Interrater Reliability of Clinical Examination Measures for Identification of Lumbar Segmental Instability

Gregory E. Hicks, PhD, PT, Julie M. Fritz, PhD, PT, ATC, Anthony Delitto, PhD, PT, FAPTA, John Mishock, DC, PT

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

Objective: To determine the interrater reliability of common clinical examination procedures proposed to identify patients with lumbar segmental instability.

Design: Single group repeated-measures interrater reliability study.

Setting: Outpatient physical therapy (PT) clinic and university PT department.

Participants: A consecutive sample of 63 subjects (38 women, 25 men; 81% with previous episodes of low back pain [LBP]) with current LBP was examined by 3 pairs of raters.

Interventions: Not applicable.

Main Outcome Measures: Repeat measurements of clinical signs and tests proposed to identify lumbar segmental instability.

Results: Kappa values for the trunk range of motion (ROM) findings varied (range, .00–.69). The prone instability test (κ=.87) showed greater reliability than the posterior shear test (κ=.22). The Beighton Ligamentous Laxity Scale (LLS) for generalized ligamentous laxity showed high reliability (intraclass correlation coefficient=.79). Judgments of pain provocation (κ range, .25–.55) were generally more reliable than judgments of segmental mobility (κ range, −.02 to .26) during passive intervertebral motion testing.

Conclusions: The results agree with previous studies suggesting that segmental mobility testing is not reliable. The prone instability test, generalized LLS, and aberrant motion with trunk ROM demonstrated higher levels of reliability.

Key Words: Diagnosis; Low back pain; Rehabilitation; Reliability and validity.

TO DIAGNOSE PATIENTS with low back pain (LBP) according to the traditional medical model, one would have to successfully identify an underlying pathologic mechanism.[1] Clinicians and researchers have had minimal success in identifying specific, structural faults in most cases of LBP. As a result, the diagnosis and subsequent treatment of LBP tends to be lumped into 1 large homogenous group. Several researchers have suggested that the population of patients with LBP is not a homogenous group; rather, the LBP population should be classified into subgroups that share similar characteristics, impairments, and dysfunction.[2,3] Such a classification system could guide diagnosis and treatment and could improve the overall decision-making processes in the management of LBP patients within each subgroup. The identification of valid subgroup populations of patients with LBP has been listed as the top priority of an international forum of primary care researchers on LBP.[4]

One subgroup identified in the literature is patients thought to have lumbar segmental instability (LSI).[5-8] Many clinicians believe that patients with LSI may preferentially respond to a particular rehabilitation approach, and therefore accurately identifying these patients could improve treatment outcomes. Several definitions of LSI have been proposed.[7-10] Recently, Panjabi[7,8] proposed a definition that may provide a useful framework within which to approach the problem. Panjabi[7,8] proposed that the total range of motion (ROM) of the spine consists of the neutral zone and the elastic zone. The neutral zone is the flexible part of the total ROM in which there is minimal resistance to intervertebral motion from passive structures. The elastic zone is near the end ROM where there is significant resistance to motion from passive structures. Furthermore, the spinal stabilizing system is described as consisting of the active and passive subsystems, as well as the neural control unit. The passive component includes the intervertebral disks, ligaments, and facets of the spinal column, whereas the active component includes the muscles surrounding the spinal column. The neural control unit uses kinesthetic input to coordinate the muscles' stabilizing function.[11] From this perspective, LSI can be defined as a decrease in the capacity of the spinal stabilizing system to maintain intervertebral neutral zones within physiologic limits so that there is no major deformity, neurologic deficit, or incapacitating pain.[7,8]

Although LSI is becoming increasingly recognized as an important subgroup of patients with LBP, the identification of reliable and valid clinical diagnostic tools has thus far been elusive. The diagnostic standard for LSI has traditionally centered on identifying excessive translational or rotational movements between lumbar vertebrae by using lateral flexion and extension radiographs.[12,13] Arriving at accurate radiographic diagnostic criteria, however, has been complicated by high false-positive rates and significant variation among asymptomatic persons.[14,15] There may be other factors as well, such as neuromuscular control of spinal movement and midrange motion characteristics, that may also indicate LSI.

