiners from a different discipline (two physical therapists *vs.* two medical physicians) using the same tests.[53] Although physical therapists were more reliable than physicians in using segmental vertebral motion tests, they were otherwise comparable in terms of reliability of other tests.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 3. Quality Scores, Study Characteristics, Intra- and InterEx Reliability for Pain Provocation Tests

Study	Quality Score	Examiners, Subjects	Type of Reliability, Spinal Region, Pain Provocation Test, and Results	Interpretation*
Strender *et al*[48]	79.0	2 PT; 50 (25 Sx and 25 ASx) subjects	InterEx, cervical digital pressure K = 0.31–0.52; 58–76% agreement	Low to medium reliability; no difference between Sx and ASx subjects
Schops *et al*[49]	77.5	5 Physicians; 20 Sx subjects	InterEx, cervical and thoracic digital pressure K = 0.2–0.6 C-spine; K = 0.6–0.75 T1; K = 0.2–0.75 muscles	C: low to medium reliability T1: medium reliability Muscles: low reliability, except SCM which had medium reliability
Hsieh *et al*[47]	69.0	8 examiners: 1 expert MD; 4 trained: 2 DC, 1 DO and 1 MD; 4 untrained: 2 DC and 2 MD; 52 (26 Sx and 26 ASx) subjects	InterEx, lumbar referred pain upon digital pressure on trigger point InterEx: Trained K = 0.435; Untrained K = 0.320 Agreement with expert: Trained K = 0.337; Untrained K = 0.292	Low reliability overall except for medium reliability between trained examiners, but not with expert
Lundberg *et al*[52]	68.0	2 PT; 150 UMS subjects	InterEx, thoracic and lumbar digital pressure L4–L5: K = 0.71 L5-S1: K = 0.67	L4–L5 and L5–S1: medium reliability Data for thoracic and other lumbar segments not reported
Keating *et al*[46]	67.5	3 DC; 46 (21 Sx and 25 ASx) subjects	InterEx, thoracic and lumbar bony and soft tissue digital pressure K = 0.22–0.42 for soft tissue pain; K = 0.34–0.65 for osseous pain (mean 0.48)	Low to medium reliability; L4–L5 and L5-S1 had greater concordance for osseous pain (mean K > 0.6); no significant difference between Sx vs. ASx subjects
Maher *et al*[66]	66.0	6 PT; 90 Sx subjects	InterEx, lumbar predictive reliability of subject's pain upon palpation 27–57% agreement; ICC: 0.27–0.85	Low to occasionally reliable
McPartland *et al*[88]	66.0	2 DO; 18 (7 Sx and 11 ASx) subjects	InterEx, cervical digital pressure on 'Strain-counterstrain' tenderpoints Sx subjects: 72.7% agreement; K = 0.45; ASx subjects: 59.43% agreement; K = 0.19	Medium reliability in Sx subjects; low reliability in ASx subjects
McPartland *et al* [88]	66.0	18 DO students; 18 ASx subjects	InterEx, cervical digital pressure on 'Strain-counterstrain' tenderpoints 64.2% agreement; K = 0.2	Low reliability
Deboer *et al*[69]	64.5	3 DC; 40 ASx subjects	IntraEx and InterEx, cervical digital pressure IntraEx: C1–C3: 55–80% agreement, Kw = 0.3–0.56; C4–C7: 60–68% agreement, Kw = 0.2–0.43; InterEx: C1–C3: 43–66% agreement, Kw = 0.08–0.48; C4–C7: 34–53% agreement, Kw = - 0.04–0.18	Both IntraEx and InterEx: low to medium reliability; IntraEx more reliable than InterEx reliability; both more reliable at C1–C3 than C4–C7
Strender *et al*[53]	62.5	2 PT; 50 Sx subjects	InterEx, lumbar paravertebral digital pressure and segmental, lateral bend, extension, flexion, foramen compression passive motion tests 78–98% agreement; K = 0.27 for paravertebral tenderness; K = 0.43–0.76 for regional lateral bend, flexion, extension pain and segmental lumbosacral and 'one segment above' lumbosacral pain; foramen compression test: 94% agreement Sensibility at L4: 98% and L5: 97% agreement; all 3 tests: prevalence < 10%†	Training made no difference; paravertebral tenderness: low reliability; segmental, lateral bend, extension and flexion pain, foramen compression test, and sensibility at L4 and L5 upon digital pressure all had medium to high reliability
Strender *et al*[53]	62.5	2 MD; 21 Sx subjects	InterEx, lumbar paravertebral digital pressure, and segmental, lateral bend, extension and flexion, foramen compression passive motion tests Lateral bend pain: 73% agreement; K = 0.06. Extension and flexion pain: 86% agreement; K = 0.71. Paravertebral tenderness: 76%, K = 0.22. Lumbosacral segment and 'one above it' tenderness: 71% agreement; K = 0.40 Foramen compression test: 98% agreement; sensibility at L4 and L5–L100% agreement; prevalence < 10%†	Lateral bend pain and paravertebral tenderness: low reliability Extension and flexion pain: medium reliability Lumbosacral segment and 'one segment above it': medium reliability Foramen compression test and sensibility at L4–L5: high reliability
Hubka *et al*[89]	62.0	2 DC; 30 Sx subjects	InterEx, cervical digital pressure 76.6% agreement; K = 0.68	Medium reliability
Boline *et al*[72]	60.0	2 DC (1 student); 50 (23 Sx and 27 ASx) subjects	InterEx, lumbar digital pressure Sx subjects: L2–L3 and L3–L4 only: 96% agreement; K = 0.65; Other lumbar levels: 81% (L5–S1)-91% (T12–L1 and L1–L2) agreement; K = 0–0.06 Both ASx and Sx subjects combined: 90–96% agreement; K = −0.03–0.37 at T12–L2 and L3–S1; K = 0.49 at L2–L3	Sx subjects at L2–L3 and L3–L4: medium reliability; rest of L-spine: low reliability With both Sx and Asx subjects at L2–L3: medium reliability; rest of L-spine: low reliability
Viikari-Juntura *et al* [90]	58.5	1 MD and 1 PT; 52 Sx subjects	InterEx, cervical (C5–C8) digital pressure tenderness, sensitivity and foramen compression passive motion test K = 0.24–0.56 for tenderness to palpation; K = 0.41–0.64 for sensitivity testing; K = 0.28–0.77 for segmental foramen compression test for radiculopathy	Tenderness: low to medium reliability; sensitivity: medium reliability Foramen compression test: low to high reliability; most reliable for radicular symptoms to the forearm
Nice *et al*[91]	52.0	12 PT; 50 Sx subjects	InterEx, lumbar trigger point digital pressure 76–79% agreement, K = 0.29–0.38	Low reliability; improved reliability noted when examiners followed proper technique per protocol and subjects reported Sx immediately prior to examination
Boline *et al*[92]	43.0	3 DC; 28 Sx subjects	InterEx, lumbar osseous and soft tissue digital pressure Osseous pain provocation: 79–96% agreement, K = 0.48–0.90; Soft-tissue pain provocation: 75–93% agreement, K = 0.40–0.78	Both had medium to high reliability

