Spinal Motion Palpation: A Review of Reliability Studies

Peter A. Huijbregts, DPT, OCS, FAAOMPT

Abstract: Spinal motion palpation is a diagnostic tool used by a number of professions. Research studies available on the reliability of motion palpation studies have statistical and methodological flaws affecting their statistical conclusion validity, external validity, and construct validity. Further research is necessary to determine the reliability of motion palpation techniques bearing in mind these statistical and methodological flaws.

Key Words: Spinal Motion Palpation, Reliability, Research Validity

Spinal motion palpation is a diagnostic tool used by physical therapists, chiropractors, osteopaths, and medical doctors[1]. Table 1 lists the different types of spinal motion palpation[1,2]. Four assumptions form the rationale for the use of motion palpation as a diagnostic tool[3-5]:

1. Spinal segmental motion abnormalities cause or contribute to functional limitation and disability.
2. Motion palpation is a reliable indicator of these motion abnormalities.
3. Motion palpation is a valid indicator of these abnormalities.
4. Motion palpation is sensitive to clinically important changes in these motion abnormalities.

The goal of this article is to provide the clinician with a critical analysis of the research into the intra- and interrater reliability of spinal motion palpation. Intrarater reliability refers to the stability of measurements taken by one rater across two or more trials; interrater reliability is concerned with the level of agreement between findings of two or more raters measuring the same group of subjects[6].

Address all correspondence and request for reprints to:
Dr. Peter A. Huijbregts
Olympic Physiotherapy
#245-3066 Shelbourne Street
Victoria, BC V8R 6T9 Canada
huijbregts_peter@hotmail.com

Method

The MEDLINE and CINAHL databases and the computerized index to the holdings of Western States Chiropractic College (WSCC) were searched for the period 1980-2000 using the keywords *motion palpation, accessory motion,* and *intervertebral motion*. This was followed by a hand search of the reference lists of the retrieved articles. Studies have been included in this review if they reported on original research related to reliability of diagnostic motion palpation of the cervical, thoracic, or lumbar spine. This article discusses the research validity of reliability studies, presents all studies retrieved, discusses research validity specific to these studies, and concludes with clinical implications and suggestions for further research.

Research Validity

Domholdt[7] defined research validity as the extent to

The Journal of Manual & Manipulative Therapy
Vol. 10 No. 1 (2002), 24 - 39

Table 1. Types of spinal motion palpation.

Active motion palpation[1]	• Assessment technique in which the clinician palpates bony landmarks while guiding the patient through cardinal plane motions of the trunk
Passive motion palpation (PIVM)[2]	
• PPIVM	• Assessment technique whereby one vertebra is moved in physiological ranges on another
• PAIVM	• Assessment technique whereby segmental mobility is assessed through the translatory motions associated with physiological motions

PIVM = passive intervertebral motion

PPIVM = passive physiological intervertebral motion palpation

PAIVM = passive accessory intervertebral motion palpation

which the conclusions of a study are believable and useful. Discussed here are three different areas in which the research validity of reliability studies can be threatened: statistical conclusion validity, external validity, and construct validity.

Statistical Conclusion Validity

Using inappropriate statistical tools for data analysis is a threat to statistical conclusion validity[7]. Reliability studies quantitatively express the level of reliability by way of an index of agreement. The simplest index of agreement is the *percentage agreement value*[6], which is defined as the ratio of the number of agreements to the total number of ratings made[8]. It is most commonly used for nominal and ordinal scale data but can also be used with higher scale data. Because it does not correct for chance agreement, it may provide a misleadingly high estimate of reliability[6,8,9].

The *kappa statistic* (κ) is a chance-corrected index of agreement for use with nominal and ordinal data[6,8]. When used with higher scale data, it tends to underestimate reliability[6]. All variations of the κ statistic are inappropriate for use as a reliability statistic when there is limited variation in the data set. Limited variation occurs when there is a large proportion of agreement or when most agreement is limited to one of the possible rating categories[8]. This can be the result of a study population that is highly homogenous on the variable of interest; it can also occur as a result of rater bias or when the raters use only a limited portion of a multi-point rating scale[6,10]. High percent agreement but low κ values are indicative of limited variation[8]. Use of the κ statistic with small samples can also lead to misleading results[6]. Lantz[10] suggested that interpretation of κ is not possible in the absence of the percent agreement values or the contingency tables from which it was derived. Theoretically, κ can be negative if agreement is worse than chance. Practically, in clinical reliability studies, κ usually varies between 0.00 and 1.00[6]. Table 2 contains benchmark values used for interpretation of κ values[5,6].

Some of the rating scales used in motion palpation reliability studies are multi-point scales: one end of the scale represents hypomobility while the other end represents hypermobility. A single-point disagreement between raters may not have great implications for patient management, e.g., if both raters perceive the presence of a hypomobility, albeit of a different magnitude, both will decide to use mobilization. Conversely, if both raters perceive the presence of a hypermobility, albeit again of a slightly different magnitude, both will likely decide to use stabilization. However, greater point differences may put the raters on opposite sides of the scale midpoint of normal mobility, and this will have important implications for their patient management choices. The κ statistic does not differentiate among disagreements; it assumes that all disagreements are of equal clinical importance[6]. The *weighted kappa statistic* (κ_w) is a modification of κ for use with ordinal level data; by assigning different weights to the different cells used to calculate k, the researcher accounts for the relative seriousness of disagreements[6,8]. Interpretation of a study using κ_w is only possible when the study provides data on the assignment and value of the weights[10].

The κ and κ_w statistics are chance-corrected indices of agreement for use with two ratings or two raters. With more than two ratings or raters, researchers may choose to use the mathematical mean of multiple κ statistics to represent reliability. However, combining κ statistics to calculate a *mean kappa* (κ_m) is allowed only if standard errors of κ are similar in magnitude[8].

Another variation of κ that can be used to evaluate reliability in the case of more than two raters is the *generalized kappa statistic* (κ_g). A κ_g is the weighted average of pair-wise κ's with lower weights assigned to rater-pairs where the expected agreement based on chance is high[11]. As with κ_w, interpretation of a study using κ_g requires data regarding the weights assigned.

Another chance-corrected index of agreement is *Scott's π*: This is used to determine the percent agreement beyond that expected to occur by chance. Scott's π is the ratio of the actual difference between obtained and chance agreement[12].

The *Pearson product-moment correlation coefficient* (*r*) quantitatively describes the strength and direction of the relationship between two variables. It is designed for use with continuous data with underlying normal distributions on an interval or ratio scale. The *Spearman rank correlation coefficient* and *Phi coefficient* are correlation coefficients intended for use with ordinal and nominal data, respectively. A limitation of correlation coefficients as indices of agreement is that they are designed to assess only bivariate relationships, i.e., two ratings or two raters[6]. In fact, correlation coefficients are not really appropriate as an index of agreement as they do not reflect agreement, but rather they are a measure of covariance. They express the degree to which two variables vary in similar patterns[5,6,9]. Despite low actual agreement, a consistent difference between ratings will produce a large value for *r*, giving the misleading impression of high reliability. Correlation coefficients vary from –1.00 indicating a perfect negative correlation to 1.00 indicating a perfect positive correlation; a value of 0.00 indicates total absence of correlation[6]. Table 3 contains benchmarĸ values for interpretation of *r* values[6].

The *intraclass correlation coefficient* (ICC) is a reliability coefficient calculated with variance estimates obtained through an *analysis of variance* (ANOVA). It can be used for two or more raters or ratings, and it does not require the same number of raters per subject. Although designed for interval or ratio scale data, it can also be used for ordinal scale data, provided the intervals between the ratings are assumed to be equivalent[6]. Portney and Watkins[6] described six different equations for ICC calculation. ICC (1,1) designates the equation used when each subject is assessed by a different set of two or more raters, randomly chosen from a larger population of raters, and when the ratings used are single ratings and not the mean of several measurements. ICC (2,1) designates the equation used when each rater assesses each subject; raters are again randomly chosen and the ratings used are single ratings. Because the choice of ICC used affects the numerical value of ICC with the same data set used, the type of ICC used should be reported in research studies[6]. Limited variation within the data set also makes the ICC an unreliable indicator of reliability[6,8]. In case of limited variation, ICC can exceed 1.00, but normally ICC varies between 0.00 and 1.00. Portney and Watkins[6] provided benchmark values for using ICC in reliability studies (Table 4).

