Reliability of Joint Mobility and Pain Assessment of the Thoracic Spine and Rib Cage in Asymptomatic Individuals

BRYAN HEIDERSCHEIT, PT, PhD¹, WILLIAM BOISSONNAULT, PT, DHSc, FAAOMPT²

The examination and evaluation of joint mobility dysfunction is an important factor in the differential diagnosis of thoracic and chest pain¹.

Mobility dysfunction of the thoracic vertebrae and rib articulations has been linked to a variety of pain syndromes of the upper quarter, low back, and neck re-

gions as well as atypical chest pain²⁻⁴. The presence of a symptomatic joint dysfunction would suggest a musculoskeletal origin of, or contribution to, a patient's chief complaint.

While a variety of joint mobility tests have been described^{5,6}, the application of a posteroanterior (PA) pressure to the joint in question is routinely used⁷. During this test, the clinician assesses the magnitude of the motion with respect to the applied force, thus qualitatively determining the joint stiffness. In addition, the clinician is observant of any accompanying pain response by the patient, often-times directly questioning the patient as to whether the applied pressure provoked pain. This combined information can then be integrated by the clinician with additional examination findings to determine the appropriate diagnosis and develop a treatment plan. A survey of manual physical therapists revealed that 98% of the therapists based at least part of their treatment decision on the results of segmental motion assessment⁸. For example, an identified hypermobile or hypomobile joint based on accessory motion evaluation can be used to classify the nature of non-specific back pain⁹ as well

ABSTRACT: Despite the importance of correctly diagnosing a spinal dysfunction, limited research exists related to physical therapists' ability to reliably identify a joint exhibiting signs of dysfunction. The purpose of this investigation was to determine the inter- and intra-examiner reliability of a thoracic spine and rib cage joint mobility and pain assessment between two experienced manipulative physical therapists. Nine healthy subjects (3 male, 6 female; ages 23-35) without history of mid- or low back pain participated. Posterior-to-anterior pressures were applied to the thoracic spine and rib articulations with anterior-to-posterior pressures applied to the costosternal joints of each subject by two examiners to evaluate joint mobility and pain provocation. Both examiners assessed all subjects twice and were blinded to subject identity. Kappa statistics were calculated using a strict and expanded definition of agreement to determine the between- and within-examiner reliability for each outcome. Intra-examiner reliability of joint mobility assessment ranged from *slight* to *fair* based on the strict agreement but improved to *good* when findings were compared across ± 1 spinal/rib level. Pain provocation reliability increased to *very good* under the expanded agreement; however, this finding should be viewed with caution due to limited pain prevalence in the subject sample. Selected clinical prediction rules, applied to the care of individuals with back pain, characterize the patient's regional mobility simply as *hypomobile*, *normal*, or *hypermobile*; consequently, we feel the results of an expanded definition of agreement may be more appropriate for clinic practice. Further research is needed to determine the reliability in individuals with thoracic spine and rib cage symptoms.

KEYWORDS: Joint Flexibility, Observer Variation, Rib, Thoracic Vertebrae

¹Assistant Professor, Physical Therapy Program, Department of Orthopedics and Rehabilitation, University of Wisconsin-Madison, Madison, WI.

²Associate Professor, Physical Therapy Program, Department of Orthopedics and Rehabilitation, University of Wisconsin-Madison, Madison, WI.

The authors wish to acknowledge financial support provided through an OPTP grant from the American Academy of Orthopaedic Manual Physical Therapists.

Address all correspondence and requests for reprints to: Dr. Bryan Heiderscheid, heider@orthorehab.wisc.edu

as serving as an initial point of intervention. The treatment classification system summarized by Fritz et al¹⁰ includes lumbar spine hypomobility or hypermobility as one of the criteria associated with deciding whether spinal thrust manipulation or stabilization exercises, respectively, are appropriate. The use of this treatment classification system among individuals with acute low back pain resulted in a significant improvement in disability during the initial weeks of care compared to treatment based on clinical practice guidelines¹¹.