Numerous findings from the clinical examination have been proposed as signs of LSI. Some researchers[2,5,16] have focused on historical data such as frequent previous episodes brought on by minimal perturbations or a reduction in pain with previous treatment with bracing. Clinical signs proposed as diagnostic of LSI include palpation of vertebral malalignment and excessive passive intervertebral motion.[17,18] Several researchers[10,18] have described an "instability catch," or other move-

From the Clinical Research Branch, National Institute on Aging, Baltimore, MD (Hicks); Department of Physical Therapy, University of Pittsburgh, Pittsburgh, PA (Fritz, Delitto); and Joyner Sports Medicine Institute, Altoona, PA (Mishock).

Supported in part by the Foundation for Physical Therapy (grant no. P5397).

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the author(s) or upon any organization with which the author(s) is/are associated.

Correspondence to Gregory E. Hicks, PhD, PT, 3001 S Hanover St, Harbor Hospital, 5th Fl, Baltimore, MD 21225, e-mail: *hicksgr@grc.nia.nih.gov.* Reprints are not available from the author.

0003-9993/03/8412-7945$30.00/0
doi:10.1016/S0003-9993(03)00365-4

ment alterations observed during active trunk ROM, that may indicate LSI. Several special tests for LSI have been described, including the posterior shear[2] and prone instability tests.[19,20] It has also been suggested that patients with general ligamentous laxity may be at increased risk for LSI,[2] and therefore the finding of generalized laxity in a person with LBP may increase the suspicion of LSI. Little is known about the reliability of any of these tests for LSI. The purpose of the present study was to evaluate the interrater reliability of clinical diagnostic tests that are commonly advocated for use in the evaluation of patients suspected of having LSI.

METHODS

Participants

Subjects with current complaints of LBP without radiation of symptoms below the knee were recruited for this study. Patients were excluded if their LBP could be attributed to current pregnancy, acute fracture, tumor, or infection. Previous lumbar surgical fusion was also an exclusion criterion. To assess the generalizability of the results, subjects were recruited from 2 different sources, and 3 pairs of raters were used. Subjects were recruited either as consecutive participants in research studies on LBP or as patients referred to an outpatient physical therapy (PT) clinic. Subjects recruited from the research study populations were evaluated by rater pairs 1 and 2; subjects from the clinical population were evaluated by rater pair 3. Subjects evaluated by rater pair 1 and rater pair 3 were recruited solely to evaluate test-retest reliability. Subjects evaluated by rater pair 2 were recruited as part of a larger project examining outcomes of a stabilization exercise program. A description of the subjects evaluated by each rating pair is in table 1. Before participating in the study, each subject read and signed a consent form approved by the University of Pittsburgh Health Sciences Institutional Review Board.

Raters

Four physical therapists participated as raters. Each rater routinely used the clinical tests included in this study in clinical practice or in academic instruction. Rater 1 (PT1) was a physical therapist and chiropractor with 3 years of experience as a chiropractor and 2 years as an orthopedic physical therapist.

Table 1: Characteristics of the Subjects

Variable	Entire Sample (N=63)	Rater Pair 1 (n=20)	Rater Pair 2 (n=28)	Rater Pair 3 (n=15)
Age (y)				
Mean ± SD	36.0±10.3	36.8±10.5	32.8±8.5	37±12.1
Range	20–66	20–51	22–59	20–66
Gender				
Male	25	9	10	6
Female	38	11	18	9
Prior LBP episodes?*				
Yes	51	20	20	11
No	12	0	8	4
Oswestry score				
Mean ± SD	17.8±11.3	13.5±8.5	12.6±10.3	28.5±10.8
Range	92-52	2-32	2-38	14-52

Abbreviations: Oswestry, Oswestry Disability Questionnaire[39]; SD, standard deviation.
*Episode is a period of LBP during which the subject must modify daily activities due to pain.