(*Table continues*)

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 3. Continued

Study	Quality Score	Examiners, Subjects	Type of Reliability, Spinal Region, Pain Provocation Test, and Results	Interpretation*
Richter *et al*[83]	40.0	5 MD; 61 Sx subjects	Intra- and InterEx, lumbar digital pressure IntraEx: K = 0.8; InterEx: K = 0.00–0.65	IntraEx: high reliability InterEx: low to medium reliability
Waddell *et al* [93]	37.0	4 MD; 810 (475 Sx and 335 ASx) subjects	InterEx, lumbar digital pressure K = 1.0 in ASx subjects (*i.e.,* agreed on lack of pain)	ASx subjects: high reliability
Van Suijlekom *et al*[87]	33.5	2 MD; 24 Sx subjects	InterEx, cervical extension and right rotation passive motion tests and digital pressure Pain with movement: K = 0.53–0.67; Vertebral joint pain with digital pressure: K = 0.15–0.37; Posterior SCM:K = 0.6–1.0	Pain upon extension and right rotation had medium to medium reliability Palpation posterior to SCM: high reliability Joint pain provoked with digital pressure: low reliability
McCombe *et al*[33]	25.0	2 MD; 50 UMS subjects	InterEx, lumbar paravertebral and midline digital pressure Paravertebral: K = 0.11 Midline: K = 0.38	Both had low reliability
McCombe *et al*[33]	25.0	1MD, 1PT; 33 UMS subjects	InterEx, lumbar paravertebral and midline digital pressure Paravertebral: K = 0.38 Midline: K = 0.47	Paravertebral soft tissue tenderness: low reliability; midline tenderness: medium reliability

PT = physical therapist; DO = doctor of osteopathic medicine; DC = doctor of chiropractic; MD = medical doctor; Sx = symptomatic; Asx = asymptomatic; UMS = undefined medical status; IntraEx = intraexaminer; InterEx = interexaminer; K = kappa; C = cervical; T = thoracic; L = lumbar; S = sacral; SCM = sternocleidomastoid muscle.

*The examiners' reliability rating indicated as reliable or unreliable is based on measures of association such as kappa (K) or weighted kappa (K (w)), Pearson r, or Index of Association. The kappa value is the difference between observed and expected agreement (K = observed agreement-expected agreement/1 − expected agreement). kappa values range from −1 to 1, with 1 signifying complete agreement, 0 signifying agreement no better than by chance, and −1 signifying complete disagreement. Commonly accepted interpretations of the kappa statistic are 0.00–0.39 = poor or low reliability; 0.40–0.74 = fair to good, or medium reliability; 0.75–1.00 = excellent or high reliability. The authors determined a test to have acceptable reliability if the kappa value was 0.40 or greater. If kappa values were provided in addition to percent agreement, the more rigorous kappa value was used as the preferred statistic to determine level of reliability. For percent agreement and intraclass correlation coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analyses required a case-by-case analysis to make the determination of degree of reliability.

†K not calculated for > 90% agreement or prevalence < 10%.

There are informative trends noticeable among the higher quality quartile studies that used the same statistical analysis. In those studies that used kappa statistics, a higher percentage of the pain provocation studies (7 of 11; 64%) demonstrated acceptable reliability followed by motion studies (7 of 12; 58%), landmark studies (1 of 3; 33%), and soft tissue studies (0 of 11; 0%). No spinal region affected pain provocation palpatory diagnostic test reliability. Among motion studies, regional range of motion was more reliable than segmental range of motion assessments. Overall, intraexaminer reliability was better than interexaminer reliability.

Paraspinal soft tissue palpatory tests had low interexaminer reliability in all regions, even though they are one of the most commonly used palpatory diagnostic procedures in clinical practice, especially by manual medicine practitioners.

The level of clinical experience of the examiners did not improve the reliability of the procedures; *i.e.*, experienced clinicians faired no better than students in terms of palpatory test reliability. Contrary to common belief, examiners' consensus on procedure used, training just before the study, or use of symptomatic subjects, did not consistently improve reliability of spinal palpatory diagnostic tests, confirming conclusions made previously by other researchers.[54]

■ Discussion

This is the most comprehensive systematic review on the intra- and interexaminer reliability of spinal palpatory procedures used in the evaluation and management of back and neck pain. The primary findings of this systematic review indicate that, in general, the quality of the research on inter- and intrareliability of spinal palpatory diagnostic procedures needs to be improved. Subject description, study design, and presentation of results were the weakest areas. Pain provocation, regional motion, and landmark location tests have acceptable reliability (kappa = 0.40 or greater), but they were not always reproducible by other examiners under similar conditions.

Among the tests reviewed, pain provocation tests are the most reliable and soft tissue paraspinal palpatory diagnostic tests are the least reliable. Regional range of motion tests are more reliable than segmental range of motion tests, and intraexaminer reliability is better than interexaminer reliability. The results of several of the lower quality articles differed from those of the higher quality articles (*i.e.*, compare Fjellner *et al*[44] with Marcotte and Normand[55] in regards to "end feel" reliability).