Table 2. κ benchmark values[6].

< 40%	Poor to fair agreement
40-60%	Moderate agreement
60-80%	Substantial agreement
> 80%	Excellent agreement
100%	Perfect agreement

Table 3. *r* benchmark values for health sciences[6].

0.00-0.25	Little or no relationship
0.25-0.50	Fair relationship
0.50-0.75	Moderate to good relationship
> 0.75	Good to excellent relationship

Table 4. ICC benchmark values[6].

< 0.75	Poor to moderate agreement
> 0.75	Good agreement
> 0.90	Reasonable agreement for clinical measurements

The standard error of measurement (SEM) is the standard deviation of the distribution of measurement results on one subject; in the case of rater reliability, it reflects the expected error in the scores of the different raters. It is expressed in the same units as the original measure[6,13]. SEM is often estimated using the standard deviation of measurements and the correlation coefficient of these measurements[7].

Some reliability studies report tests of clinical significance. A *z-score* is a statistic expressed in terms of standard deviation units; calculating a z-score assumes a normal distribution of ratio scale data[6]. A *Chi-square* (χ^2) test is a non-parametric test of significance for use with nominal or ordinal scale data. It cannot distinguish a significant relationship predominated by agreement from one predominated by disagreement: deviation from chance in either direction contributes to the magnitude of χ^2 [8]. Showing that κ significantly differs from zero is of little value: large samples tend to produce small, yet significant κ values, whereas small samples may cause even large κ values to be statistically insignificant[8]. Sample size also affects significance of *r*: large samples produce statistical significance despite a low *r*[6]. We discussed above how k, *r*, and ICC values are generally compared to benchmark values rather than being tested for significance[5,6].

External Validity

External validity deals with the degree to which study results can be generalized to different subjects, settings, and times[7]. Similarity in subjects, raters, motion palpation technique, rating scale, and setting allows for a greater degree of generalization of motion palpation reliability studies to the clinical setting.

As mentioned above, one of the assumptions underlying the use of diagnostic spinal motion palpation is that segmental motion abnormalities are (partly) responsible for functional limitations and disabilities; lack of symptoms would imply absence of motion abnormalities. If this assumption is true, then the results of motion palpation studies in asymptomatic subjects cannot be generalized to a patient population. Patient body type greatly affects the clinician's ability to reliably palpate relative motion of bony landmarks; study results should only be generalized to subjects with a similar body type[12]. Other patient characteristics, e.g., gender, age, and medical history, should also be matched to maximize external validity.

The experience level of the rater may affect reliability. However, the nature of this relationship is unknown. One might assume that increased practice increases skill level and thus reliability. However, Mior et al[14] found higher interrater reliability in students versus experienced clinicians for sacroiliac motion palpation tests. Thus, idiosyncratic behavior might negatively affect reliability in experienced clinicians[14,15]. External validity is greatest for studies in which rater experience and skill level are comparable to that of the rater in the clinical setting. This leaves clinicians and researchers with the challenge of defining experience and skill level. Technique and method of rating used may also depend on the professional training of the rater; e.g., results of a study using chiropractors may be more easily generalized to a chiropractor using motion palpation techniques.

Table 1 lists the different types of spinal motion palpation. Spinal motion palpation techniques have other parameters in addition to being active, passive, physiological, or accessory. Lee and Svensson[16] showed that an increased loading rate decreased the amount of multi-level spinal displacement occurring as a result of a postero-anterior pressure (PA) test; this should logically affect the perceived level of PA stiffness. Viner and Lee[17] compared the direction of applied force during L1-S1 PA tests; they stated that the interrater variation in the direction of force might account for differences in stiffness, e.g., 10% or more at L3. Maher and Adams[18] found that the grip used for a central PA test affected the perceived magnitude of stiffness stimuli; though equally sensitive to changes in physical stiffness, the thumb grip method made the stimuli appear stiffer than the pisiform grip method. Maher and Adams[19] found that visual occlusion did not affect the ability to discriminate between stiffness stimuli, but that the absence of visual feedback caused stiffness to be judged as significantly higher. Edmonston et al[20] found that stiffness during an L5 PA was significantly greater in flexion than in extension as well as significantly greater in extension than in a neutral position; at L3, PA stiffness was significantly greater in flexion than in neutral. Maher et al[21] found that a padded surface significantly reduced stiffness parameters during PA testing when compared to a rigid surface. In a separate article, Maher[22] mentioned that stiffness perceived during PA testing will be affected by series and contrast effects and by the use of reference stimuli. Table 5 summarizes the variables affecting perceived stiffness during PA tests. Closely matching all motion palpation variables used in the study to those used in the clinical situation should maximize external validity.

The rating scales used in reliability studies vary widely. Some studies used a dichotomous nominal level rating scale wherein raters are asked to indicate absence or presence of a *"joint fixation"*. Others used an ordinal level multi-point scale, varying from 3- to 13-points. Maher et al[23] used known reference stimuli for all 11 points on their rating scale in part of their study, effectively transforming their scale into a ratio level scale. Ratings on a scale can be related to mobility, but they can also be related to presence versus absence or magnitude of pain. The discussion of the importance of verbal feedback from the patient regarding pain during motion palpation testing has yet to be resolved[5]. Matching a rating to a dichotomous or numerical value on the rating scale used requires a clear definition of the different values of the scale. External validity increases if both rating scale and definition of this scale are similar to those used clinically.

Motion palpation studies are often done in a highly controlled research setting; this may affect generalization to the true clinical setting with many confounding variables.

Construct Validity

A construct is an artificial framework that is not directly observable[7]. The main threat to construct validity in reliability research is the discrepancy between the construct as labeled and the construct as implemented[7].

Table 5. Factors affecting perceived PA stiffness.

- Loading frequency
- Direction of force
- Type of grip used (pisiform or thumb grip)
- Visual feedback
- Patient position
- Plinth padding
- Series and contrast effects
- Use of reference stimuli

These two concepts are illustrated with examples in the discussion section below.

Reliability Studies

Raters, motion palpation technique, study protocol, rating scale, subjects, and statistical analysis for the reliability studies are discussed below per spinal region. Table 6 contains intra- and interrater reliability results from studies on the cervical and cervicothoracic spine, Table 7 contains the results from thoracic and thoracolumbar studies, and Table 8 contains results from lumbar spine studies.

Cervical Spine

Mior et al[24] studied two blindfolded senior chiropractic student raters using supine C1-C2 rotation (ROT) and sidebending (SB) PPIVM tests after three months of specialized instruction. The rating scale was dichotomous: absence or presence of a fixation was defined as a loss of joint play with a hard endfeel. Subjects were 59 asymptomatic chiropractic students. Data analysis was done with percent agreement and κ values.

DeBoer et al[25] reported on three chiropractors using seated motion palpation for C1-T1 flexion (FL), extension (EXT), ROT, and SB. A 3-point rating scale was used:

Table 6. Reliability studies of cervical and cervicothoracic motion palpation.