Examiner reliability of PA pressures during spinal and rib joint assessments has been previously investigated, with between and within-examiner reliability ranging from poor to moderate^{6,12,13}. However, methodological limitations of these prior studies have been identified (e.g., inappropriate statistical analysis of agreement between and within examiners; influence of spinal level identification on examiner reliability), suggesting that additional evaluation of this common clinic measure is warranted¹⁴. Further, the reliability of both motion stiffness and pain provocation has not been simultaneously evaluated in both the thoracic spine and rib cage. Given the routine use of this examination technique, the determination of reliability is necessary. Therefore, the purpose of this investigation was to determine the inter- and intra-examiner reliability of thoracic spine and rib cage joint mobility assessment and associated pain provocation between two experienced manipulative physical therapists. To address the influence of spinal level identification on examiner reliability, both a strict and expanded definition of agreement was employed.

Methods

Subject Selection

Nine healthy subjects (3 male, 6 female; ages 23–35) without history of mid- or low back pain participated. Due to the examination techniques, individuals were excluded if they reported being previously diagnosed with a bone or joint condition/disease such as rheumatoid arthritis or other systemic connec-

tive tissue disease, severe joint effusion, osteopenia or osteoporosis, or unhealed fracture. Subjects without mid- or low back pain were selected as the degree of spinal segmental stiffness has been found to be retained over time in asymptomatic individuals^{15–17}, thus reducing the likelihood that subjects' joint mobility would change during the data collection procedures. Further, the use of asymptomatic subjects is not without clinical merit, as thoracic spine and rib cage mobility assessment are frequently performed on patients void of back pain but who are symptomatic in adjacent body regions (e.g., shoulder or cervical pain)^{2,3}. Each subject provided written informed consent consistent with a protocol approved by the University of Wisconsin's Health Sciences Institutional Review Board.

Assessment

The thoracic spine and rib cage mobility characteristics of each subject were assessed twice by each of 2 examiners within a 4-hour period. Each examiner had over 10 years experience in clinic practice using the employed assessment techniques. Prior to the examination period, the examiners practiced the assessment procedures that would be performed, standardizing both the sequence of tests and the relative magnitude of pressures. To prevent the examiners from recalling subject-specific mobility characteristics, examiners were masked to subject identity through the use of draping (Figure 1) and a random exam-

ination table assignment. Further, each examiner remained masked to the other's findings.

Passive accessory joint motions within the thoracic spine and rib cage were determined by examiner-applied manual pressures (Figure 1). With subjects positioned in prone, five consecutive PA pressures were applied to the spinous processes (central) and transverse processes (unilaterals) of the thoracic vertebrae 1 through 12 (T_1 – T_{12}). All pressures were systematically applied from cephalad to caudad, beginning with the central pressures, followed by right unilaterals and ending with left unilaterals. Posterior rib mobility was then assessed through five consecutive PA pressures applied to the rib angles of ribs 1 through 12 beginning with the right side. Subjects were then repositioned in supine for the assessment of anterior rib mobility with five consecutive anterior-to-posterior (AP) pressures applied to the costosternal joints of ribs 1 through 7. Mobility of each joint was judged as being *hypomobile*, *hypermobile*, or *normal*. A research assistant accompanied each examiner to record all findings.

Despite the fact that the participants were asymptomatic, pain provocation that occurred during the PA and AP pressures was also recorded. Pain was defined for the participants as *a sensation other than just pressure* and included both local and referred sensations. A standard set of questions regarding pain provocation was asked by each examiner. Participants were instructed to ver-

FIGURE 1. Passive accessory joint mobility was assessed through posterior to anterior pressures with standardized technique between examiners. Joint mobility was rated as *hypomobile*, *normal*, or *hypermobile*. Pain provocation during this assessment was recorded as *present* or *absent*.

bally state whether pain was provoked (present or absent) during the pressure applications at each joint. In the event that pain was perceived, the examiner reapplied the pressure at a non-painful thoracic location to ensure that the perceived pain exceeded the sensation of pressure only.

Following the initial assessment of each subject by the two examiners, the subjects were randomly assigned a different examination table and re-draped for the second assessment. All procedures for the second assessment were consistent with the first.