Table 2: Clinical Examination Measures Proposed to Identify LSI

Observations of trunk AROM
Painful arc in flexion
Painful arc on return from flexion
Instability catch
Gower sign
Reversal of lumbopelvic rhythm
Aberrant movement pattern
Special tests
Posterior shear test
Prone instability test (fig 1)
Beighton LLS
Passive intervertebral motion testing
Segmental mobility judgment
Pain provocation judgment

The second rater (PT2) was a physical therapist with 6 years of experience in an orthopedic setting. The third rater (PT3) was an orthopedic physical therapist with 8 years of experience. Rater 4 (PT4) was a physical therapist with 4 years of experience in the orthopedic environment. Three rater pairs were used during the study, paired as follows: rater pair 1 included PT1 and PT2; rater pair 2 included PT2 and PT3; and rater pair 3 consisted of PT1 and PT4. Training of the raters consisted of (1) a group review of operational definitions for each evaluative procedure, and (2) a single 1-hour practice session together. During the practice session, raters performed the diagnostic tests on each other and on several PT students to ensure that procedures would be performed in the same manner by each rater. No further training was implemented.

Procedures

The same testing protocol was used for all subjects regardless of recruitment source. For each pair of raters, the first rater performed all the clinical examination measures on each subject. The second rater, who was blinded to the results of the first evaluation, then performed the same examination procedures. The clinical examination included the assessment of various movement aberrations during lumbar active ROM (AROM), 2 special tests for LSI (posterior shear and prone instability tests), the Beighton Ligamentous Laxity Scale (LLS), and assessment of passive intervertebral motion in the prone position (table 2). A minimum 15-minute time delay between evaluations was used to minimize the chance that the patient's clinical presentation may have changed as a result of repeat evaluation procedures, which would confound the interpretation of the reliability coefficients. The operational and procedural definitions and the grading criteria for each clinical examination measure are described in appendix 1.

Data Analysis

A κ statistic[21] was used to calculate interrater reliability for all diagnostic tests considered to have ordinal level measurements, which included all measures except the Beighton LLS. The κ statistic represents the percentage agreement beyond chance between raters and is therefore the appropriate statistic for this purpose. A weighted κ statistic[22] was used to calculate the reliability for the passive intervertebral motion tests. The weighted κ is the appropriate statistic for use when disagreements of varying degrees are to be weighted accordingly. Equal weights were assigned to each interval. An intraclass correlation coefficient (ICC), model 1,1,[23] was used to determine the interrater reliability for the Beighton LLS, which was analyzed

Table 3: Reliability Coefficients for Clinical Measures Used in the Identification of Patients for the LSI Subgroup

Variable	Reliability Coefficient, κ (95% CI)	Percentage Agreement	Distribution of Ratings (negative/positive)	
			Rater 1	Rater 2
Painful arc in flexion	.69 (.54–.84)	92%	53/10	54/9
Painful arc on return from flexion	.61 (.44–.78)	90%	54/9	54/9
Instability catch	.25 (−.10 to .60)	92%	61/2	58/5
Gower sign	.00 (−1.09 to 1.09)	98%	63/0	62/1
Reversal of lumbopelvic rhythm	.16 (−.15 to .46)	87%	61/2	55/8
Aberrant movement pattern	.60 (.47–.73)	84%	45/18	47/16
Posterior shear test	.35 (.20–.51)	74%	42/21	52/11
Prone instability test	.87 (.80–.94)	91%	36/27	36/27
	$ICC_{1,1}$		Mean ± SD (Range)	Mean ± SD (Range)
Beighton LLS	.79 (.68–.87)	—	1.37±1.86 (0–6)	1.46±1.87 (0–7)

as a continuous variable. Reliability coefficients with 95% confidence intervals (CIs) and percentage agreement were calculated for the entire subject sample for each clinical test. Furthermore, percentage agreements and reliability coefficients with 95% CI were calculated for each of the 3 rater pairs. The extent of overlap of the CIs for each rater pair were examined to assess the consistency of reliability judgments among the different pairs of examiners.