Authors	Intrarater reliability	Interrater reliability
Mior et al (1985)	• 71% (κ=0.37) • 79% (κ=0.52)	• 61% (κ=0.15)
De Boer et al[25]	• C1-C3: 63-70% (κ_w=0.43-0.76) • C3-C6: 45-50% (κ_w=0.01-0.20) • C6-T1: 58-75% (κ_w=0.36-0.45)	• C1-C3: 21-56% (κ_w=-0.03-0.23) • C3-C6: 25-36% (κ_w= 0.01-0.05) • C6-T1: 44-58% (κ_w= 0.40-0.45)
Bronemo & Van Steveninck[26]	Average intrarater agreement 88.2-94.7%	Average interrater agreement in sitting 84.4%, in supine 84.8%
Nansel et al[27]		Irrespective of rater order or severity indicated by first rater, agreement rates near 50% (κ=0.013)
Schoensee et al[28]	Asymptomatic subjects: • PAIVM κ=0.81 • PPIVM κ=0.72	• Asymptomatic subjects: PAIVM κ=0.45, PPIVM κ=0.38 • Patients: PAIVM κ=0.79, PPIVM κ=0.52
Jull et al[29]		• κ=0.80 (2 rater pairs) to 1.00 (6 pairs) on inclusion in trial • C0-C1: κ=0.34 to 0.78 (ranking joints on magnitude of restriction) • C1-C2: κ=0.37-1.00 • C2-C3: κ=0.25-0.78
Schoeps et al[30]		• κ=0.03-0.44 (mobility) • κ=0.09-0.59 (pain)
Smedmark et al[31]		• C1-C2: 87% (κ=0.28) • C2-C3: 70% (κ=0.43) • C7-T1: 79% (κ=0.36)
Smith et al[32]	• Agreement 51.9-100.0% (κ=0.291-1.00) • Mean agreement per segment 71.6-80.3% (κ_m=0.572 for T1-T2 to 0.672 for T2-T3)	• Pair-wise agreement 33.3-92.6% (κ=-0.057 to 0.602) • Combined per segment κ_m=0.118 (T2-T3) to 0.239 (C7-T1)

absent, slight, or obvious fixation in any direction. The subjects were 40 asymptomatic male chiropractic students (21-44 years old). Pair-wise percent agreement and κ_w values were calculated for three condensed regions; details regarding the assignment of weights were not provided. The κ_w values were analyzed for significance with a z-score. All κ_w values for C6-T1, but none for C3-C6, were significant at $P<0.05$; for C1-C3, all intrarater κ_w values and one interrater pair-wise κ_w reached significance at $P<0.05$.

Bronemo and Van Steveninck[26] reported on two blindfolded senior chiropractic students using seated and supine C2-C7 PPIVM in an oblique-posterior-lateral direction. The rating scale was dichotomous: absence or presence of fixation defined as decreased joint play with a hard endfeel. Subjects were 102 chiropractic students for the interrater, 34 for the intrarater study. Agreement, defined as agreement on absence or presence of fixation, was expressed with percent agreement values. Raters agreed mainly on the absence of fixation: 70.8% of agreements in sitting and 76.8% in supine. The lowest interrater agreement was at C2-C3 with a progressive increase of agreement towards the lower cervical segments.

Nansel et al[27] reported on two rater pairs (three chiropractors and one chiropractic student) using mid- to low-cervical SB PPIVM after some practice sessions. The rating scale was dichotomous: left or right. Asymptomatic male and female chiropractic students (25-45 years old) served as subjects. The first of a rater pair palpated a subject to identify an obvious side-to-side endfeel difference and graded its severity on a 3-point scale; the second rater then indicated the side of greatest restriction on this marked segment. The order of the raters was alternated. One pair tested 76 subjects in sitting; the other 88 in supine. Interrater agreement was analyzed with percent agreement and κ values (for pooled data from seated and supine tests). A z-score was used to determine significance of the percent agreement values; all z-scores failed to reach significance at $P=0.05$.

Schoensee et al[28] reported on two physical therapists using prone central C2-C3 and unilateral C1-C3 PA; supine upper cervical FL, EXT, and SB PPIVM; and seated C1-C2 ROT PPIVM. Tests were rated on a 3-point scale. Normal on PAIVM testing was defined as free movement in the normal range of motion (ROM), limited as some resistance to movement but movement through partial ROM, and severely limited as minimal to no mobility with immediate resistance. The PPIVM scale defined normal as a good chin tuck, 15° EXT, 25-35° SB and 45° ROT. Limited was defined by a 15-75% restriction in ROM, severely limited by restriction > 75%. Ten asymptomatic subjects were examined 2-3 days apart for intrarater reliability data; two immediately consecutive examinations with a varied order of raters established interrater data. A second study reported interrater data on five cervicogenic headache patients. Results were analyzed with κ.

Jull et al[29] studied seven manipulative physical therapists blinded to subject symptom status using manual examination of C0-C3 that was not restricted to specific techniques. One rating scale was dichotomous: presence or absence of joint dysfunction, sufficient to include the patient in a trial for painful upper cervical dysfunction. The raters also recorded and ranked the upper cervical joints in magnitude of restriction based

Table 7. Reliability studies of thoracic and thoracolumbar motion palpation.

Authors	Intrarater reliability	Interrater reliability
Loram[33]	• In sitting mean agreement 93.3+/-7.2% • In prone mean agreement 95+/-6.3%	
Haas et al[4]	• κ=0.43 • κ=0.55 (both based on partial repeated measures) subjects without restriction)	• κ=0.14 (segmental level and direction of restriction) • κ=0.19 (on level alone) • κ=0.35 (for identifying
Love & Brodeur[34]	r = 0.302-0.684	R = 0.023-0.085
Keating et al[35]		Segmental agreement: • Passive motion palpation κ_m=-0.03 to 0.23 • Active motion palpation κ_m=0.00-0.25

Table 8. Reliability studies of lumbar motion palpation.

Authors	Intrarater reliability	Interrater reliability
Gonella et al[15]	Reasonable to good	No interrater reliability
Larsson[36]	• Average agreement in first 3 ratings: 78.6% (based on level and direction) and 66.6% (on level alone) • Average agreement in all 5 ratings: 69.0% (level and direction) and 53.3% (level alone)	• 56.4% (level and direction) • 34.3% (level alone)
Grant & Spadon[37]	85-90% (on level and direction)	• Segmental agreement: 60.8% (L1-L2); 73.3% (L2-L3); 80.8% (L3-L4); 65.0% (L4-L5); 52.5% (L5-S1) • Agreement between 3 raters: 71.5-74.0% • Agreement between 2 raters: 79.5-84.7%
Bergstroem & Courtis[38]	Mean intrarater agreement: • 95.4+/-3.2% • 96.0+/-3.2%	• Mean agreement on level and direction: 81.8+/-4.6% • Segmental agreement on level and direction: 79% (L1-L2); 80.5% (L2-L3); 86% (L3-L4); 88% (L4-L5); 75.5% (L5-S1) • Mean agreement on level alone: 74.4+/-6.2% • Segmental agreement on level alone: 77% (L1-L2); 72% (L2-L3); 77% (L3-L4); 81% (L4-L5); 65% (L5-S1)
Jull & Bullock[39]	Perfect agreement 87.5% (r=0.81-0.91)	Perfect agreement 86% (r=0.82-0.94)
Boline et al[40]		• Agreement on fixation 60-90% (κ=0.05-0.31) • Agreement on muscle spasm 65-70% (κ=0.10-0.31) • Agreement on pain with motion palpation 90-96% (κ=0.03-0.49) • Combined scores: 44-70%; κ=0.00-0.26; κ_w=0.08-0.33; r=-0.02-0.38
Mootz et al[41]	Segmental reliability: • Rater #1: κ=0.05-0.39 • Rater #2: κ=-0.09-0.48 Collapsed segments (L2-L4; L4-S1): • Rater #1: κ=0.26; κ=0.46 • Rater #2: κ=-0.11; κ=0.41	• Interrater κ=-0.17-0.17 • Interrater for collapsed segments $\kappa \leq 0.17$
Leboeuf et al[42]	• Perfect agreement ≈ 50% • Perfect and partial agreement ≈ 90% (≈ = approximately)	• Perfect agreement at first visit ≈ 20%; at fifth visit ≈45% • Perfect and partial agreement at first visit ≈ 85%; at fifth visit ≈ 100%