Statistical Analysis

The intra- and inter-examiner reliability for joint mobility and pain provocation for the thoracic spine and rib cage were assessed using percent agreement and kappa statistics with 95% confidence intervals (CI). Central and unilateral pressures were pooled across spinal levels to create a composite score for the thoracic spine. Pressures to the anterior and posterior ribs were similarly pooled for the rib cage. Only the initial assessment data were used to determine inter-examiner reliability. To account for potential segment-level identification inaccuracies, an expanded definition of agreement was also used¹³. Using this expanded definition, agreement with regard to the localization of findings was present if it was reproduced during the second examination session and located in the exact same spinal level or in a neighboring level (± 1 spinal segment). All kappa values were interpreted using previously defined categories¹⁸: *slight* (≤ 0.20), *fair* (0.21–0.40), *moderate* (0.41–0.60), *good* (0.61–0.80), and *very good* (0.81–1.00). In addition, the maximum attainable kappa, prevalence index, and bias index were calculated to assist in the interpretation of the corresponding kappa values^{19,20}. The maximum attainable kappa is the maximum value that kappa could attain for the set of data, with the difference between it and kappa being reflective of the unachieved agreement beyond chance given the study's pre-existing factors¹⁹. Prevalence index reflects differences in positive and negative agreement while bias index reflects dif-

ferences in the number of disagreements, both of which can influence the magnitude of kappa²⁰.

Results

Joint Mobility

Due to the minimal presence of a hypermobility dysfunction (Table 1), the corresponding intra- and inter-examiner reliability values (Table 2) were calculated by collapsing the hypermobility and hypomobility observations into a single mobility dysfunction category. Intra-examiner reliability for joint mobility assessment of the thoracic spine ranged from *slight* (examiner 2, $\kappa=0.17$) to *fair* (examiner 1, $\kappa=0.26$) based on the strict agreement calculation. Inter-examiner reliability was *slight* ($\kappa=0.15$). Under the expanded definition of agreement, intra-examiner reliability for both examiners increased to *good* (examiner 1, $\kappa=0.75$; examiner 2, $\kappa=0.61$) while inter-examiner reliability increased to *moderate* ($\kappa=0.59$).

Similar reliability was present for joint mobility assessment of the rib cage, with intra-examiner reliability increasing from *fair* ($\kappa=0.26$ and 0.29) to *good* ($\kappa=0.71$ and 0.76, respectively) under the

expanded agreement. Inter-examiner reliability was consistently less than intra-examiner reliability (strict agreement, $\kappa=0.11$; expanded agreement, $\kappa=0.64$).

Pain Provocation

Intra-examiner reliability for pain provocation assessment of the thoracic spine ranged from *fair* (examiner 2, $\kappa=0.28$) to *good* (examiner 1, $\kappa=0.66$) based on the strict agreement calculation (Table 3). Inter-examiner reliability was *fair* ($\kappa=0.24$). Under the expanded definition of agreement, intra-examiner reliability for both examiners increased to *very good* (examiner 1, $\kappa=0.89$; examiner 2, $\kappa=0.86$) while inter-examiner reliability increased to *good* ($\kappa=0.62$). However, the limited number of painful joints observed (Table 1) must be considered when interpreting these values.

Reliability of examiner 2 during pain provocation assessment of the rib cage increased from *moderate* ($\kappa=0.49$) to *good* ($\kappa=0.76$) under the expanded agreement. Due to the identification of a limited number of painful rib articulations ($n=5$; Table 1), examiner 1 showed no agreement ($\kappa=0.00$) under the strict definition but increased to perfect agreement ($\kappa=1.00$) under the expanded definition.

TABLE 1. Examiner observations of assessment categories.

	Examiner 1		Examiner 2	
	Time 1	Time 2	Time 1	Time 2
Joint Mobility				
Thoracic Spine (n=324)				
Hypermobility	0	0	2	1
Hypomobility	83	66	55	34
Normal	241	258	267	289
Rib Cage (n=342)				
Hypermobility	2	1	26	20
Hypomobility	47	40	31	29
Normal	293	301	284	294
Pain Provocation				
Thoracic Spine (n=324)				
Present	1	4	7	5
Absent	341	338	335	337
Rib Cage (n=342)				
Present	5	4	3	4
Absent	319	320	321	320

n = total number of joints examined

TABLE 2. Intra-examiner and inter-examiner reliability of joint mobility assessment using both strict and expanded definitions of agreement.

