Title

Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature.

Permalink
https://escholarship.org/uc/item/5z8886vw

Journal
Spine, 29(19)

ISSN
1528-1159

Authors
Seffinger, Michael A
Najm, Wadie I
Mishra, Shiraz I
et al.

Publication Date
2004-10-01

Peer reviewed

RELIABILITY OF SPINAL PALPATION FOR DIAGNOSIS OF BACK AND NECK PAIN: A SYSTEMATIC REVIEW OF THE LITERATURE

Michael A. Seffinger, D.O.[1], Wadie I. Najm, M.D.[2], Shiraz I. Mishra, M.D., Ph.D.[3]. Alan Adams, D.C., M.S.[4], Vivian M. Dickerson, M.D.[5], Linda S. Murphy, M.L.I.S.[6] & Sibylle Reinsch, Ph.D.[7]

[1]Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine of the Pacific, Western University of Health Sciences, 309 E. 2nd St., Pomona, CA 91766-1854 mseffinger@westernu.edu

[2]Department of Family Medicine, University of California, Irvine, Medical Center, 101 City Drive, Orange, CA 92868 winajm@uci.edu

[3]Department of Epidemiology and Preventive Medicine, School of Medicine, University of Maryland, Baltimore, MD 21201 smishra@som.umaryland.edu

[4]Office for Academic Affairs and Office of the Provost, Florida State University, 212 Westcott Building, Tallahassee, Florida 32306 aadams@mailer.fsu.edu

[5]Department of Obstetrics & Gynecology, University of California, Irvine, Medical Center, 101 City Drive, Orange, CA 92868 vdickerson@socal.rr.com

[6]Science Library Reference Department, University of California, Irvine, P.O. Box 19557, Irvine, CA 926233-9557 lmurphy@uci.edu

[7]Department of Physical Medicine & Rehabilitation, University of California, Irvine, Medical Center, 101 City Drive, Orange, CA 92868 sreinsch@uci.edu

Send correspondence to:

Michael A. Seffinger, D.O.
Assistant Professor, Dept. of OMM
College of Osteopathic Medicine of the Pacific
Western University of Health Sciences
Pomona, CA 91766-1854
Tel.: (909) 469-5634
Fax: (909) 469-5289
E-mail: mseffinger@westernu.edu

Abstract:

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.

Introduction

Health care professionals examine and diagnose patients with cervical, thoracic and lumbar back pain on a daily basis. Back pain, in fact, 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 (musculo-ligamentous) pain accounts for at least 70% of etiologies of low back pain[3]. Approximately 85% of neck pain is attributed to chronic musculo-ligamentous 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 neuro-vascular 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 regional and segmental motion, paraspinal soft tissue abnormalities and tenderness upon provocation. The ability to arrive at an accurate palpatory assessment depends mainly upon 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[11’12’13]. Several narrative reviews of the literature on spinal palpatory diagnostic procedures have been published[14’15’16’17]. 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 inter-examiner reliability of spinal palpatory diagnostic procedures"?

Materials and Methods

A multi-disciplinary 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 1-1-1966 and 10-1-2001. 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 inter-examiner reliability of manual spinal palpation; published between January 1, 1996 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 articles out of 797 were identified by our search. Upon further review and discussion, eleven of these articles were excluded for the following reasons: no separate data analysis for each procedure [12,15,24,25,26,27,28,29,30]; 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 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[21,34,35,36,37,38], including the Centre for Reviews and Dissemination (CRD) 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 2 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 inter-examiner reliability studies. Thus, the total number of studies included in the 49 articles is 53. Description of the characteristics of the studies are summarized in Table 1.

Table 1: Characteristics of studies reviewed

Characteristics	N	Percentage[ϒ]
Study type		
Inter-rater reliability	39	74
Intra- and inter-rater 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.)	9	17
Medical Doctor (M.D.), practitioner and/or student	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 (Cervico-thoracic, Thoraco-lumbar, Full spine)	11	21
Number 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
Intra-class correlation coefficient	5	9
Chi-square (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

[Υ] 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 parenthesis.

Paired reviewers initially disagreed on the quality score of 16 (33%) of the 49 articles. Quality scores of the 49 articles ranged from 25-79/100. 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 9/53 (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 67.5/100 or greater). Most (2/3) of the higher quality articles employed 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 2-5. 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.