Given that the majority of palpatory tests studied, regardless of the study conditions, demonstrated low reliability, one has to question whether the palpatory tests are indeed measuring what they are intending to measure. That is to say, is there content validity of these tests? Indeed, there is a paucity of research studies addressing the content validity of these procedures.[56] If spinal palpatory procedures do not have content validity, it is unlikely they will be reproducible (reliable). Obviously, those spinal palpatory procedures that are invalid or unreliable should not be used to arrive at a diagnosis, plan treatment, or assess progress.

Many argue that assessment for bony or soft tissue sensitivity or tenderness is a patient subjective evaluation and not a true physical finding. However, since it is the same patient that responds to each examiner's prodding, there is, of course, a higher reproducibility of these pro-

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 4. Quality Scores, Study Characteristics, Intra- and InterEx Reliability for Soft Tissue Tests

Study	Quality Score	Examiners Subjects	Type of Reliability, Spinal Region, Soft Tissue Test and Results	Interpretation*
Strender *et al*[48]	79.0	2 PT; 50 (25 Sx and 25 ASx) subjects	InterEx, cervical consistency of occipital muscles and C2-C3 facet capsule 36–70% agreement, K = −0.18–0.24	Low reliability
Schops *et al*[49]	77.5	5 MD; 50 (25 Sx and 25 ASx) subjects	InterEx, cervical and thoracic paraspinal soft tissue tone K = 0.2–0.4	Low to medium reliability
Rouwmaat *et al*[94]	73.5	12 PT; 12 ASx subjects	IntraEx and InterEx, thoracic skin fold thickness test IntraEx: ICC:0.25–0.28; InterEx: ICC: 0.08–0.12	Both IntraEx and InterEx had low reliability; practice time and marking spinal levels were not helpful in improving reliability
Ghoukassian *et al*[95]	69.5	10 DO (Australia), 'senior post graduate'; 19 ASx subjects	InterEx, thoracic segmental tissue feel of compliance upon percussion K = 0.07	Low reliability
Hsieh *et al*[47]	69.0	8 examiners: 1 expert MD; 4 trained: 2 DC, 1 DO and 1 MD; 4 untrained: 2 DC and 2 MD; 52 (26 Sx and 26 ASx) subjects	InterEx, lumbar Taut band and local twitch response test Taut band: Trained K = 0.108 Untrained K = −0.019 With expert: Trained K = 0.238 Untrained K = 0.042 Twitch: Trained K = −0.001 Untrained K = 0.022 With expert: Trained K = 0.147 Untrained K = 0.104	Low reliability regardless of training or experience level
Keating *et al*[46]	67.5	3 DC; 46 (21 Sx and 25 ASx) subjects	InterEx, thoracic and lumbar muscle tension palpation Mean K = −0.07–0.21	Low reliability
Deboer *et al*[69]	64.5	3 DC; 40 ASx subjects	IntraEx and InterEx, cervical muscle tension palpation IntraEx: 38–93% agreement; Kw = 0.19–0.47 InterEx: 24–45% agreement; Kw = −0.1–0.53	Both IntraEx and InterEx had low to medium reliability
Boline *et al*[72]	60.0	2 DC (1 student); 50 (23 Sx and 27 ASx) subjects	InterEx, lumbar paraspinal muscle hypertonicity Both Sx and ASx subjects combined: 65–70% agreement; K = 0.10–0.31; Sx only: 51–74% agreement; K = −0.16–0.33	Low reliability; no difference in reliability between Sx vs. ASx subjects
Viikari-Juntura *et al*[90]	58.5	1 MD, 1 PT; 52 Sx subjects	InterEx, cervical paraspinal muscle tone K = 0.4	Medium reliability
Johnston *et al* [96]	54.0	6 DO (5 students); 30 UMS subjects	InterEx, thoracic paraspinal soft tissue tension assessed by percussion (finger tapping) Expected agreement 20.75 vs. Observed agreement 61; 79–86% agreement	Medium reliability
Comeaux *et al*[79]	52.5	3 DO; 54 UMS subjects	InterEx, cervical and thoracic paraspinal muscle tone assessed by finger pressure or percussion K = 0.16–0.43	Low to medium reliability
Eriksson *et al*[97]	47.0	2 PT; 19 ASx subjects	InterEx, thoracic and lumbar paraspinal muscle tone Thoracic muscles: 73.6% agreement; K = 0.16; Lumbar muscles: 94.7% agreement; K = 0.82	Thoracic: low reliability; Lumbar: high reliability

PT = physical therapist; DO = doctor of osteopathic medicine; DO(Australia) = diplomate of osteopathy in Australia; DC = doctor of chiropractic; MD = medical doctor; Sx = symptomatic; Asx = asymptomatic; UMS = undefined medical status; IntraEx = intraexaminer; InterEx = interexaminer; K = kappa; C = cervical; T = thoracic; L = lumbar.

*The examiners' reliability rating indicated as reliable or unreliable is based on measures of association such as kappa (K) or weighted kappa (K (w)), Pearson r, or Index of Association. The kappa value is the difference between observed and expected agreement (K = observed agreement-expected agreement/1 − expected agreement). kappa values range from −1 to 1, with 1 signifying complete agreement, 0 signifying agreement no better than by chance, and −1 signifying complete disagreement. Commonly accepted interpretations of the kappa statistic are 0.00–0.39 = poor or low reliability; 0.40–0.74 = fair to good, or medium reliability; 0.75–1.00 = excellent or high reliability. The authors determined a test to have acceptable reliability if the kappa value was 0.40 or greater. If kappa values were provided in addition to percent agreement, the more rigorous kappa value was used as the preferred statistic to determine level of reliability. For percent agreement and intraclass correlation coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analyses required a case-by-case analysis to make the determination of degree of reliability.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 5. Quality Scores, Study Characteristics, Intra- and InterEx Reliability for Landmark Tests