		κ (95% CI)		κ maximum ^a	Percent agreement	Prevalence ^b	Bias ^c
Strict Agreement							
Thoracic Spine							
	Examiner 1	0.26	(0.14–0.38)	0.85	74%	0.54	0.05
	Examiner 2	0.17	(0.04–0.30)	0.72	80%	0.72	0.07
	Inter-rater	0.15	(0.04–0.27)	0.77	71%	0.57	0.08
Rib Cage							
	Examiner 1	0.26	(0.12–0.40)	0.90	83%	0.73	0.02
	Examiner 2	0.29	(0.16–0.42)	0.89	81%	0.69	0.03
	Inter-rater	0.11	(–0.01–0.23)	0.89	77%	0.69	0.03
Expanded Agreement							
Thoracic Spine							
	Examiner 1	0.75	(0.66–0.84)	0.79	91%	0.56	0.08
	Examiner 2	0.61	(0.46–0.75)	0.61	90%	0.75	0.10
	Inter-rater	0.59	(0.47–0.70)	0.67	85%	0.60	0.11
Rib Cage							
	Examiner 1	0.71	(0.57–0.85)	0.73	93%	0.79	0.06
	Examiner 2	0.76	(0.65–0.88)	0.76	94%	0.74	0.06
	Inter-rater	0.64	(0.46–0.82)	0.70	90%	0.81	0.08

^a κ maximum is the proportions of positive and negative ratings by each examiner expressed relative to the greatest possible agreement by adjusting the distribution of paired ratings.

^bPrevalence index is the absolute difference between the positive and negative agreements, relative to the number of paired ratings.

^cBias index is the absolute difference between the positive and negative disagreements relative to the number of paired ratings.

TABLE 3. Intra-examiner and inter-examiner reliability of pain provocation assessment using both strict and expanded definitions of agreement.

		κ (95% CI)		κ maximum ^a	Percent agreement	Prevalence ^b	Bias ^c
Strict Agreement							
Thoracic Spine							
	<i>Examiner 1</i>	0.66	(0.30–1.00)	0.89	99%	0.97	0.00
	<i>Examiner 2</i>	0.28	(–0.16–0.72)	0.86	99%	0.98	0.01
	<i>Inter-rater</i>	0.24	(–0.16–0.64)	0.75	98%	0.98	0.00
Rib Cage							
	<i>Examiner 1</i>	0.00	(–0.01–0.00)	0.40	99%	0.99	0.01
	<i>Examiner 2</i>	0.49	(0.14–0.84)	0.83	98%	0.98	0.02
	<i>Inter-rater</i>	0.00	(–0.01–0.00)	0.25	98%	0.96	0.01
Expanded Agreement							
Thoracic Spine							
	<i>Examiner 1</i>	0.89	(0.67–1.00)	0.89	99%	0.97	0.00
	<i>Examiner 2</i>	0.86	(0.46–1.00)	0.86	99%	0.98	0.01
	<i>Inter-rater</i>	0.62	(–0.06–1.00)	0.62	99%	0.98	0.00
Rib Cage							
	<i>Examiner 1</i>	1.00	(1.00–1.00)	1.00	100%	0.99	0.00
	<i>Examiner 2</i>	0.76	(0.45–1.00)	0.76	99%	0.97	0.01
	<i>Inter-rater</i>	0.76	(–0.63–1.00)	0.88	99%	0.99	0.00

^a κ maximum is the proportions of positive and negative ratings by each examiner expressed relative to the greatest possible agreement by adjusting the distribution of paired ratings.

^bPrevalence index is the absolute difference between the positive and negative agreements, relative to the number of paired ratings.

^cBias index is the absolute value of the difference between the positive and negative disagreements relative to the number of paired ratings.

Agreement between examiners in pain provocation of the rib cage increased from *none* ($\kappa=0.00$) to *good* ($\kappa=0.76$) under the expanded definition.

Discussion

The purpose of this investigation was to assess the intra- and inter-examiner reliability of joint mobility and pain provocation assessment of the thoracic spine and rib cage in asymptomatic adults. In general, the intra- and inter-examiner reliability for joint mobility assessment was *slight* to *fair* under the strict definition of agreement while pain provocation assessment ranged from *no* agreement to *perfect* agreement. When the expanded definition of agreement was applied to adjust for errors due to spinal level identification, the intra-examiner reliability of joint mobility assessment increased to *good* with the inter-examiner reliability considered *moderate*. Similarly, the intra- and inter-examiner reliability of pain provocation increased to *very good* and *good*, respectively, when the expanded definition of agreement was applied.