Table 8. (continued)

Authors	Intrarater reliability	Interrater reliability
Richter & Lawall[43]	Intrarater κ's on average 0.3-0.8 higher than interrater values	• Total agreement PPIVM: FL κ=0.18-0.33; EXT κ=0.14-0.36; left SB κ=0.12-0.72; right SB κ=0.08-0.47; left ROT κ=0.22-0.29; right ROT κ=0.09-0.29 • PA: κ=0.08-0.18 (mobility); κ=0.21-0.55 (pain)
Phillips & Twomey[44]		• PPIVM FL 62-97% (κ=0.00-0.30) • PPIVM EXT 62-95% (κ=0.04-0.22) • Unilateral PA 40-97% (κ=0.31-1.00) • Central PA 30-100% (κ=0.16-0.87) • Transverse pressure test 33-100% (κ=0.19-0.93)
Maher & Adams[9]		• Pain: 31-43%; ICC (1,1) = 0.67-0.72 • Stiffness: 21-29%; ICC (1,1) = 0.03-0.37
Binkley et al[45]		• Agreement on mobility ratings for marked segment: 29%; κ_g = 0.09; ICC (1,1)=0.25; SEM=1.2 • Agreement on decision to treat marked level κ=0.09
Inscoe et al[12]	Mean agreement: • 66.67% (π=41.89%) • 75% (π=61.29%)	Mean agreement: 48.61% (π=18.35%)
Phillips & Twomey[46]		• Mobility: PPIVM FL 55-98% (κ_w=-0.11-0.32); PPIVM EXT 61-99% (κ_w=-0.02-0.23); unilateral PA 81-99% (κ_w=-0.10-0.11); central PA 74-99% (κ_w=-0.14-0.24); transverse pressure test 76-100% (κ_w=-0.15-0.23) • Tissue response: unilateral PA 43-99% (κ=-0.09-0.28); central PA 60-99% (κ=-0.15-0.19); transverse pressure test 51-100% (κ=-0.16-0.22)
Maher et al[23]		• Study #1: ICC (2,1)=0.50-0.62; SEM=1.35-1.58 • Study #2: ICC (2,1)=0.77; SEM 0.72

on an overall rating of joint motion and subject-reported pain with manual examination. The study used 40 subjects with or without neck pain and headache (20-58 years old); the order of testing was varied. Pair-wise agreement was expressed with κ. There was no interrrater disagreement on the left-right decision, so results for the two joints of one motion segment were summated for the second interrater study. The authors noted that limited variation in the data set affected κ values in this portion of the study.

Schoeps et al[30] reported on five medical doctors using C0-C1 SB and ROT, C1-C2 FL-ROT, C2-C3 SB-ROT, undescribed C3-C6 segmental tests, and C6-T1 segmental ROT. The raters used two dichotomous rating scales: one for absence or presence of hypomobility as compared to the contralateral motion, the other for absence or presence of pain during motion palpation testing. The subjects were 20 asymptomatic volunteers (ten male, ten female; 20-49 years old) and 20 patients with cervical region complaints (eight male, 12 female; 21-55 years old). Volunteers and patients were selected at random from a larger group. All five raters evaluated all subjects; the order of testing was randomized and the raters were blind to subject symptom status. Agreement was expressed with κ values.

Smedmark et al[31] studied two manipulative physical therapists using seated C1-C2 ROT, supine C2-C3 SB, and sidelying C7-T1 FL/EXT PPIVM after a pilot study for standardization. The rating scale was dichotomous: hypomobility versus normo-/hypermobility. The C1-C3 tests were positive if ROM right was smaller than left with a stiff endfeel; the C7-T1 test was positive if judged stiff when compared to adjacent levels. The subjects were 61 patients (21-70 years old) with non-specific neck problems. Rater order was random. Data were analyzed with percent agreement, κ, and percentage of positive findings. The percentage of positive findings were 3%, 32%, and 69%, respectively.

Cervicothoracic Spine

Smith et al[32] reported on three physical therapists using C6-T4 FL PPIVM. The rating scale was a 7-point scale; the raters tended to use only 4 points. The subjects were 27 patients with upper-quarter disorders. The results were analyzed with pair-wise percent agreement, κ, and κ_m values.

Thoracic Spine

Loram[33] reported on a senior chiropractic student using seated active motion palpation of thoracic FL and EXT, and prone central PA with hypothenar eminence contact. The rating scale was dichotomous: absence or presence of a fixation defined as decreased approximation and separation as well as a loss of joint play with active motion palpation; decreased "springiness" on PAIVM. Agreement, defined as agreement on the absence or presence of a fixation, was expressed with percent agreement values. Ten chiropractic students were subjects.

Haas et al[4] studied two chiropractors using seated T3-L1 ROT PPIVM after practice sessions. The rating scale was dichotomous: presence or absence of a hard, restricted endplay sufficient to indicate manipulative treatment. The subjects were 73 freshman chiropractic students with and without symptoms; spinous processes were marked. An ANOVA failed to show significance for the effects of rater, repeated examinations, upper or lower region, and the presence and absence of symptoms. Pooled data for multiple tests on each patient were analyzed to determine κ because of the limited number of positive findings at any segment.

Thoracolumbar Spine

Love and Brodeur[34] reported on eight senior chiropractic students asked to choose the most hypomobile segment with a seated T1-S1 active motion palpation scan. Subjects were 32 male chiropractic students (22-34 years old) with their spinous processes marked. The authors used r to determine intrarater reliability and the index of association R for pair-wise interrater reliability; χ^2-analysis determined significance of these values. Six of the eight raters had significant intrarater reliability at $P<0.05$; no R was significant.

Keating et al[35] studied three chiropractors using seated active and passive motion palpation of T11-S1 after practice sessions. The rating scale was dichotomous: absence or presence of fixations defined as lacking movement of adjacent spinous processes on active and as hard endfeel on passive motion palpation. The subjects were 21 patients with low back pain (LBP) and 25 asymptomatic chiropractic students; the subjects ranged in age from 23 to 60 years. Their spinous processes were marked. Pair-wise κ and κ_m values were calculated; an unspecified test determined significance of κ values. Five of 21 passive motion palpation pair-wise κ's and two of 21 active motion palpation pair-wise κ's reached significance.

Lumbar Spine

Gonnella et al[15] studied five physical therapists using sidelying T12-S1 FL, SB, and ROT PPIVM; the raters first reached consensus on grading criteria. Eyesight was occluded for the first examination of the intrarater study. The rating scale was a 7-point scale, but adding half-point scores expanded it to a 13-point scale. However, the raters used only grades between 1 and 4. Subjects were five asymptomatic endomorphic female physical therapy students (22-27 years old). The authors provided only summary descriptive statistics. Reliability was highest at L1-L3 and lowest at L5-S1.

Larsson[36] reported on four raters (one chiropractor, one senior chiropractic student and two sophomore chiropractic students) using active seated motion palpation of combined L1-S1 FL and EXT, and SB after previous supervised instruction. The rating scale was dichotomous: absence or presence of fixation defined as the absence of movement of adjacent vertebra. The subjects were 32 asymptomatic chiropractic students (18-40 years old) with vertebral levels marked. In the intrarater study, three blindfolded raters each rated five subjects five times. Agreement, defined as agreement on absence or presence of fixation, was expressed with percent agreement values; in the interrater study, all four raters needed to agree. Of 271 cases of interrater agreement, 6 agreed on the presence and 265 on the absence of fixation.

Grant and Spadon[37] studied two senior and two junior chiropractic students using prone PPIVM L1-S1 SB on a table with a free-floating rear section after four weeks of training using this table. The rating scale was dichotomous: absence or presence of the "normally" coupled rotatory motion of the spinous process. The subjects were 60 chiropractic students (18-52 years old). Three blindfolded raters examined six subjects three times for the intrarater study. Agreement, defined as agreement on absence or presence of fixation, was expressed with percent agreement values. Agreement was on the presence of fixation in 14 cases and on the absence in 386 cases.