Study	Quality Score	Examiners Subjects	Type of Reliability, Spinal Region, Landmark Test, and Results	Interpretation*
Downey *et al*[50]	72.0	6 PT; 60 Sx subjects	InterEx, lumbar location of nominated lumbar spinal level K = 0.44–0.88 for agreement on one nominated level; Kw = 0.86–0.98 (scale and criteria not reported)	Medium to high reliability; selected examiners trained and educated in manipulative therapy, and accepted a range of determinations as being concordant; improved agreement by design: allowed for a range of selections for a landmark (*i.e.,* within 25 mm of each other) as opposed to discrete identification of a part of a bony landmark
Byfield *et al*[51]	67.5	2 DC; 42 ASx subjects	IntraEx and InterEx, lumbar location of bony landmarks IntraEx: 9–62% agreement; InterEx: 55–79% (sitting), 69–81% agreement (prone)	IntraEx: low reliability; InterEx: better reliability, especially at L4.
Keating *et al*[46]	67.5	3 DC; 46 (21 Sx and 25 ASx) subjects	InterEx, thoracic and lumbar misalignment of landmarks Mean K = −0.08–0.03	Low reliability
Binkley *et al*[81]	47.0	6 PT; 18 Sx subjects	InterEx, lumbar identification of a marked spinal segment K = 0.3 ICC = 0.69 (95% CI = 0.53–0.82)	Low reliability
McKenzie *et al*[98]	41.5	17 PT; 10 ASx subjects	IntraEx and InterEx, lumbar location of bony landmarks IntraEx: 84–96% agreement, K = 0.61–0.90; InterEx: 56% agreement, K = 0.28	IntraEx: medium to high reliability InterEx: low reliability

PT = physical therapist; DO = doctor of osteopathic medicine; DC = doctor of chiropractic; MD = medical doctor; Sx = symptomatic; Asx = asymptomatic; UMS = undefined medical status; IntraEx = intraexaminer; InterEx = interexaminer; K = kappa; C = cervical; T = thoracic; L = lumbar.
*The examiners' reliability rating indicated as reliable or unreliable is based on measures of association such as kappa (K) or weighted kappa (K (w)), Pearson r, or Index of Association. The kappa value is the difference between observed and expected agreement (K = observed agreement-expected agreement/1 − expected agreement). kappa values range from −1 to 1, with 1 signifying complete agreement, 0 signifying agreement no better than by chance, and −1 signifying complete disagreement. Commonly accepted interpretations of the kappa statistic are 0.00–0.39 = poor or low reliability; 0.40–0.74 = fair to good, or medium reliability; 0.75–1.00 = excellent or high reliability. The authors determined a test to have acceptable reliability if the kappa value was 0.40 or greater. If kappa values were provided in addition to percent agreement, the more rigorous kappa value was used as the preferred statistic to determine level of reliability. For percent agreement and intraclass correlation coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analyses required a case-by-case analysis to make the determination of degree of reliability.

cedures. In a systematic review of the content validity of spinal palpatory tests, the authors found that pain scales were one of only a few validated instruments that can be used in these types of studies.[56]

The spinal examination, with its small joints and limited mobility, may be more difficult for most clinicians than more prominent joints. The larger joints of the extremities fare slightly better (*i.e.*, physical therapists assessing shoulder motion restriction, kappa = 0.62–0.76).[57] However, the smaller joints of the extremities, like the vertebral spine, are less reliable (*i.e.*, kappa = 0.24–0.60 among rheumatologists palpating for hard tissue enlargement of hand and finger joints).[58]

Evaluation of the reliability of physical examination procedures in general poses a number of methodologic challenges. Examiner bias and inconsistency create variability in procedures. Although palpation for pedal pulses has medium to high reliability (kappa = 0.54–0.87),[59] many physical examination procedures used commonly in clinical practice have low to medium reliability.[60,61] This includes lung auscultation (kappa = 0.32 for bronchial breath sounds and 0.51 for wheezes)[62] and heart auscultation (31%–72% agreement among physicians).[63]

The primary research articles on the reliability of spinal palpatory procedures are difficult to compare due to variability in the palpatory tests, terminology, research design, study conditions, and statistical analysis used. The

Table 6. Quality Assessment Instrument

Criteria	Weight	Score
Study subjects		
Study subjects adequately described	1	8
Inclusion/exclusion criteria described	1	2
Subjects naive/without vested interest	1	2
No. of subjects in study given	1	4
Drop-outs described	1	1
Subjects not informed of findings	1	1
Examiners		
Selection criteria for examiners described	2	1
Background of examiners described (*e.g.,* education/clinical experience)	5	1
Examiners blind to clinical presentation of subjects	8	1
Examiners blind to previous findings	10	1
Study conditions		
Consensus on test procedures and training of examiners	4	2
Description of test/retest procedure and time interval	3	1
Study conditions described (*e.g.,* facilities and setup)	1	1
Description of palpation test technique (position of hands of examiner, etc.)	8	1
Uniform description of test outcome	5	1
Data analysis		
Appropriate statistical method used	10	1
Selection of significance level of *P* value described	8	1
Precision of examiner agreement calculated and displayed	7	1
Results		
Results displayed appropriately (*e.g.,* figures, tables)	1	1
Results adequately described	2	1
Potential study biases identified	4	1