In a study with a design similar to ours, Christensen et al¹³ included both the strict and expanded definitions of agreement and observed reliability values for the thoracic spine consistent with the current study. While they investigated thoracic levels from T₁ to T₈ in asymptomatic individuals, our results suggest that the inclusion of the lower thoracic spine (T₉–T₁₂) does not affect the examiner reliability. In addition, our findings indicate that examiner reliability of the posterior and anterior rib cage, at least for joint mobility assessment, is comparable to the thoracic spine. Further, the reliability values observed for the thoracic spine are generally consistent with those reported in the cervical and lumbar regions of the spine when similar assessment techniques have been employed^{21–23}.

While Potter et al⁶ found good ($ICC_{(1,1)} = 0.70$) intra-examiner reliability in identifying a joint dysfunction in the thoracic spine of asymptomatic individuals, their examination procedures were more comprehensive than those

used in the present study. In addition to passive accessory motion assessment, active and passive physiological motions about all cardinal planes were considered in the determination of the joint dysfunction location. Thus, the use of PA and AP pressures in addition to other standard motion assessment measures, a combination more typical of routine clinical practice, appears to improve the overall examiner reliability. As the application of force magnitude and direction during PA and AP pressures has been reported to be inconsistent between examiners²⁴, the use of a standardized technique, as was implemented in the current study, is recommended.

Because perfect agreement ($\kappa=1$) may not be achievable in true clinical environments, the maximum attainable kappa provides an additional, and arguably more appropriate, standard for comparison that reflects these pre-existing constraints. In our study, the maximum kappa values were generally below 0.90. Interpreting the kappa values in comparison to their corresponding maximum attainable kappa results in improved reliability for all of the parameters tested. This was especially true for the joint mobility assessment of the thoracic spine and rib cage under the expanded definition of agreement, where the kappa values were consistently within 10% of the maximum attainable kappa. Thus, when one considers the study's pre-existing factors and potential errors in spinal level identification, the actual intra- and inter-rater reliability is quite favorable.

Prevalence and bias of the clinical measure must be considered in the interpretation of kappa, as both can influence its magnitude^{25,26}. The magnitude of kappa will decrease if prevalence is high, or increase if bias is high²⁰. In this study, bias was consistently low and thus likely had minimal effect on the calculated kappa values. However, the prevalence indices were moderate to high, particularly for the pain provocation outcome. This resulted from the limited number of painful joints compared to non-painful joints detected by the examiners. While the use of asymptomatic subjects reduced the likelihood of change

in the joint mobility or pain provocation between examination sessions, it also reduced the number of painful joints present in our sample. Thus, the findings from the present investigation may not be generalizable to specific symptomatic patient populations. Future studies specifically investigating the reliability of pain provocation assessment may opt to enroll individuals with spinal pain as an increased reporting of pain can be expected. It is worth noting that prior investigations involving a more diverse subject pool have reported the reliability of pain provocation to be good for the thoracic spine¹³ and fair for the anterior chest wall¹². Similarly, improved reliability in the detection of joint mobility dysfunctions has also been found when patient populations are examined compared to asymptomatic subjects²⁷.

The limited number of painful joints reported also produced a substantial discrepancy between the percent agreement and kappa values. For example, during pain provocation assessment of the rib cage, the percent agreement for examiner 1 and between examiners was 99% and 98%, respectively; yet the corresponding kappa value was 0.00 for both. This base rate problem²⁸ (high percent agreement and low kappa value) was a result of near perfect agreement in the cases that did not report pain, with the absence of agreement in those few cases where pain was reported. This was also reflected in the high prevalence indices, owing to the difference in positive and negative agreements. The inclusion of subjects known to have painful spinal segments in addition to the asymptomatic subjects that were enrolled would have likely minimized these issues.