Bergstrom and Courtis[38] studied two blindfolded senior chiropractic students using seated L1-S1 SB PPIVM. The rating scale was dichotomous: absence or presence of a fixation indicated by a hard endfeel. The subjects were 100 chiropractic students; 20 of these subjects also participated in the intrarater study. Vertebral levels were marked. Data were analyzed with percent agreement values. Of 818 agreements, 99 were on the presence of a fixation.

Jull and Bullock[39] reported on two manipulative physical therapists using T12-S1 FL, EXT, unilateral SB and ROT PPIVM, and a prone L1-L5 PA. The rating scale was a 5-point scale. Twenty pain-free subjects (12 females, eight males; 20-63 years old) were used for the intrarater study, ten asymptomatic subjects (six females and four males; 22-54 years old) for the interrater study. Percent agreement and r values were calculated.

Boline et al[40] studied a chiropractor and a senior chiropractic student using seated T12-S1 passive motion palpation after 20 hours of practice. Muscle spasm, pain, and fixation indicated by a hard endfeel during motion palpation were all graded on a dichotomous scale: absent or present. The study used 23 patients with LBP and 27 asymptomatic subjects. Order of testing was random. Agreement, defined as agreement on absence or presence of the three dimensions listed above, was expressed with percent agreement and κ values. Combined scores, derived by summing across all three dimensions for each segment, were analyzed with percent agreement, κ, κ_w, and r values. Collapsing data to three regions yielded κ's of ≤ 0.28 for agreement on T12-L2 fixation, ≤ 0.40 for pain at T12-L2, and ≤ 0.29 for muscle spasm at L4-S1.

Mootz et al[41] studied two chiropractors using seated L1-S1 PPIVM after practice sessions. The rating scale was dichotomous: presence or absence of a fixation indicated by a hard endfeel in any of the six directions tested. The subjects were 60 chiropractic students, some symptomatic; order of testing was random. The results were analyzed with κ; significance of κ was determined with an unspecified test. Four segmental intrarater κ's and three collapsed intrarater κ's were statistically significant at $P<0.05$ (or lower). No interrater κ's for the collapsed segments reached statistical significance.

Leboeuf et al[42] reported on four senior chiropractic students using lumbosacral motion palpation. The rating scale was not defined, but likely dichotomous. The subjects were 45 patients with LBP of at least six months' duration (18-79 years old). Interrater reliability was studied by comparing findings of two raters on the first and the fifth visit; intrarater reliability was determined by comparing the findings of one rater on the first visit and the second visit prior to therapy. Percent agreement values were calculated for perfect agreement on either presence or absence of findings and for partial agreement, defined as agreement on the presence of a finding but disagreement on its location. Using normal approximation to the binomial distribution, the authors determined that for a sample size of 40, 68% agreement was necessary for statistical significance. The majority of agreement was on absence of findings.

Richter and Lawall[43] reported on five medical doctors who used seated L1-S1 FL, EXT, SB, and ROT PPIVM and prone L1-L5 PA. The findings of four of the raters were collapsed into a hypothetical second rater. All PPIVM tests were rated on a dichotomous scale: normal or decreased mobility. The PA test was rated on this same dichotomous scale and on absence or presence of pain. The interrater study used 35 patients with LBP; 26 patients were subjects for the intrarater study. Agreement was analyzed with κ. For the PPIVM tests, κ's were provided for total interrater agreement and agreement on normal or decreased ROM. With five exceptions in 30 PPIVM tests, the κ values for normal mobility were higher than the κ's for detection of decreased ROM.

Phillips and Twomey[44] reported on two manipulative physical therapists using lumbar FL and EXT PPIVM, central and unilateral PA, and transverse pressures. Rating was a 3-point scale. The subjects were LBP patients (24-70 years old). The results were analyzed with percent agreement and κ values. A large proportion of subjects had symptoms at the lowest two lumbar segments; the authors suggested this might have led to the higher percent agreement values in the upper lumbar spine where there were fewer abnormal motion palpation findings as well as to the limited variation affecting κ values.

Maher and Adams[9] reported on six manipulative physical therapists using central L1-L5 PA. Mobility and pain during the test were recorded on 11-point scales; raters tended to collapse the 11-point stiffness scale to a 6-point scale. Each rater pair evaluated 30 patients with LBP in their own clinic; skin markings indicated spinal levels and order of testing was varied. Data were analyzed with ICC (1,1) and percent agreement scores.

Binkley et al[45] studied six physical therapists, including three manipulative therapists, using PA on an arbitrarily marked spinous process between L1 and S1 after training sessions with and prior use of the rating scale. Rating was on a 9-point mobility scale. The subjects were 18 patients (23-62 years old) with nonspecific mechanical LBP of two months' to ten years' duration. Order of testing was random. The results were analyzed with percentage agreement, κ_g, ICC (1,1), and SEM.

Inscoe et al[12] reported on two physical therapists using right sidelying double leg T12-S1 FL PPIVM. Rating was on a 3-point mobility scale. Six patients with recurrent LBP and a slender (intermediate) build (24-34 years old) served as subjects. The L5 spinous process was marked. The results were analyzed with percent agreement and Scott's π. In the interrater study, disagreements occurred involving a 2-grade difference in 8.33% and a 1-grade difference in 43.05%. No 2-grade differences occurred in the intrarater study.

Phillips and Twomey[46] studied two manipulative physical therapists using lumbar FL and EXT PPIVM, central and unilateral PA, and transverse pressures. All tests were rated on a 3-point mobility scale; PAIVM tests were also rated on a 2-point scale for the absence or presence of tissue resistance through range or at endrange. The subjects were 72 patients with mechanical LBP and nine subjects without a history of LBP; all subjects had a mean age of around 50. The raters were blinded to the subjects' general mobility and to the others' examination and findings; the first rater examined without getting feedback regarding pain, the second rater examined with patient feedback on pain. Results were analyzed with percentage agreement, κ, and κ_w values. The authors implicated limited variation for low κ values.

Maher et al[23] reported on two studies in which physical therapists used a prone central pisiform grip L3 PA. The perceived stiffness was rated on an 11-point scale. In the first study, three raters compared the PA stiffness of 13 asymptomatic subjects (26-41 years old) to a mechanical stiffness stimulus rated as normal; then they rated the subject's PA stiffness on the rating scale. Subjects were always tested in the same order of rater. In the second study, two manipulative therapists rated 27 asymptomatic subjects (18-43 years old) on the rating scale with mechanical reference stimuli provided for all 11 points of the rating scale. In this second study, all variables known to affect stiffness judgments were controlled. The results were analyzed with ICC (2,1) and SEM.

Discussion

Inevitably, the research validity of even methodologically sound research can be questioned. Statistical conclusion validity, external validity, and construct validity specific to the studies presented above are discussed below.

Statistical Conclusion Validity

A number of studies[26,33,36-38,42] have exclusively used percentage agreement values as an index of agreement. Because percentage agreement values do not correct for agreement based on chance, conclusions regarding reliability of motion palpation based on the results of these studies lack statistical conclusion validity.

Love and Brodeur[34] and Jull and Bullock[39] used Pearson's r to quantify agreement. However, as discussed earlier, correlation coefficients express covariance rather than agreement. Also, in both studies the data analyzed were not continuous, nor on an interval or ratio level scale: Pearson's r is an inappropriate statistic for use with ordinal level data. Therefore, both studies lack sufficient statistical conclusion validity to allow for conclusions regarding reliability.

A number of studies[25,27,34,35,41,42] used tests to determine statistical significance of κ or correlation coefficient values. The influence of sample size on statistical significance of κ and r values has been discussed above. Tables 2 and 3 list the benchmark comparison values commonly used for κ and r values obtained in reliability research.