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

Table 7. Reliability Articles Weighted Mean Quality Scores

Reliability Article [listed by author(s) (year of publication)]	Subjects (18)*	Examiners (25)*	Condition (25)*	Analysis (25)*	Results (7)*	Overall (total 100)*
Strender *et al* (1997)[48]	5.0	25.0	25.0	17.0	7.0	79.0
Schops *et al* (2000)[49]	5.5	25.0	23.5	18.0	5.5	77.5
Fjellner (1999)[44]	5.0	17.0	21.0	25.0	6.0	74.0
Rouwmaat *et al* (1998)[94]	4.0	17.0	20.5	25.0	7.0	73.5
Downey *et al* (1999)[50]	3.0	17.0	21.0	25.0	6.0	72.0
Love *et al* (1987)[45]	4.0	25.0	21.0	18.0	4.0	72.0
Johnston *et al* (1982)[42]	0.0	25.0	20.0	25.0	1.0	71.0
Ghoukassian *et al* (2001)[95]	2.5	17.0	25.0	18.0	7.0	69.5
Hsieh *et al* (2000)[47]	5.0	25.0	22.0	10.0	7.0	69.0
Lundberg *et al* (1999)[52]	2.0	17.0	24.0	18.0	7.0	68.0
Byfield *et al* (1992)[51]	3.5	25.0	14.0	18.0	7.0	67.5
Keating *et al* (1990)[46]	5.0	20.0	17.5	18.0	7.0	67.5
Johnston *et al* (1980)[41]	0.0	23.0	22.0	15.0	7.0	67.0
Maher *et al* (1994)[66]	7.5	17.0	17.0	17.5	7.0	66.0
McPartland *et al* (1997)[88]	7.0	17.0	20.0	18.0	4.0	66.0
Grant *et al* (1985)[67]	1.0	25.0	23.5	10.0	6.0	65.5
Haas *et al* (1995)[68]	7.0	25.0	19.5	10.0	3.0	64.5
Deboer *et al* (1985)[69]	1.5	25.0	13.0	18.0	7.0	64.5
Phillips *et al* (1996)[70]	5.0	23.0	10.0	18.0	7.0	63.0
Strender *et al* (1997)[53]	3.5	12.0	25.0	17.0	5.0	62.5
Hubka *et al* (1994)[89]	4.5	17.0	13.0	25.0	2.5	62.0
Mastriani *et al* (1991)[71]	6.0	25.0	23.5	0.0	7.0	61.5
Boline *et al* (1988)[72]	4.0	7.0	17.0	25.0	7.0	60.0
Inscoe *et al* (1995)[73]	6.0	17.0	21.0	10.0	5.0	59.0
Nansel *et al* (1989)[74]	4.0	22.5	18.5	10.0	3.5	58.5
Viikari-Juntura *et al* (2000)[90]	4.5	15.0	25.0	10.0	4.0	58.5
Marcotte *et al* (2001)[55]	3.0	17.0	17.0	18.0	3.0	58.0
Johnston *et al* (1982)[75]	0.0	18.0	18.0	17.5	3.0	56.5
Bergstrom (1986)[76]	1.5	25.0	22.0	0.0	7.0	55.5
Mior *et al* (1985)[13]	2.5	22.5	15.5	10.0	5.0	55.5
Mootz *et al* (1989)[77]	2.0	5.0	25.0	18.0	5.0	55.0
Johnston *et al* (1983)[96]	−1.0	18.5	20.0	13.5	3.0	54.0
Johnston *et al* (1982)[78]	−2.0	25.0	21.0	9.0	1.0	54.0
Comeaux *et al* (2001)[79]	3.5	25.0	10.0	10.0	4.0	52.5
Nice *et al* (1992)[91]	6.0	5.0	25.0	10.0	6.0	52.0
Maher *et al* (1998)[80]	1.5	17.0	9.0	17.0	7.0	51.5
Eriksson *et al* (2000)[97]	1.5	2.0	22.5	18.0	3.0	47.0
Binkley *et al* (1995)[81]	4.0	7.0	13.0	17.0	6.0	47.0
Boline *et al* (1993)[92]	6.0	2.0	10.0	18.0	7.0	43.0
Smedmark *et al* (2000)[82]	3.0	6.0	20.0	10.0	3.0	42.0
McKenzie *et al* (1997)[98]	2.5	6.0	9.0	18.0	6.0	41.5
Richter *et al* (1993)[83]	2.0	10.0	4.0	17.0	7.0	40.0
Olson *et al* (1998)[84]	3.5	5.0	13.0	10.0	6.0	37.5
Waddell *et al* (1982)[93]	5.0	7.0	5.0	18.0	2.0	37.0
Lindsay *et al* (1995)[85]	−1.0	7.0	16.0	10.0	3.0	35.0
Rhudy *et al* (1988)[86]	2.0	12.0	10.0	10.0	0.0	34.0
Van Suijlekom *et al* (2000)[87]	3.5	2.0	17.0	10.0	1.0	33.5
Johnston *et al* (1976)[11]	−0.5	6.0	21.5	0.0	3.0	30.0
McCombe *et al* (1989)[33]	2.0	5.0	1.0	10.0	7.0	25.0

Articles are ranked in order of highest to lowest overall score.
*Maximum possible score for that category.

quality scoring instrument helped to evaluate the relative value of their results. The quality assessment form can also provide a template with which future higher quality reliability studies can be designed (Tables 6 and 7).

Limitations of this review include the retrospective design, the search strategy, databases used[64], and article quality scoring. The authors conducted a retrospective review with current standards and expectations for scientific rigor that might not have been expected at the time these studies were conducted and published. Authors and indexers are not always on the same page when choosing titles and keywords.[20] Online database searches were inadequate in locating all articles that met the inclusion criteria.[20] Content expert and selective manual searches were necessary in finding many of the articles[20]. The article reviewers each had different education and training backgrounds, accounting for the initial disagreement in scoring in one third of the articles. Before reviewer consensus, there was variability in interpretation of the quality scoring instrument terms as well as in judgments regarding how well an article addressed the issues being evaluated. In using a quality assessment instrument, some quality scoring criteria are more detailed/differentiated than others, which introduces an inherent bias. Scores/assigned weights may be biased toward rigor of research methodology and presentation. Since the

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

quality assessment instrument focused on the internal validity of the studies, the quality scores cannot be extrapolated to measure the studies' significance or impact (in terms of findings, relevance to the discipline).

There are several strengths, however. The authors formed a multidisciplinary team, paying special attention to minimizing bias by the Doctor of Osteopathic Medicine and Doctor of Chiropractic on our team who did not review studies in their respective professions. The authors combined information (studies) obtained from different professions (PT, DO, DC, MD) in a systematic manner. The quality assessment instrument is comprehensive and was developed after careful consideration and discussion of prior instruments and guidelines. Reviewers were blinded to author(s) and journal, minimizing bias. Because of the current electronic search capabilities, the authors were able to survey a wider number of literature databases (13) than feasible in earlier reviews.