RESULTS

The reliability coefficients with corresponding 95% CIs and percentage agreements for the observations of trunk ROM, the special tests, and generalized LLS for the entire sample are in table 3. The κ values for the observations of trunk ROM ranged from .00 to .69. During the study, we noted that making distinctions among the different categories was sometimes difficult and that prevalence was often low; therefore, we decided to collapse all 5 observational elements into a single category. The new category, aberrant movement pattern during trunk flexion, was defined as positive in the presence of any of the 5 observational elements and negative in the complete absence of all elements. The κ value for this new category was .60 (95% CI, .43–.73).

The results for the 2 components of passive intervertebral motion testing are reported in table 4. The weighted κ values for mobility testing were low for each lumbar segment (range, −.02 to .26). Pain provocation judgments demonstrated higher κ values at each spinal level (range, .25–.55). We hypothesized that the low interrater reliability for the mobility judgments might be due in part to difficulty in accurately locating and naming the lumbar level and not wholly due to the judgment of mobility itself. Because the treatment for LSI is generally not directed at a specific spinal level, but at the entire lumbar spine, we collapsed the segmental mobility tests into a dichotomous rating for each subject. The subject was rated either as hypermobile (at least 1 segmental level in the entire lumbar spine rated as hypermobile) or not hypermobile (no lumbar levels rated as hypermobile). The same procedure was followed to examine the reliability of ratings for the presence of any hypomobility in the lumbar spine versus no hypomobility. Reliability of this dichotomous judgment of hypermobility was also low (κ=.30; 95% CI, .13–.47); however, the percentage agreement was fairly high (76%). Reliability of the judgment of any hypomobility was also low (κ=.18; 95% CI, .05–.32) (table 5).

The reliability coefficients and percentage agreements for each rater pair are in table 6. The reliability coefficients were generally consistent across the rater pairs for each clinical measure, as noted by the overlap of CIs. In the case of the prone instability test (κ=1.00; 95% CI, 1.00–1.00), the extremely narrow 95% CI around the reliability coefficient for rater pair 1 did not overlap with the CIs for the other rater pairs. This is because there was perfect agreement between raters in pair 1.

DISCUSSION

Although we examined the reliability of several aspects of the clinical examination that have been proposed as useful in the identification of patients with LSI, we did not examine the validity of these findings, and, therefore, no conclusions can be drawn about the diagnostic accuracy of these tests. Although no globally acknowledged standards for interpretation of reliability coefficients exists, Landis and Koch[24] suggested the following interpretation for the Cohen κ statistic: less than 0.0 is poor; 0.0 to .20 is slight; .21 to .40 is fair; .41 to .60 is moderate; .61 to .80 is substantial; and .81 to 1.0 is almost perfect. The first group of tests we examined included 5 items based on the observation of trunk AROM. Judgments of a

Table 4: Reliability of Segmental Mobility and Pain Provocation Judgments

Variable	Mobility,* κ (95% CI)	Percentage Agreement	Distribution of Ratings (hypo/norm/hyper)		Provocation, κ (95% CI)	Percentage Agreement	Distribution of Ratings (positive/negative)	
			Rater 1	Rater 2			Rater 1	Rater 2
L1	.26 (−.01 to .53)	68%	17/44/2	9/50/4	.36 (.12–.59)	87%	7/56	7/56
L2	.17 (−.13 to .47)	69%	15/46/2	8/52/3	.45 (.26–.63)	85%	8/55	11/52
L3	−.02 (−.25 to .28)	52%	17/37/9	4/54/5	.30 (.12–.47)	76%	15/48	12/51
L4	.11 (−.26 to .35)	58%	29/38/5	6/52/5	.25 (.11–.40)	65%	26/37	20/43
L5	.18 (−.03 to .49)	65%	12/41/10	6/53/4	.55 (.43–67)	78%	42/21	52/11

Abbreviations: hyper, above; hypo, below; norm, normal.
*Weighted k.