Table 2: Quality Scores, Study Characteristics, Intra- and Interexaminer reliability for **Motion Palpation Tests**

Study	Quality score	Examiners Subjects	Type of Reliability, Spinal Motion Tests and Results	Interpretation[a]
Strender et al. (1997)[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. (2000)[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 (1999)[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. (1987)[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. (1982)[42]	71.0	3 DO; 307 (153 hypertensive) subjects	InterEx, cervical and thoracic segmental Higher level of InterEx agreement in subsample with more hypertensives (X^2 = 27.75, df = 1, $p < 0.001$)	more reliable in hypertensive subjects
Lundberg et al. (1999)[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. (1990)[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. (1980)[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, $p < .05$	more reliable in hypertensive subjects
Maher et al. (1994)[66]	66.0	6 PT; 90 Sx subjects	InterEx, lumbar segmental 13-43% agreement ICC = -0.4 -0.73	low reliability
Grant et al. (1985)[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. (1995)[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. (1985)[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. (1996)[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. (1997)[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. (1997)[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. (1991)[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. (1988)[72]	60.0	2 DC (1student); 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. (1995)[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. (1989)[74]	58.5	4 DC (1 student); 270 Asx subjects	InterEx, cervical segmental K = 0.013	low reliability

Marcotte et al. (2001)[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<.01
			InterEx: 82.3-93.2% agreement;
			K=0.57-0.85, p<.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. (1982)[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. For student #1: D (mean)=15.2, SD=2.0, p<.01; for student #2: D (mean)=13.2, SD=3.5, p<.15; for student #3: D (mean)=15.6, SD=3.5, p<.35	significant InterEx reliability for 1 of the 3 student examiners when compared with the 2 osteopathic physicians
Bergstrom (1986)[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. (1985)[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
Mootz et al. (1989)[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. (1982)[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=.0005 Sidebending: observed agreement=12, expected agreement=5, z=2.5, α=.03	Regional ROM: reliable (for rotation and sidebending)

Comeaux et al. (2001)[79]	52.5	3 DO; 54 UMS subjects	InterEx, cervical and thoracic segmental K=0.16-0.43	low to medium reliability
Maher et al. (1998) [80]	51.5	3 graduate PT students; 13 Asx subjects	InterEx, lumbar segmental ICC=0.50-0.62 (p<0.05)	Posterior-Anterior pressure test at L3 (stiffness): low reliability
Maher et al. (1998) [80]	51.5	2 PT; 27 ASx subjects	InterEx, lumbar segmental ICC=.77 (p<0.05)	Posterior-Anterior pressure test at L3 (stiffness): medium reliability; experience level, training and consensus may have improved reliability.
Binkley et al. (1995)[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 L1-5: low reliability

Smedmark et al. (2000)[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. (1993)[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 L1-2: 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 L1-2 which was medium reliability.

Olson et al. (1998)[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. (1995)[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, though 3 tests had 100% agreement (kappa not calculated with 100% agreement)

Rhudy et al. (1988)[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. (2000)[87]	33.5	2 MD; 24 Sx subjects	InterEx, cervical segmental K=0.27-0.46	low to medium reliability
Johnston et al. (1976)[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.

Abbreviations Key: PT = physical therapist; DO = doctor of osteopathic medicine; DC = doctor of chiropractic; MD = medical doctor. Sx = Symptomatic; Asx = Asymptomatic; UMS = undefined medical status; IntraEx = Intra-examiner; InterEx = Inter-examiner; K = Kappa; C = cervical; T = thoracic; L = lumbar

[a] 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 (designated as "L") reliability; 0.40-0.74= fair to good, or medium (designated as "M") reliability; 0.75-1.00= excellent or high (designated as "H") reliability[43]. 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 Intra-class Correlation Coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analysis required a case by case analysis to make the determination of degree of reliability.

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 [a]
Strender et al. (1997)[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. (2000)[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. (2000)[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. (1999)[52]	68.0	2 PT; 150 UMS subjects	InterEx, thoracic and lumbar digital pressure L4-5: K=0.71; L5-S1: K=0.67	L4-5 and L5-S1: medium reliability (Data for thoracic and other lumbar segments not reported)
Keating et al. (1990)[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; L4-5 and L5-S1 had greater concordance for osseous pain (mean K > 0.6); No significant difference between Sx vs. ASx subjects
Maher et al. (1994)[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. (1997)[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. (1997) [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. (1985)[69]	64.5	3 DC; 40 ASx subjects	IntraEx and InterEx, cervical digital pressure IntraEx: C1-3: 55-80% agreement, Kw = 0.3-0.56; C4-7: 60-68% agreement, Kw = 0.2-0.43; InterEx: C1-3: 43-66% agreement, Kw = 0.08-0.48; C4-7: 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 C1-3 than C4-7.
Strender et al. (1997)[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. (1997)[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 100% 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. (1994) [89]	62.0	2 DC; 30 Sx subjects	InterEx, cervical digital pressure 76.6% agreement; K=0.68	medium reliability

Boline et al. (1988)[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=-.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. (1987) [90]	58.5	1 MD and 1 PT; 52 Sx subjects	InterEx, cervical (C5-8) 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-.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. (1993) [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. (1993) [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
Richter et al. (1993)[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. (1982) [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. (2000)[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 (1989)[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 (1989)[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

Abbreviations Key: PT = physical therapist; DO = doctor of osteopathic medicine; DC = doctor of chiropractic; MD = medical doctor. Sx = Symptomatic; Asx = Asymptomatic; UMS = undefined medical status; IntraEx = Intra-examiner; InterEx = Inter-examiner; K = Kappa; C = cervical; T = thoracic; L = lumbar; S = sacral; SCM = sternocleidomastoid muscle

[a] 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 (designated as "L") reliability; 0.40-0.74= fair to good, or medium (designated as "M") reliability; 0.75-1.00= excellent or high (designated as "H") reliability[43]. 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 Intra-class Correlation Coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analysis required a case by case analysis to make the determination of degree of reliability.