The expanded definition of agreement was implemented to account for discrepancies in spinal level determination between examiners and between examination sessions and has been previously used in the reliability of thoracic spine mobility assessment¹³. Recent work has indicated that while experienced physiotherapists are only approximately 50% accurate in identifying a specific lumbar spinous process, they are typically within one spinal level (72%)²⁹. In addition, a recent study

raised questions regarding the accuracy of the long-standing *rule of threes*, a method utilized to identify the corresponding thoracic spine transverse processes with the same-level spinous processes³⁰. Application of the new proposed method for identifying these bony landmarks compared to the *rule of threes* method would result in a stated difference of one spinal level. Thus, the absence of spinal level standardization in the current study would likely have introduced a source of error into the reliability estimates resulting in an underestimation of the true reliability. Further, we acknowledge that the greater agreement rates and kappa values resulting from the expanded agreement may also not reflect the true assessment reliability. Therefore, we chose to report findings based on the strict and expanded definitions of agreement with the expectation that the true reliability lies somewhere in between.

Taking the concept of expanded definition one step further, Hicks et al³¹ collapsed their spinal segmental mobility data for each patient into a dichotomous rating: the presence or absence of segmental hypermobility. Their rationale was that spinal stabilization exercises are typically applied to the spine as a unit as opposed to specific spinal levels. This approach would be consistent with the proposed spinal manipulation clinical prediction rules for patients with low back pain that include spinal segmental hypomobility at any lumbar level as one of the criteria³². These studies illustrate the contribution of multiple examination findings, which by themselves may have less than excellent reliability properties but when clustered can provide predictive clinical decision-making value. The relative strength of our results may lend support to future studies attempting to develop clinical prediction rules for patients more likely to benefit from treatment of the thoracic spine and rib cage regions.

Using within-day repeated measures may have influenced our results. Although the study design minimized the likelihood of change in joint mobility or pain provocation due to daily activities, each motion segment was as-

sessed passively with five consecutive manual pressures applied twice by each examiner. This repeated application of pressures may have mobilized a previously stiff joint resulting in greater joint mobility as the data collection progressed. Further, the perceived magnitude of mobility by the examiner may have been biased by the continual assessment design. That is, the pressure applied by each examiner or the examiner's perception of mobility between the two examination sessions may not have been constant throughout the data collection. Additionally, the repeat pressures to each joint may have also progressively increased (or decreased) local pain provocation, serving as a mechanical source of irritation or relief.

Conclusion

Based on strict definition of agreement, the reliability of joint mobility assessment for the thoracic spine and rib cage ranged from *slight* to *fair*. However, *moderate* to *good* reliability was observed when the expanded definition of agreement was applied. Further, the reliability between two examiners was consistently less than the within examiner reliability. Finally, pain provocation assessment showed similar reliability but should be viewed with caution due to the limited prevalence in the sample investigated.

Acknowledgements

The authors wish to thank Lindsey Dold, MPT and Dean Orvis, PT for their assistance in data collection. This project was supported by an OPTP grant from the American Academy of Orthopaedic Manual Physical Therapists.

REFERENCES

1. American Physical Therapy Association. Guide to Physical Therapist Practice. 2nd ed. *Phys Ther* 2001;81:9-746.
2. Menck JY, Requejo SM, Kulig K. Thoracic spine dysfunction in upper extremity complex regional pain syndrome type I. *J Orthop Sports Phys Ther* 2000;30:401-409.
3. Arroyo JF, Joliet P, Junod AF. Costoverte-

bral joint dysfunction: Another misdiagnosed cause of atypical chest pain. *Postgrad Med J* 1992;68:655-659.