Table 5: Reliability of Dichotomous Assessment of Segmental Mobility Judgment

Variable	Reliability Coefficient, κ (95% CI)	Percentage Agreement	Distribution of Ratings (no/yes)	
			Rater 1	Rater 2
Any hypermobility in the lumbar spine?	.30 (.13–.47)	76%	47/16	52/11
Any hypomobility in the lumbar spine?	.18 (.05–.32)	59%	31/32	49/14

painful arc in flexion and on return from flexion both demonstrated substantial agreement (κ values, .69 and .61). The other observations associated with trunk AROM (Gower sign, instability catch, reversal of lumbopelvic rhythm) demonstrated poor to fair reliability. One reason for low reliability was the low prevalence of these observations in the sample. For example, the Gower sign was identified only once by 1 rater out of 126 possible ratings (see table 4). Minimal variability in ratings leads to a high percentage of chance agreement between examiners and low κ values.[24] To overcome this problem, we collapsed the observations into a single category called aberrant movement during trunk flexion and found a κ value of .60. This collapsed category overcomes the difficulties presented by the low prevalence of some findings and the difficulty in distinguishing between observations that are often similar in appearance. We believe this is an acceptable approach because each type of variation in trunk ROM is thought to indicate an inability of the subject to adequately control lumbar ROM in a manner that does not produce symptoms. Clinically, the observation of any of these aberrant movements may lead to a similar decision to categorize the patient as a member of the LSI subgroup and to consider the inclusion of stabilization exercises in the treatment program. Presently, no evidence exists to suggest that any of these observations is more diagnostic or prognostic than another; however, if such evidence were forthcoming, then analyzing these observations separately would be required.

We examined the interrater reliability of 2 special tests: the posterior shear and prone instability tests. The posterior shear test had an adequate ratings distribution but showed only fair reliability (κ=.35). The prone instability test (fig 1) demonstrated almost perfect reliability (κ=.87), and the narrow width of the CI suggests that the κ value is fairly precise. The high reliability coefficients and relatively narrow CIs found in each of the individual rater pairs for the prone instability test further strengthen the notion that this test is generalizable to a broad spectrum of clinicians as a reliable tool. The prone instability test is based on the hypothesis that if pain is present on passive provocation testing but disappears when the patient activates the spinal extensors, then the muscle activity must be able to effectively stabilize the segment, thereby indicating the presence of LSI (fig 1, step 2). Further testing on the validity of this test appears warranted based on the high degree of reliability achieved.

The Beighton LLS for generalized ligamentous laxity also showed substantial interrater reliability (ICC=.79). We are not aware of other reports on the reliability of this scale in the literature. Although the Beighton LLS has been associated with an increased risk for musculoskeletal injury,[25,26] a link to LBP, or the particular condition of LSI, has yet to be investigated.

Several other investigators[27-30] have examined the reliability of passive intervertebral motion testing. Gonnella et al,[28] who used a 7-point mobility scale, performed segmental mobility testing with the subject in the side-lying position and reported poor agreement between examiners, although no reliability coefficients were presented. Maher and Adams[29] used an 11-point scale for grading segmental mobility and pain provocation. Testing was performed with the patient prone, and the authors reported greater interrater reliability for the pain provocation tests (ICC range, .67–.72) than for the mobility tests (ICC range, .03–.37).[29] Binkley et al[27] used a 9-point scale and also found poor interrater reliability for segmental mobility testing performed in the prone position (ICC=.25). Our results are in agreement with previous studies that have found poor interrater reliability for segmental mobility testing. Our results also support the finding of Maher and Adams[29] that pain provocation is a more reliable judgment than segmental mobility.