The findings of this comprehensive systematic review have implications for research, clinical practice, and policy. Researchers across disciplines need to incorporate more rigor in study design and presentation of results. Clinical trials using spinal palpatory diagnostic procedures need to assess the reliability and, if possible, the content validity of the procedures, which is akin to calibrating validated laboratory instruments before an experiment. Clinicians need to be cognizant that pain provocation tests are most reliable and soft tissue paraspinal palpatory diagnostic tests are not reliable. Given that spinal palpatory procedures are a cornerstone of diagnostic and therapeutic interventions across disciplines for patients with nonspecific low back and neck pain, professional societies and organizations need to enact continuing medical education programs and establish research guidelines to address the reliability of spinal palpatory procedures.[65]

■ Key Points

- A multidisciplinary team performed a comprehensive review of the primary research literature and assessed the reliability of spinal palpatory procedures used to diagnose neck or back pain.
- The majority of spinal palpatory diagnostic procedures are unreliable.
- Pain provocation tests are most reliable; soft tissue tests are not reliable.
- Regional range of motion is more reliable than segmental range of motion, and intraexaminer reliability is better than interexaminer reliability.
- Overall, examiners' discipline, experience level, consensus on procedure used, training just prior to the study, or use of symptomatic subjects does not consistently improve reliability.

Acknowledgments

The authors thank Joseph Scherger, MD, MPH, Clinical Professor, Department of Family & Preventive Medicine, UCSD for his support of the multidisciplinary team and fostering training in systematic reviews; Raymond J. Hruby, MS, DO, FAAO, and H. James Jones, DO, for reviewing articles and critiquing the manuscript; Wolfgang Gilliar, DO, for assistance in translation of the German articles; and D.V. Gokhale, PhD, and Arnold Goodman, PhD, for their statistical input.

References

1. Hartvigsen J, Christensen K, Frederiksen H. Back pain remains a common symptom in old age: a population-based study of 4486 Danish twins aged 70–102. *Eur Spine J* 2003;14:14.
2. Guo HR, Tanaka S, Halperin WE, et al. Back pain prevalence in US industry and estimates of lost workdays. *Am J Public Health* 1999;89:1029–35.
3. Deyo RA, Weinstein NJ. Low back pain. *N Engl J Med* 2001;344:363–70.
4. Narayan P, Haid R. Neurologic treatment: treatment of degenerative cervical disc disease. *Neurol Clin* 2001;19:217–29.
5. Atlas S, Deyo R. Evaluating and managing acute low back pain in the primary care setting. *J Gen Intern Med* 2001;16:120–31.
6. Carey T, Garrett J, Jackman A, et al. The outcomes and costs of care for acute low back pain among patients seen by primary care practitioners, chiropractors, and orthopedic surgeons. *N Engl J Med* 1995;333:913–7.
7. Goldstein M. *The Research Status of Spinal Manipulative Therapy. NINCDS Monograph No. 15* [DHEW Publication No. NIH 76–998]. Bethesda, MD: U.S. Department of Health, Education and Welfare; 1975.
8. Koes B, Tulder MV, Ostelo R, et al. Clinical guidelines for the management of low back pain in primary care: an international comparison. *Spine* 2001; 26:2504–14.
9. Winter G. A comparative discussion of the notion of 'validity' in qualitative and quantitative research. *The Qualitative Report*. 2000;4(3, 4) Available: http://www.nova.edu/ssss/QR/QR4-3/winter.html.
10. Haas M. The reliability of reliability. *J Manipulative Physiol Ther* 1991;14: 199–208.
11. Johnston W. Inter-examiner reliability in palpation. *J Am Osteopath Assoc* 1976;76:286–7.
12. Gonnella C, Paris SV, Kutner M. Reliability in evaluating passive intervertebral motion. *Phys Ther* 1982;62:436–44.
13. Mior S, King R, McGregor M, et al. Intra and inter-examiner reliability of motion palpation in the cervical spine. *J Can Chiropractic Assoc* 1985;29: 195–9.
14. Johnston W. Inter-examiner reliability studies spanning a gap in medical research: Louisa Burns Memorial Lecture. *J Am Osteopath Assoc* 1982;81: 43–53.
15. Beal MC, Goodridge JP, Johnston WL, et al. Inter-examiner agreement on long-term patient improvement: an exercise in research design. *J Am Osteopath Assoc* 1982;81:322–8.
16. Panzer DM. The reliability of lumbar motion palpation. *J Manipulative Physiol Ther* 1992;15:518–24.
17. Huijbregts P. Spinal motion palpation: a review of reliability studies. *J Manipulative Physiol Ther* 2002;10:24–39.
18. Hestboek L, Leboeuf-Yde C. Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic critical literature review. *J Manipulative Physiol Ther* 2000;23:258–75.
19. Van der Wurff PMW, Hagmeyer RHM. Clinical tests of the sacroiliac joint: a systematic methodological review. 1. Reliability. Manual Therapy 1999; 5:30–6.
20. Murphy LS, Reinsch S, Najm WI, et al. Spinal palpation: the challenges of information retrieval using available databases. *J Manipulative Physiol Ther* 2003;26:374–82.
21. Mulrow C, Oxman A, eds. *Cochrane Collaboration Handbook [updated September 1997]*. Update Software, Issue 4. ed. Oxford: Cochrane Library [database on disk and CDROM], 1997.
22. Irwig LTA, Gatsonis C, Lau J, et al. Guidelines for meta-analyses evaluating diagnostic tests. *Ann Intern Med* 1994;120:667–76.
23. Mulrow C, Linn W, Gaul M. Assessing quality of a diagnostic test evaluation. *J Gen Intern Med* 1989;4:288–95.
24. McConnell DG, Beal MC, Dinnar U, et al. Low agreement of findings in neuromusculoskeletal examinations by a group of osteopathic physicians using their own procedures. *J Am Osteopath Assoc* 1980;79:441–50.
25. Beal MC, Goodridge JP, Johnston WL, et al. Inter-examiner agreement on patient improvement after negotiated selection of tests. *J Am Osteopath Assoc* 1980;79:432–40.
26. Beal M, Dvorak J. Palpatory examination of the spine: a comparison of the