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

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[a]
Strender et al. (1997)[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. (2000)[49]	77.5	5 MD; 20 Sx subjects	InterEx, cervical and thoracic paraspinal soft tissue tone K=0.2-0.4	low to medium reliability
Rouwmaat et al. (1998)[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. (2001)[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. (2000)[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=-.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. (1990)[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. (1985)[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. (1988)[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. (1987)[90]	58.5	1 MD, 1 PT; 52 Sx subjects	InterEx, cervical paraspinal muscle tone K = 0.4	medium reliability
Johnston et al. (1983)[96]	54.0	6 DO (5 students); 30 UMS subjects T	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. (2001)[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. (2000)[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

Abbreviations Key: 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 = Intra-examiner; InterEx = Inter-examiner; K = Kappa; C = cervical; T = thoracic; L = lumbar

[a] 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 (designated as "L") reliability; 0.40-0.74= fair to good, or medium (designated as "M") reliability; 0.75-1.00= excellent or high (designated as "H") reliability[43]. 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 Intra-class Correlation Coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analysis required a case by case analysis to make the determination of degree of reliability.

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[a]**

Downey et al. (1999)[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. (1992)[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. (1990) [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. (1995)[81]	47.0	6 PT;
		18 Sx subjects
			InterEx, lumbar identification of a marked spinal segment
			K= 0.3
			ICC= O.69 (95% C.I.=0.53-0.82)
				low reliability

McKenzie et al. (1997) [98]	41.5	17 PT; 10 ASx subjects	Intra- 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.

Abbreviations Key: PT = physical therapist; DO = doctor of osteopathic medicine; DC = doctor of chiropractic; MD = medical doctor. Sx = Symptomatic; Asx = Asymptomatic; UMS = undefined medical status; IntraEx = Intra-examiner; InterEx = Inter-examiner; K = Kappa; C = cervical; T = thoracic; L = lumbar

[a] 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 (designated as "L") reliability; 0.40-0.74= fair to good, or medium (designated as "M") reliability; 0.75-1.00= excellent or high (designated as "H") reliability[43]. 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 Intra-class Correlation Coefficient, 70% or greater or 0.75 or greater, respectively, was required to determine reliability. The other types of analysis required a case by case analysis to make the determination of degree of reliability.

Using EBM to Answer CAM Questions and How to Teach It: The majority of spinal palpatory diagnostic tests demonstrated low reliability. Data from the higher quality studies (quality score 67.5/100 or greater) showed acceptable reliability for the following spinal palpatory diagnostic procedures: 1) **inter-examiner** regional range of motion of the cervical spine[44]; 2) **intra-examiner** thoracic and lumbar segmental vertebral motion tests[45]; 3) **inter-examiner** 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 inter-examiner cervical, thoracic and lumbar segmental vertebral motion tests. One study showed 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 1 study compared the reliability of examiners from one discipline with the reliability of examiners from a different discipline (2 physical therapists vs. 2 medical doctors) using the same tests[53]. Although physical therapists were more reliable than physicians in employing segmental vertebral motion tests, they were otherwise comparable in terms of reliability of other tests.

There are informative trends noticeable amongst the higher quality quartile studies that utilized the same statistical analysis. In those studies that used kappa statistics, a higher percentage of the pain provocation studies (7/11; 64%) demonstrated acceptable reliability followed by motion studies (7/12; 58%), landmark studies (1/3; 33%) and soft tissue studies (0/11; 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, intra-examiner reliability was better than inter-examiner reliability.

Paraspinal soft tissue palpatory tests had low inter-examiner 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 prior to 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 inter-examiner 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 intra-reliability 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 (K=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 intra-examiner reliability is better than inter-examiner 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[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 procedures. 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 exam, 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 K= 0.62-0.76)[57]. However, the smaller joints of the extremities, like the vertebral spine, are less reliable (i.e., K=0.24-0.60 amongst 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 methodological challenges. Examiner bias and inconsistency create variability in procedures. Although palpation for pedal pulses has medium to high reliability (Kw=0.58-0.87)[59], many physical exam procedures used commonly in clinical practice have low to medium reliability [60’61]. This includes lung auscultation (K=0.32 for bronchial breath sounds and 0.51 for wheezes)[62] and heart auscultation (44-72% 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 utilized. The 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 (see tables 6 and 7 below).

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	

Number 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 & 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	

Table 7: Reliability Articles Weighted Mean Quality Scores

Using EBM to Answer CAM Questions and How to Teach It:

Reliability Article Listed by author(s)(year of publication)[reference]	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

* maximum possible score for that category

Limitations of this review include the retrospective design, the search strategy, databases used 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 1/3 of the articles. Prior to 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 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 multi-disciplinary 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 (P.T., D.O., D.C., M.D.) 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. Due to 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 utilizing 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[64].

References