That chance-corrected indices of agreement are the statistics of choice in reliability research has also been discussed. Many of the studies presented here have used a variation of the κ statistic. As suggested by Lantz[10] some studies[24,25,27,31,40,44-46] provided both percentage agreement and κ values; other studies[4,28-30,32,35,41,43], however, only provided κ values. This makes it hard to determine the possible influence of limited variation in the data set on the κ values obtained. Some studies used κ_w values[25,40,46]; Binkley et al[45] analyzed their data with κ_g. None of these studies provided data regarding assignment and value of the weights used in calculating these statistics. Smith et al[32] and Keating et al[35] used the κ_m statistic: however, it was unclear from these studies if a similar magnitude of SEM of κ allowed for the use of κ_m. Inscoe et al[12] used Scott's π as a chance-corrected index of agreement, but information on the correct use of this statistic is lacking in commonly used statistics textbooks[6,7].

Limited variation in the data set renders the κ statistic inappropriate for use as an index of agreement. A number of the studies discussed[15,23-25,27,28,36] have used asymptomatic subjects. If assumption #1 made in the introduction above on the relation between spinal motion abnormalities and symptom status is correct, the use of asymptomatic subjects may result in a highly homogenous study population predominantly devoid of motion abnormalities; therefore, κ would be an inappropriate measure of agreement.

The reliability of a rating scale increases as the number of points available and the number used by the raters increases[45]. Chance agreement increases if few categories on the rating scale are used by the raters[9]. Many of the studies[24,26,27,29-31,33,35-38,40-43,46] used dichotomous rating scales. In other studies[9,15,32], the raters tended to use only a limited portion of the rating scale. Dichotomous scales and limited use of the rating scale weaken the statistical conclusion validity of a research study.

If interrater agreement between two raters is examined, agreement as a result of chance will occur more frequently than in a study of interrater agreement between more than two raters. Most studies, even the studies with multiple raters, have calculated pair-wise indices of agreement. Only three studies presented here[23,36,37] have truly studied reliability between more than two raters.

Leboeuf et al[42] provided their data in a very summary format and did not present segmental data; in general, a summary presentation of quantitative research data weakens the support for any conclusions presented[47].

External validity

How similarity in subjects, raters, motion palpation technique, rating scale, and setting affect external validity was discussed earlier. We also discussed how research that used asymptomatic subjects[15,23-25,27,28,36] may not allow for generalization of the study results to a symptomatic population. Many of the studies[4,15,24-27,33-39] used students as subjects; experienced subjects may respond differently than subjects without experience with the technique studied. Studies have used patients as subjects with cervicogenic headache[28,29], neck pain[29-31]upper-quarter disorders[31], thoracic pain[4], LBP[9,35,40,41,43,44,46], recurrent LBP[12], and prolonged LBP[42,45]. Matching the study population to the population of interest will increase external validity.

The unclear relationship between rater experience level and reliability has also been discussed. Many of the studies[24,26,27,33,34,36-38,40,42] used student raters. Raters received various levels of training prior to the study. The studies presented here used chiropractors[4,24-27,33-38,40-42], medical doctors[30,43], and physical therapists[9,12,15,23,28,29,31,32,39,44-46] as raters.

Despite the large number of studies presented here, only a limited number of commonly employed spinal motion palpation techniques have been investigated; no conclusions can be made regarding reliability of techniques not researched in the studies reported. Table 5 lists a number of parameters affecting the perceived stiffness during PA testing; the experienced clinician will easily come up with a list of parameters affecting findings on PPIVM testing. Matching all motion palpation technique parameters will maximize external validity; however, most studies give insufficient detail regarding the specific parameters.

The rating scale used in the clinical or research setting should be similar to the scale used in the study to increase external validity. The use of dichotomous scales versus multi-point ordinal scales has also been discussed. The studies presented used mainly mobility rating scales, but some studies also used pain scales[9,30,40,43], a muscle spasm scale[40], and clinically oriented scales regarding the decision to treat or to include a subject in a treatment trial[4,29,45]. In the conclusion section operant definitions of mobility rating scales will be discussed.

Most studies take place in a highly controlled research environment. Only Maher and Adams[9] did reliability research in a clinical setting; the confounding variables of this setting may have produced the low reliability observed in this study.

Construct Validity

In motion palpation reliability studies, the construct as labeled is always the reliability for a specified motion palpation technique. The study protocol will frequently, however, produce a quite different construct as implemented.

Blindfolding raters[15,24,36-38], draping sheets over the subjects, and placing restrictions on talking to prevent clinicians from recognizing subjects may eliminate confounding variables in motion palpation reliability studies. Clinically, this multi-sensory feedback during testing may be an important factor in helping clinicians to confirm their assessment[12]. The construct as implemented then becomes reliability of a technique devoid of potentially clinically crucial sensory feedback.

Previous training sessions might be expected to help standardize techniques among raters. The studies presented used various levels of rater training and standardization of techniques and rating scale. Training the raters prior to the study introduces the effect of this training into the research equation; the construct as implemented at least partly becomes the effect of rater training on the reliability of the technique studied rather than just reliability of the technique used in the study.

Motion palpation testing sometimes inadvertently leads to manipulation of the segment tested. Repeated motion palpation testing will most likely affect tissue compliance. The changes induced in the system by diagnostic procedures may be the reason for low reliability of these techniques[1]. Some studies[9,27-31,40,41] attempted to compensate for this effect by randomizing testing order. If the diagnostic motion palpation affects segmental mobility, then establishing reliability becomes an exercise in futility. After all, reliability can only exist if the status of the subject being examined has not changed between examinations[48]. The construct as implemented here becomes the effect of repeated mobilizing stress on segmental mobility as measured by motion palpation.

The literature provides unclear evidence as to the ability of physical therapists to correctly palpate spinal levels. McKenzie and Taylor[49] showed that physical therapists without advanced manipulative therapy qualifications had poor interrater reliability in correctly locating spinal levels. Binkley et al[45] also showed poor interrater reliability in a group of physical therapists that included three manipulative physical therapists. In contrast, Downey et al[50] demonstrated that manipulative physical therapists have good interrater reliability in locating lumbar spinal levels, and they suggested that advanced spinal therapy training positively affects this ability. Motion palpation reliability studies on subjects with unmarked spinal levels may, therefore, introduce another source of poor reliability in the study. The construct as implemented becomes the ability of the raters to reliably palpate a spinal level in addition to their ability to reliably use a specified motion palpation technique at the correctly identified spinal level. Some of the studies presented here[4,34,35,45] attempted to eliminate this confounding factor by marking spinal levels. Inscoe et al[12] warned about the effect of improper segment identification when marking in one position and testing in another position. Other studies have collapsed individual segments into multi-segment units[25,40,41] or created a category of partial agreement that allowed for agreement on presence but not necessarily on exact level[42]. For discussion on the reasoning behind this, see the conclusion.

Conclusion

To draw a conclusion regarding reliability of a specific motion palpation technique, it is important to find a study that matches as much as possible the clinical situation to which we would like to generalize the study findings, bearing in mind the methodological flaws and subsequent threats to the research validity of the study as discussed above. It will be obvious that in order to draw conclusions regarding many clinically used motion palpation techniques, much research remains to be done. This research needs to use appropriate patients as subjects, raters with different experience levels and post-graduate qualifications, and clinically usable rating scales. Research also needs to be done in the clinical setting rather than in a strictly controlled research environment.

Based on the studies some general conclusions can be drawn:

- Intrarater agreement varies from *less than chance* to generally *moderate* or *substantial* agreement.
- Interrater agreement only rarely exceeds *poor* to *fair* agreement.
- Rating scales measuring absence versus presence or magnitude of pain response yield higher agreement values than mobility rating scales.