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

results of two methods and their relationship to visceral disease. *Manual Med* 1984;1:25–32.
27. French S, Green S, Forbes A. Reliability of chiropractic methods commonly used to detect manipulable lesions in patients with chronic low-back pain. *J Manipulative Physiol Ther* 2000;23:231–8.
28. Hawk C, Phongphua C, Bleecker J, et al. Preliminary study of the reliability of assessment procedures for indications for chiropractic adjustments of the lumbar spine. *J Manipulative Physiol Ther* 1999;22:382–9.
29. Jull G, Zlot G, Trott P, et al. Inter-examiner reliability to detect painful upper cervical joint dysfunction. *Aust J Physiother* 1997;43:125–9.
30. Tuchin P, Hart C, Johnson C, et al. Inter-examiner reliability of chiropractic evaluation for cervical spine problems: a pilot study. 1. Graduates from one institution. *Australas Chiropractic Osteopathy* 1996;5:23–9.
31. Hardy G, Napier J. Inter and Intratherapist reliability of passive accessory movement technique. *NZ J Physiother* 1991;19:22–4.
32. Leboeuf C, Gardner V, Carter A, et al. Chiropractic examination procedures: a reliability and consistency study. *J Aust Chiropractors Assoc* 1989;19: 101–4.
33. McCombe PF, Fairbank JC, Cockersole BC, et al. 1989 Volvo Award in clinical sciences: reproducibility of physical signs in low-back pain. *Spine* 1989;14:908–18.
34. NHS Centre for Reviews and Dissemination. *Undertaking Systematic Reviews of Research on Effectiveness. CRD's Guidance for Those Carrying Out or Commissioning Reviews*. CRD Report Number 4 (2nd ed). 2001 NHS Centre for Reviews and Dissemination, University of York; March 2001.
35. Cook DJ SD, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. *J Clin Epidemiol* 1995;48:167–71.
36. Deeks J. Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *Br Med J* 2001;323:157–62.
37. Juni PAD, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. *Br Med J* 2001;323:42–46.
38. Juni PWA, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054–60.
39. Shekelle PG AA, Chassin MR, Hurwitz EL, et al. Spinal manipulation for low-back pain. *Ann Intern Med* 1992;117:590–8.
40. Koes BAW, van der Heijden GJ, Bouter LM, et al. Spinal manipulation and mobilization for back and neck pain: a blinded review. *Br Med J* 1991;303: 1298–303.
41. Johnston W, Hill J, Sealey J, et al. Palpatory findings in the cervicothoracic region: variations in normotensive and hypertensive subjects. A preliminary report. *J Am Osteopath Assoc* 1980;79:300–8.
42. Johnston W, Hill J, Elkiss M, et al. Identification of stable somatic findings in hypertensive subjects by trained examiners using palpatory examination. *J Am Osteopath Assoc* 1982;81:830–6.
43. Fleiss J. *Statistical Methods for Rates and Proportions*, 2nd ed. New York: John Wiley & Sons, 1981.
44. Fjellner A, Bexander C, Faleij R, et al. Inter-examiner reliability in physical examination of the cervical spine. *J Manipulative Physiol Ther* 1999;22: 511–6.
45. Love RM, Brodeur RR. Inter-examiner and intra-examiner reliability of motion palpation for the thoracolumbar spine. *J Manipulative Physiol Ther* 1987;10:1–4.
46. Keating JC Jr, Bergmann TF, Jacobs GE, et al. Inter-examiner reliability of eight evaluative dimensions of lumbar segmental abnormality. *J Manipulative Physiol Ther* 1990;13:463–70.
47. Hsieh C-YJ, Hong C-Z, Adams AH, et al. Inter-examiner reliability of the palpation of trigger points in the trunk and lower limb muscles. *Arch Phys Med Rehabil* 2000;81:258–64.
48. Strender LE, Lundin M, Nell K. Inter-examiner reliability in physical examination of the neck. *J Manipulative Physiol Ther* 1997;20:516–20.
49. Schops P, Pfingsten M, Siebert U. Reliability of manual examination techniques at the cervical spine: study on quality assessment of manual diagnosis [in German]. *Z Orthop Grenzgeb* 2000;138:2–7.
50. Downey BJ, Taylor NF, Niere KR. Manipulative physiotherapists can reliably palpate nominated lumbar spinal levels. *Manual Ther* 1999;4:151–6.
51. Byfield D, Humphreys K. Intra- and inter-examiner reliability of bony landmark identification in the lumbar spine. *Eur J Chiropractic* 1992;72:13–17. *0*
52. 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 *0 1999;31:197–206.*
53. Strender LE, Sjoblom A, Sundell K, et al. Inter-examiner reliability in physical examination of patients with low back pain. *Spine* 1997;22:814–20.
54. Gjorup T. Reliability of diagnostic tests. *Acta Obstet Gynecol Scand* 1997; 166(suppl):9–14.
55. Marcotte J, Normand MC. Standardizing dynamic palpation in chiropractic: a reliability study for treatment of the neck area [in French]. *J Can Chiropractic Assoc* 2001;45:106–12.
56. Najm WI, Seffinger MA, Mishra SI, et al. Content validity of manual spinal palpatory exams: a systematic review. *BMC Complement Altern Med* 2003;3:1.
57. Chesworth B, MacDermid J, Roth J, et al. Movement diagram and "end-feel" reliability when measuring passive lateral rotation of the shoulder in patients with shoulder pathology. *Phys Ther* 1998;78:593–601.
58. Bellamy N, Klestov A, Muirden K, et al. College of Rheumatology classification criteria for hand, knee and hip osteoarthritis (OA): observations based on an Australian Twin Registry study of OA. *J Rheumatol* 1999;26:2654–8.
59. Lawson I, Ingman S, Masih Y, et al. Reliability of palpation of pedal pulses as ascertained by the kappa statistic. *J Am Geriatr Soc* 1980;28:300–3.
60. Koran L. The reliability of clinical methods, data and judgments. Part I. *N Engl J Med* 1975;293:642–6.
61. Koran L. The reliability of clinical methods, data and judgments. Part II. *N Engl J Med* 1975;293:695–701.
62. Spiteri M, Cook D, Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet* 1988;8590:873–5.
63. Raftery E, Holland W. Examination of the heart: an investigation into variation. *Am J Epidemiol* 1967;85:438–444.
64. Aker PD, McDermaid C, Opitz BG, et al. Searching chiropractic literature: a comparison of three computerized databases. *J Manipulative Physiol Ther* 1996;19:518–24.
65. FIMM S. Reproducibility and validity studies of diagnostic procedures in manual/musculoskeletal medicine for low back pain patients [Protocol formats]. Available at: *http://www.fimm-online.org/Home.html*.
66. Maher C, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Phys Ther* 1994;74:801–9.
67. Grant A, Spadon R. An inter- and intra-examiner reliability study, using lateral flexion motion palpation of the lumbar spine in the prone position, *Dissertation*, Anglo-European College of Chiropractic, 1985.
68. Haas M, Raphael R, Panzer D, et al. Reliability of manual end-play palpation of the thoracic spine. *Chiropractic Tech* 1995;7:120–4.
69. Deboer K, Harmon R, Tuttle C, et al. Reliability study of detection of somatic dysfunctions in the cervical spine. *J Manipulative Physiol Ther* 1985; 8:9–16.
70. Phillips DR, Twomey LT. A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure: this study was presented in part at the 8th Biennial Conference of the MPAA, in 1993. *Manual Ther* 1996;1:82–7.
71. Mastriani P, Woodman K. *Reliability of Passive Lumbar Segmental Motion*, Boston, MA: MGH Institute of Health Professions, 1991.
72. Boline P, Keating J, Brist J, et al. Inter-examiner reliability of palpatory evaluations of the lumbar spine. *Am J Chiropractic Med* 1988;1:5–11.
73. Inscoe E, Witt P, Gross M, et al. Reliability in evaluating passive intervertebral motion of the lumbar spine. *J Manual Manipulative Ther* 1995;3: 135–43.
74. Nansel DD, Peneff AL, Jansen RD, et al. Inter-examiner concordance in detecting joint-play asymmetries in the cervical spines of otherwise asymptomatic subjects. *J Manipulative Physiol Ther* 1989;12:428–33.
75. Johnston WL, Beal MC, Blum GA, et al. Passive gross motion testing: III. Examiner agreement on selected subjects. *J Am Osteopath Assoc* 1982;81: 309–13.
76. 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.
77. Mootz RD, Keating JC, Kontz HP, et al. Intraobserver and interobserver reliability of passive motion palpation of the lumbar spine. *J Manipulative Physiol Ther* 1989;12:440–5.
78. Johnston WL, Elkiss ML, Marino RV, et al. Passive gross motion testing. II. A study of inter-examiner agreement. *J Am Osteopath Assoc*1982;81: 304–8.
79. Comeaux Z, Eland D, Chila A, et al. Measurement challenges in physical diagnosis: refining interrater palpation, perception and communication. *J Bodywork Movement Ther* 2001;5:245–53.
80. Maher CG, 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;78:829–37.
81. Binkley J, Stratford PW, Gill C. Interrater reliability of lumbar accessory motion mobility testing. *Phys Ther* 1995;75:786–92; discussion 793–5.
82. Smedmark V, Wallin M, Arvidsson I. Inter-examiner reliability in assessing passive intervertebral motion of the cervical spine. *Manual Ther* 2000; 5:97–101.
83. Richter T, Lawall J. Reliability of diagnostic findings in manual medicine [in