Previous studies have used 7- to 11-point scales for judging segmental mobility. We chose to use a 3-point scale (hypomobile-normal-hypermobile). Reducing the number of potential ratings will tend to diminish reliability; however, we felt that this scale more accurately reflected judgments made in clinical practice. If a clinician judges a segment as hypomobile, some type of mobilization treatment may be indicated. Conversely, if a segment is judged as hypermobile, a stabilization treatment approach may be used. Gradations within each judgment are

Table 6: Reliability Coefficients for Clinical Measures Used to Identify Patients for Each Rater Pair

Variable	Rater Pair 1 (n=20)		Rater Pair 2 (n=28)		Rater Pair 3 (n=15)	
	Reliability Coefficient, κ (95% CI)	Percentage Agreement	Reliability Coefficient, κ (95% CI)	Percentage Agreement	Reliability Coefficient, κ (95% CI)	Percentage Agreement
Painful arc in flexion	.77 (.47–1.00)	90%	.65 (.40–.90)	96%	.42 (.22–.63)	87%
Painful arc on return from flexion	.63 (.50–.75)	85%	.65 (.40–.90)	96%	.42 (.22–.63)	87%
Instability catch	.35 (.13–.56)	85%	.00 (−.72 to .72)	92%	.00 (−.52 to .52)	93%
Gower sign	.00 (−.60 to .60)	95%	UTC	100%	UTC	100%
Reversal of lumbopelvic rhythm	.00 (−.24 to .24)	80%	.16 (−.15 to .46)	87%	.42 (.22–.63)	87%
Posterior shear test	.23 (.02–.44)	80%	.31 (.14–.48)	75%	.39 (.27–.51)	67%
Prone instability test	1.00 (1.00–1.00) $ICC_{2,1}$	100%	.81 (.80–.94) $ICC_{2,1}$	93%	.74 (.64–.83) $ICC_{2,1}$	87%
Beighton LLS	.95 (.88–.98)	—	.76 (.54–.88)	—	.66 (.24–.87)	—

Abbreviation: UTC, unable to calculate due to no variability in ratings.

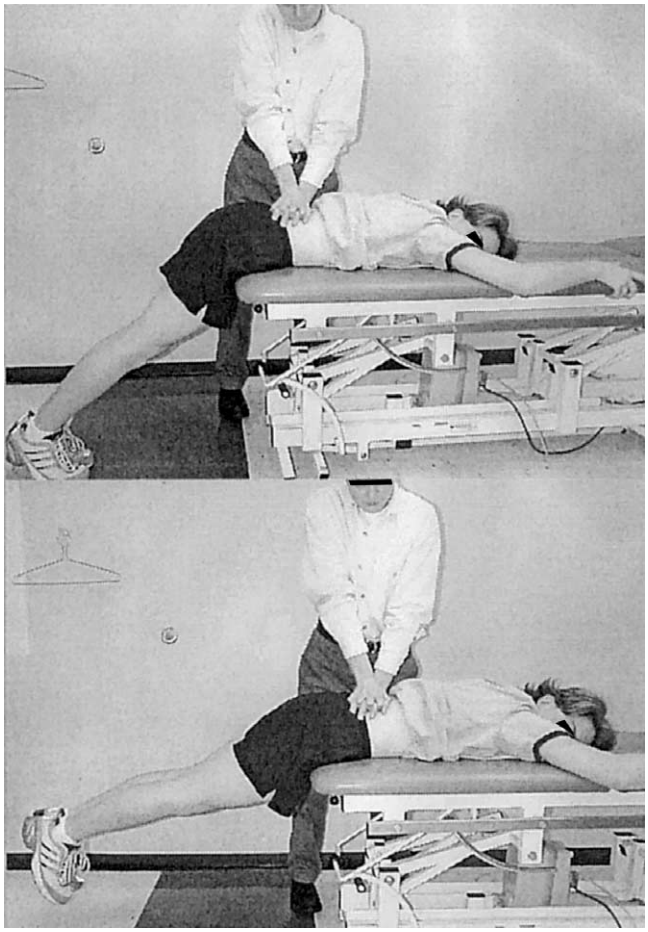

Fig 1. Prone instability test.

difficult to assess,[27,29] and it is uncertain whether they influence clinical decision making. Strender et al[30] used a similar 3-level grading scale to evaluate segmental mobility at L4 and L5 from the side-lying position and reported κ values of .75 and .54, respectively.