One possible explanation for higher intra- than interrater reliability is the effect of the potential lack of reliability when determining the spinal level, as discussed earlier. Raters might correctly identify the presence of the same segmental motion abnormality but incorrectly name the segmental level at which this abnormality was found. Raters can be expected to be very consistent in their (in)correct identification of the spinal level; this will result in higher intra- than interrater reliability values. Also discussed above, some studies provided data analysis on collapsed multi-segmental spinal units to take into account the effect of unreliable determination of segmental level. LeBoeuf et al[42] created a category of partial agreement for this same reason. In the clinical situation, however, it is less important to correctly identify the segmental level of a motion abnormality than it is to identify the presence of an abnormality. A decision to treat is not based on the exact spinal level but rather on the presence of motion abnormalities and symptom response with motion palpation testing. Incorrectly naming the segmental level is of little consequence. When evaluating reliability of motion palpation in the clinical context, we may, therefore, need to consider intra- rather than interrater data.

The assumption that there is a relationship between pain, reduced voluntary movement, and segmental spinal motion abnormalities is the reason for rating absence versus presence or magnitude of pain with motion palpation[5]. Maher and Latimer[5] suggested using pain response during motion palpation testing as the main indication for treatment of a spinal segment (if that segment were also physiologically capable of producing the patient's symptoms). In two related articles, Jull et al[51,52] pointed out the potential for false positive findings when relying too heavily on reported pain, due to the widespread pain and referred tenderness in many spinal patients.

In contrast to the relatively simple concept of pain with motion palpation, the operant definition of the concept of spinal stiffness as it is used in mobility rating scales in motion palpation studies is ambiguous. This may explain the difference in reliability when using pain versus mobility ratings. Maher et al[53] used a cluster analysis of descriptors of spinal stiffness found in manual therapy literature; Australian and US physical therapists independently identified the same three different dimensions of spinal stiffness: limited mobility, increased mobility, and nature of resistance felt in response to testing. Other dimensions may include endfeel, change in resistance relative to motion, and overall impression[45,51-53]. Research into motion palpation reliability would benefit from good operant definitions of the dimensions of spinal stiffness for use in mobility rating scales. Spinal stiffness is a multidimensional concept and is, therefore, unlikely to be satisfactorily rated on a single mobility rating scale[53].

Even with rating scales addressing the different aspects of the concept of spinal stiffness, the method for determining presence versus absence or magnitude of motion abnormality remains unclear. Spinal stiffness parameters could be graded by comparing adjacent levels or by using rater experience regarding normal values for a specific level[5]. However, differences between subjects and between adjacent segments in the same subject make this method very unreliable. Lee et al[54] demonstrated that

PA stiffness values between T4 and T5 may differ up to 125% in the same individual; stiffness values at T4 and T5 varied up to a factor of four between asymptomatic subjects. PA stiffness at L3 varied with up to almost a factor of three between asymptomatic subjects[55]. Maher[22] suggested using mechanical stiffness stimuli as reference stimuli for PA testing.

The different variables (Table 5) that affect the stiffness perceived during PA testing have been discussed. A similar table could be constructed for PPIVM testing. Intrarater reliability is probably higher than interrater reliability, because raters consistently use the same parameters between tests; different raters are less likely to use the exact same set of parameters thus confounding motion palpation findings. Identifying and controlling for these environmental and technique parameters affecting stiffness perception will likely improve reliability[22].

The suggestions here should allow for the development of motion palpation tests with greater intra- and interrater reliability. The preceding information should help assure statistical conclusion validity, external validity, and construct validity of future reliability studies. Reliability studies are not meant to be independent precursors to validity studies; when used as such, they can lead to the promotion of highly reliable but clinically useless tests. They may also exclude potentially useful tests based on a demonstrated lack of reliability[48]. Low reliability may explain a lack of accuracy; the diagnostic value of a test may be improved if the test is adapted to provide improved reliability[48].

On the other hand, prior to researching motion palpation reliability (and validity), we may need to rethink some of our assumptions. What is the role of motion palpation in the examination of the patient? Does motion palpation provide the needed crucial information on location, nature, and direction of the spinal segmental motion abnormality as proposed by some authors[56,57] or is it just one component of the complex of history, tests, and measures that direct our attention and determine our diagnosis, prognosis, and intervention? How do effective clinicians arrive at a diagnosis of spinal segmental dysfunction specific to location, nature, and even direction? In other words, perhaps it is not realistic to research motion palpation for reliability (and validity) as if it were a stand-alone and crucial diagnostic tool. Perhaps we should research the end result of using all diagnostic tools including motion palpation, i.e., the diagnosis, rather than attaching an importance to this one diagnostic tool that it does not deserve based on its clinical use and value?

Another assumption is that a specific diagnosis as to location, nature, and direction of the segmental dysfunction is needed because only a specific intervention (mobilization, manipulation, or muscle energy technique) with a specific speed, amplitude, and duration in a specific direction will restore normal segmental function[56-58]. But even if the four assumptions that form the rationale for the use of diagnostic motion palpation as mentioned in the introduction are true, what proof exists that such a specific intervention is more effective and efficient than a generic intervention, e.g., regional manipulation[58]? It should be obvious that the demands on the reliability of diagnostic motion palpation become immensely less stringent when a generic intervention proves equally or more effective than a segment-specific intervention in the management of diagnosed segmental motion abnormalities.

Acknowledgements

I would like to thank Dr. Paul Beattie, PhD, PT, OCS, for challenging my assumptions with regards to motion palpation; Dr. David Peterson, DC, from WSCC in Portland, OR, for explaining chiropractic terminology; and the libraries at Pacific University in Forest Grove, OR, and at WSCC for their help with collecting the references. ■

REFERENCES

1. Russell R. Diagnostic palpation of the spine: A review of procedures and assessment of their reliability. *J Manipulative Physiol Ther* 1983;6:181-183.
2. Meadows JTS. *Orthopedic Differential Diagnosis in Physical Therapy*. New York, NY: McGraw-Hill, 1999.
3. Alley JR. The clinical value of motion palpation as a diagnostic tool: A review. *JCCA* 1983;27:97-100.
4. Haas M, Raphael R, Panzer D, Peterson D. Reliability of manual end-play palpation of the thoracic spine. *Chiropractic Technique* 1995;7:120-124.
5. Maher C, Latimer J. Pain or resistance: The manual therapists' dilemma. *Aust J Physiother* 1992;38:257-260.
6. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice*. Norwalk, CT: Appleton & Lange, 1993.
7. Domholdt E. *Physical Therapy Research: Principles and Applications*. Philadelphia, PA: W.B. Saunders Company, 1993.
8. Haas M. Statistical methodology for reliability studies. *J Manipulative Physiol Ther* 1991;14:119-132.