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.

German][: Zur Zuverlassigkeit manualdiagnostischer Befunde]. *Manuelle Med* 1993;31:1–11.
84. Olson KA, Paris SV, Spohr C, et al. Radiographic assessment and reliability study of the craniovertebral sidebending test. *J Manual Manipulative Ther* 1998;6:87–96.
85. Lindsay DM, Meeuwisse WH, Mooney ME, et al. Interrater reliability of manual therapy assessment techniques. *Physiother Can* 1995;47:173–80.
86. Rhudy T, Sandefur M, Burk J. Inter-examiner intertechnique reliability in spinal subluxation assessment: a multifactorial approach. *Am J Chiropractic Med* 1988;1:111–4.
87. Van Suijlekom HA, De Vet HC, Van Den Berg SG, et al. Interobserver reliability in physical examination of the cervical spine in patients with headache. *Headache* 2000;40:581–6.
88. McPartland JM, Goodridge JP. Counterstrain and traditional osteopathic examination of the cervical spine compared. *J Bodywork Movement Ther* 1997;1:173–8.
89. Hubka MJ, Phelan SP. Inter-examiner reliability of palpation for cervical-spine tenderness. *J Manipulative Physiol Ther* 1994;17:591–5.
90. Viikari-Juntura E. Inter-examiner reliability of observations in physical examinations of the neck. *Phys Ther* 1987;67:1526–32.
91. Nice DA, Riddle DL, Lamb RL, et al. Intertester reliability of judgments of the presence of trigger points in patients with low back pain. *Arch Phys Med Rehabil* 1992;73:893–8.
92. Boline PD, Haas M, Meyer JJ, et al. Inter-examiner reliability of 8 evaluative dimensions of lumbar segmental abnormality. 2. *J Manipulative Physiol Ther* 1993;16:363–74.
93. Waddell G, Main CJ, Morris EW, et al. Normality and reliability in the clinical assessment of backache. *Br Med J* (*Clin Res Ed*) 1982;284:1519–23.
94. Rouwmaat PHM, Everaert D, Stappaerts KH, et al. Reliability of manual skinfold tests in a healthy male population. *J Manipulative Physiol Ther* 1998;21:327–32.
95. Ghoukassian M, Nicholls B, McLaughlin P. Inter-examiner reliability of the Johnson and Friedman percussion scan of the thoracic spine. *J Osteopath Med* 2001;4:15–20.
96. Johnston WL, Allan BR, Hendra JL, et al. Inter-examiner study of palpation in detecting location of spinal segmental dysfunction. *J Am Osteopath Assoc* 1983;82:839–45.
97. Eriksson E, Mokhtari M, Pourmotamed L, et al. Inter-rater reliability in a resource-oriented physiotherapeutic examination. *Physiother Theory Prac* 2000;16:95–103.
98. McKenzie AM, Taylor NF. Can physiotherapists locate lumbar spinal levels by palpation? *Physiotherapy* 1997;83:235–9.

Copyright © Lippincott Williams & Wilkins. Unauthorized reproduction of this article is prohibited.