One possible explanation for the poor reliability of segmental mobility testing is difficulty in identifying the level being tested.[27,31] Binkley et al[27] examined the interrater reliability of therapists determining the level of a marked spinous process and found that disagreements were common. In addition, Maher and Adams[29] controlled for the problem of disagreements on naming the spinal level by having 1 rater label the spinal levels before either rater assessed the patient to ensure that the same levels were being assessed. Even with this level of control, they found that the reliability coefficients were still low (ICC range, .03–.37) for assessment of segmental mobility. However, their judgments of pain provocation were relatively high (ICC range, .67–.73),[29] which may indicate that disagreements in judgments of pain provocation are often attributable to errors in naming the spinal level.

We attempted to eliminate the potential for level identification problems to deflate the reliability of segmental mobility judgments. Stabilization treatment approaches used for patients with LSI are typically not directed toward a particular spinal level but are applied to the lumbar spine as a whole. Therefore, when attempting to identify patients with LSI, the most important judgment would appear to be that the patient has at least 1 spinal level that is hypermobile. To assess the reliability of this level of decision making, we collapsed the segmental mobility tests for each patient into a dichotomous rating; the patient either did or did not give some evidence of having hypermobility. Interrater reliability was fair (κ=.30) for this judgment, although the percentage agreement between the examiners was relatively high (76%), indicating that the low prevalence of hypermobile ratings (21% overall) may have deflated the reliability coefficient.

Besides the problem of level identification and low prevalence, several other issues may have played a role in the low reliability found in segmental mobility judgments. Examiner training could be an issue. Each examiner had significant experience in using segmental mobility testing as an assessment tool for patients with LBP; however, the only training they underwent in this study was reading operational definitions and practicing the technique in a 1-hour training session. More intensive training might improve reliability; however, Binkley[27] found no differences in agreement between examiners with or without advanced manual therapy training. Moreover, the implementation of more intensive training would certainly limit the generalizability of this assessment tool to only those clinicians with extremely focused training in this area, thereby limiting its usefulness in the clinical setting.

Another issue that may be related to the low reliability of segmental mobility judgment was the use of the prone position for testing. Typically, when joint mobility is assessed in other regions of the body, the joint is initially tested in its resting position rather than at the end ROM.[20] However, when a patient is lying in the prone position, the spine is more likely to be extended, which may interfere with the assessment of segmental mobility. Strender et al[30] found higher κ values when testing from the side-lying position. Further examination of this side-lying technique may prove useful. A final issue is that the construct of segmental mobility itself may be inherently unstable, in which case little can be done to reduce the error, and alternative methods of identifying patients with LSI are required.

The clinical measures examined in the present study have been proposed to give clinicians some level of diagnostic information regarding the presence or absence of LSI. Other clinical measures may be developed for identifying LSI; however, the present study was a first step in examining the technical efficacy of these clinical measures for identifying LSI. We found several clinical measures to be highly reliable: the prone instability test, the Beighton LLS, and aberrant movement patterns during trunk flexion (observations of trunk AROM collapsed into 1 category). Because these clinical measures have a high level of reliability, clinicians should expect to get similar results every time these tests are performed on patients with LBP whose condition is stable. If clinicians are to confidently make treatment decisions based on test results, then the test must offer consistent results every time it is performed. With knowledge of their test-retest reliability, the next step in the evaluation of these clinical measures is to validate them as being truly diagnostic of LSI. The problem is that no true external reference or criterion standard exists by which we can accurately and definitively diagnose LSI. Typically, the condition of LSI is presumed to be present in certain patients and is treated with a lumbar stabilization program. In cases where there is no true criterion standard to make a definitive diagnosis, treatment outcome can viably be used as the criterion standard. One could argue that one of the most important functions of a clinical test is its ability to predict treatment outcome.[32] If any of these clinical measures for LSI allows us to accurately identify people who respond positively to a lum-

bar stabilization program, this feature may help us focus on those people who truly have LSI.