9. Maher C, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Phys Ther* 1994;74:801-811.
10. Lantz CA. Application and evaluation of the kappa statistic in the design and interpretation of chiropractic clinical research. *J Manipulative Physiol Ther* 1997;20:521-528.
11. French SD, 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;23:231-238.
12. Inscoe EL, Witt PL, Gross MT, Mitchell RU. Reliability in evaluating passive intervertebral motion of the lumbar spine. *J Man Manipulative Ther* 1995;3:135-143.
13. Nicholson L, Adams R, Maher C. Reliability of a discrimination measure for judgements of non-biological stiffness. *Man Ther* 1997;3:150-156.
14. Mior SA, McGregor MM, Schut B. The role of experience in clinical accuracy. *J Manipulative Physiol Ther* 1990;13:68-71.
15. Gonnella C, Paris SV, Kutner M. Reliability in evaluating passive intervertebral motion. *Phys Ther* 1982;62:436-444.
16. Lee M, Svensson NL. Effect of loading frequency on response of the spine to lumbar posteroanterior forces. *J Manipulative Physiol Ther* 1993;16:439-446.
17. Viner A, Lee M. Direction of manual force applied during assessment of stiffness in the lumbosacral spine. *J Manipulative Physiol Ther* 1995;18:441-447.
18. Maher C, Adams R. A comparison of pisiform and thumb grips in stiffness assessment. *Phys Ther* 1996;76:41-48.
19. Maher CG, Adams RD. Stiffness judgments are affected by visual occlusion. *J Manipulative Physiol Ther* 1996;19:250-256.
20. Edmonston SJ, Allison GT, Gregg CD, Purden SM, Svansson GR, Watson AE. Effect of position on the posteroanterior stiffness of the lumbar spine. *Man Ther* 1998;3:21-26.
21. Maher CG, Latimer J, Holland MJ. Plinth padding confounds measures of posteroanterior spinal stiffness. *Man Ther* 1999;4:145-150.
22. Maher C. Perception of stiffness in manipulative physiotherapy. *Physiotherapy Theory and Practice* 1995;11:35-44.
23. 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;78:829-837.
24. Mior SA, King RS, McGregor MM, Bernard M. Intra and interexaminer reliability of motion palpation in the cervical spine. *JCCA* 1985;29:195-198.
25. DeBoer KF, Harmon R, Tuttle CD, Wallace H. Reliability study of detection of somatic dysfunctions in the cervical spine. *J Manipulative Physiol Ther* 1985;8:9-16.
26. Bronemo L, Van Steveninck J. A comparison of the inter- and intraexaminer reliability of motion palpation of the lower cervical spine (C2-C7) in the oblique-posterior-lateral direction in sitting and supine positions. Thesis. Bournemouth, UK: Anglo-European College of Chiropractic, 1987.
27. Nansel DD, Peneff AL, Jansen RD, Cooperstein R. Interexaminer concordance in detecting joint-play asymmetries in the cervical spines of otherwise asymptomatic subjects. *J Manipulative Physiol Ther* 1989;12:428-433.
28. Schoensee SK, Jensen G, Nicholson G, Gossman M, Katholi C. The effect of mobilization on cervical headaches. *J Orthop Sports Phys Ther* 1995;21:184-196.
29. Jull G, Zito G, Trott P, Potter H, Shirley D, Richardson C. Interexaminer reliability to detect painful upper cervical joint dysfunction. *Aust J Physiother* 1997;43:125-129.
30. Schoeps P, Pfingsten M, Siebert U. Reliabilitaet manualmedizinischer Untersuchungstechniken an der Halswirbelsaeule. Studie zur Qualitaetssicherung in der manuellen Diagnostik. *Z Orthop Ihre Grenzgeb* 2000;138:2-7.
31. Smedmark V, Wallin M, Arvidsson I. Interexaminer reliability in assessing passive intervertebral motion of the cervical spine. *Man Ther* 2000;5:97-101.
32. Smith AR, Catlin PA, Nyberg RE. Intratester/intertester reliability of segmental motion testing of cervicothoracic forward bending in a symptomatic population. In: Paris SV, Ed. *IFOMT Proceedings*; June 1-5, 1992; Vail, CO; IFOMT, 1992;194.
33. Loram A. A comparative study of fixation findings in the thoracic spine in the sagittal plane using motion palpation in the sitting position and joint springing in the prone position. Thesis. Bournemouth, UK: Anglo-European College of Chiropractic, 1987.
34. Love RM, Brodeur RR. Inter- and intraexaminer reliability of motion palpation for the thoracolumbar spine. *J Manipulative Physiol Ther* 1987;10:1-4.
35. Keating JC, Bergmann TF, Jacobs GE, Finer BA, Larson K. Interexaminer reliability of eight evaluative dimensions of lumbar segmental abnormality. *J Manipulative Physiol Ther* 1990;13:463-470.
36. Larsson AC. A pilot study to compare the intra- and interreliability of examiners using Gillet's motion palpation methods in the lumbar spine. Thesis. Bournemouth, UK: Anglo-European College of Chiropractic, 1984.
37. Grant A, Spadon R. An inter- and intraexaminer reliability study, using lateral flexion motion palpation of the lumbar spine in the prone position. Thesis. Bournemouth, UK: Anglo-European College of Chiropractic, 1985.
38. Bergstrom E, Courtis G. An inter- and intraexaminer reliability study of motion palpation of the lumbar spine in lateral flexion in the seated position. *Eur J Chiropr* 1986;34:121-144.
39. Jull G, Bullock M. A motion profile of the lumbar spine in an ageing population assessed by manual examination. *Physiotherapy Practice* 1987;3:70-81.
40. Boline PD, Keating JC, Brist J, Denver G. Interexaminer reliability of palpatory evaluations of the lumbar spine. *Am J Chiropractic Med* 1988;1:5-11.
41. Mootz RD, Keating JC, Kontz HP, Milus TB, Jacobs GE. Intra- and interobserver reliability of passive motion palpation of the lumbar spine. *J Manipulative Physiol Ther* 1989;12:440-445.
42. LeBoeuf C, Gardner V, Carter AL, Scott TA. Chiropractic examination procedures: A reliability and consistency study. *J Aust Chiropr Assoc* 1989;19:101-104.
43. Richter T, Lawall J. Zur Zuverlaessigkeit manualdiagnostischer Befunde. *Man Med* 1993;31:1-11.
44. Phillips DR, Twomey LT. Comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. In: Singer KP, Ed. *Integrating Approaches. Proceedings of the Eighth Biennial Conference of the Manipulative Physiotherapists Association of Australia*; Nov 24-27, 1993; Perth, Western Australia; Manipulative Physiotherapists Association of Australia,

1993;55-61.

45. Binkley J, Stratford PW, Gill C. Interrater reliability of lumbar accessory motion mobility testing. *Phys Ther* 1995;75:786-795.
46. Phillips DR, Twomey LT. A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. *Man Ther* 1996;2:82-87.
47. Panzer DM. The reliability of lumbar motion palpation. *J Manipulative Physiol Ther* 1992;15:518-524.
48. Fritz JM, Wainner RS. Examining diagnostic tests: An evidence-based perspective. *Phys Ther* 2001;81:1546-1564.
49. McKenzie AM, Taylor NF. Can physiotherapists locate lumbar spinal levels by palpation? *Physiotherapy* 1996;82: 235-239.
50. Downey BJ, Taylor NF, Niere KR. Manipulative physiotherapists can reliably palpate nominated lumbar spinal levels. *Man Ther* 1999;4:151-156.
51. Jull G, Treleaven J, Versace G. Manual examination of spinal joints: Is pain provocation a major diagnostic cue for dysfunction? In: Singer KP, Ed. *Integrating Approaches*. *Proceedings of the Eighth Biennial Conference of the Manipulative Physiotherapists Association of Australia*; Nov 24-27, 1993; Perth, Western Australia; Manipulative Physiotherapists Association of Australia, 1993;40-42.
52. Jull G, Treleaven J, Versace G. Manual examination: Is pain provocation a major cue for spinal dysfunction? *Aust J Physiother* 1994;40:159-165.
53. Maher CG, Simmonds M, Adams R. Therapists' conceptualization and characterization of the clinical concept of spinal stiffness. *Phys Ther* 1998;78:289-300.
54. Lee M, Latimer J, Maher C. Manipulation: Investigation of a proposed mechanism. *Clin Biomech* 1993;8:302-306.
55. Lee M, Esler MA, Mildren J. Effect of extensor muscle activation on the response to lumbar posteranterior forces. *Clin Biomech* 1993;8:115-119.
56. Meadows J. The principles of the Canadian approach to the lumbar dysfunction patient. In: Wadsworth C, Ed. *Management of Lumbar Spine Dysfunction, Home Study Course 9.3*. LaCrosse, WI: Orthopaedic Section, APTA, 1999.
57. Huijbregts PA. Lumbopelvic region: Aging, disease, examination, diagnosis, and treatment. In: Wadsworth C, Ed. *Current Concepts of Orthopedic Physical therapy, Home Study Course 11.2*. LaCrosse, WI: Orthopaedic Section, APTA, 2001.
58. Nelson C. The subluxation question. *J Chiropr Humanities* 1997;7:46-55.