4. Gregory PL, Biswas AC, Batt ME. Musculoskeletal problems of the chest wall in athletes. *Sports Med* 2002;32:235-250.
5. Brismee JM, Gipson D, Ivie D, et al. Interrater reliability of a passive physiological intervertebral motion test in the mid-thoracic spine. *J Manipulative Physiol Ther* 2006;29:368-373.
6. Potter L, McCarthy C, Oldham J. Intraexaminer reliability of identifying a dysfunctional segment in the thoracic and lumbar spine. *J Manipulative Physiol Ther* 2006;29:203-207.
7. Lee M, Steven GP, Crosbie J, Higgs RJ. Variations in posteroanterior stiffness in the thoracolumbar spine: Preliminary observations and proposed mechanisms. *Phys Ther* 1998;78:1277-1287.
8. Abbott JH, Flynn TW, Fritz JM, Hing WA, Reid D, Whitman JM. Manual physical assessment of spinal segmental motion: Intent and validity. *Man Ther* (Nov 7, 2007: Epub, ahead of print).
9. Fritz JM, George S. The use of a classification approach to identify subgroups of patients with acute low back pain: Interrater reliability and short-term treatment outcomes. *Spine* 2000;25:106-114.
10. Fritz JM, Cleland JA, Childs JD. Subgrouping patients with low back pain: Evolution of a classification approach to physical therapy. *J Orthop Sports Phys Ther* 2007;37:290-302.
11. Fritz JM, Delitto A, Erhard RE. Comparison of classification-based physical therapy with therapy based on clinical practice guidelines for patients with acute low back pain: A randomized clinical trial. *Spine* 2003;28:1363-1371, discussion 1372.
12. Christensen HW, Vach W, Manniche C, Haghtelt T, Hartvigsen L, Hoilund-Carlson PF. Palpation for muscular tenderness in the anterior chest wall: An observer reliability study. *J Manipulative Physiol Ther* 2003;26:469-475.
13. Christensen HW, Vach W, Vach K, et al. Palpation of the upper thoracic spine: An observer reliability study. *J Manipulative Physiol Ther* 2002;25:285-292.
14. Huijbregts PA. Spinal motion palpation: A review of reliability studies. *J Man Manip Ther* 2002;10:24-39.
15. Latimer J, Lee M, Adams R, Moran CM. An investigation of the relationship between

- low back pain and lumbar posteroanterior stiffness. *J Manipulative Physiol Ther* 1996;19:587-591.
16. Shirley D, Ellis E, Lee M. The response of posteroanterior lumbar stiffness to repeated loading. *Man Ther* 2002;7:19-25.
 17. Allison G, Edmonston S, Kiviniemi K, Lani-gan H, Simonsen AV, Walcher S. Influence of standardized mobilization on the pos-teroanterior stiffness of the lumbar spine in asymptomatic subjects. *Physiother Res Int* 2001;6:145-156.
 18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
 19. Lantz CA, Nebenzahl E. Behavior and inter-pretation of the kappa statistic: Resolution of the two paradoxes. *J Clin Epidemiol* 1996;49:431-434.
 20. Sim J, Wright CC. The kappa statistic in reli-ability studies: Use, interpretation, and sample size requirements. *Phys Ther* 2005;85:257-268.
 21. Phillips DR, Twomey LT. A comparison of manual diagnosis with a diagnosis estab-lished by a uni-level lumbar spinal block procedure. *Man Ther* 1996;1:82-87.
 22. Maher CG, Latimer J, Adams R. An investi-gation of the reliability and validity of pos-teroanterior spinal stiffness judgments made using a reference-based protocol. *Phys Ther* 1998;78:829-837.
 23. Jull G, Zito G, Trott P, Potter H, Shirley D. Inter-examiner reliability to detect painful upper cervical joint dysfunction. *Aust J Physiother* 1997;43:125-129.
 24. Snodgrass SJ, Rivett DA, Robertson VJ. Manual forces applied during posterior-to-anterior spinal mobilization: A review of the evidence. *J Manipulative Physiol Ther* 2006;29:316-329.
 25. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* 1990;43:551-558.
 26. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two para-doxes. *J Clin Epidemiol* 1990;43:543-549.
 27. Schoensee SK, Jensen G, Nicholson G, Gos-smann M, Katholi C. The effect of mobiliza-tion on cervical headaches. *J Orthop Sports Phys Ther* 1995;21:184-196.
 28. Uebersax JS. A design-independent method for measuring the reliability of psychiatric diagnosis. *J Psychiatr Res* 1982;17:335-342.
 29. Harlick JC, Milosavljevic S, Milburn PD. Palpation identification of spinous processes in the lumbar spine. *Man Ther* 2007;12:56-62.
 30. Geelhoed MA, McGaugh J, Brewer PA, Murphy D. A new model to facilitate palpa-tion of the level of the transverse processes of the thoracic spine. *J Orthop Sports Phys Ther* 2006;36:876-881.
 31. Hicks GE, Fritz JM, Delitto A, Mishock J. Interrater reliability of clinical examination measures for identification of lumbar seg-mental instability. *Arch Phys Med Rehabil* 2003;84:1858-1864.
 32. Childs JD, Fritz JM, Flynn TW, et al. A clin-ical prediction rule to identify patients with low back pain most likely to benefit from spinal manipulation: A validation study. *Ann Intern Med* 2004;141:920-928.