CONCLUSION

Observation of movement patterns during trunk AROM can be reliable, particularly when the separate components are collapsed into 1 category. The Beighton LLS, as a measure of generalized ligamentous laxity, proved highly reliable, as did the prone instability test. Similar to other studies, we found poor reliability for judgments of passive segmental mobility and better reliability for judgments of pain provocation. The error related to judgments of pain provocation may be related to naming the spinal level being assessed.

In light of the diagnostic process, determining reliability is an important precursor to test validation, especially if the test is used to guide patient management strategies. Now that we know the test-retest reliability of these diagnostic tests, the next step is to look at their construct validity in terms of their ability to identify the LSI population correctly. Once this population is readily identifiable, further work can be performed in the area of treatment intervention for this classification subgroup.

APPENDIX 1: OPERATIONAL DEFINITIONS OF CLINICAL EXAMINATION MEASURES

OBSERVATION OF TRUNK AROM

In a standing position, the subject was asked to flex the trunk forward as far as possible while the examiner observed in an effort to identify any of the following abnormalities:

1. Painful arc in flexion: symptoms felt during the movement at a particular point in the motion (or through a particular portion of the range) that are not present before or after this point.[33]
2. Painful arc on return: symptoms occur only during return from the flexed to the erect position.[33]
3. Gower sign ("thigh climbing"): pushing on the thighs or another surface with the hands for assistance during return from the flexed to the erect position.[2]
4. Instability catch: any sudden acceleration or deceleration of trunk movement or movement occurring outside the primary plane of motion (eg, lateral bending or rotation during trunk flexion).[34,35]
5. Reversal of lumbopelvic rhythm: on attempting to return from the flexed position, the patient bends the knees and shifts the pelvis anteriorly before returning to the erect position.[36]

Aberrant Movement Pattern During Active Trunk Flexion

As listed above, there are 5 possible movement patterns that may be seen during trunk flexion, and if any one is present, then the score is positive. If none of the patterns is present, then the score is negative.

Generalized LLS[26,37]

Generalized ligamentous laxity was assessed on a 9-point scale described by Beighton and Horan.[37] The Beighton LLS purportedly identifies persons with generalized ligamentous laxity and defines a broader population at risk for instability problems throughout the musculoskeletal system.[26,38] Four tests are assessed separately on the right and left side, and a point is given for each test the subject can perform. The bilateral tests are passive hyperextension of the elbow greater than 10°, passive hyperextension of the fifth finger to greater than 90°, passive abduction of the thumb to contact the forearm, and passive hyperextension of the knees greater than 10°. The final test is the ability to flex the trunk and place both hands flat on the floor without flexing the knees. The range of possible scores is 0 to 9, with higher scores indicating greater laxity.

Passive Intervertebral Motion Testing[17]

With the subject in the prone position, segmental mobility testing is performed by placing the hypothenar eminence of the testing hand over the spinous process of the segment to be tested. With the elbow and wrist of the testing hand extended, the examiner applies a gentle, but firm, anteriorly directed pressure on the spinous process. Two judgments are made at each spinal level: segmental mobility and pain provocation.

Segmental mobility. Judgment is based on the passive mobility of the tested spinal segment relative to adjacent segments and the expectation of the examiner. One of the following 3 options may be selected for each spinal level:

1. Hypermobility: more motion than normally expected is found between the tested level and the adjacent segments.
2. Normal mobility: passive motion of the spinal level is within normally expected limits.
3. Hypomobility: less motion than normally expected is found between the tested level and the adjacent segments.

Pain provocation. Pain response to manually directed pressure is recorded as 1 of the following options:

1. No pain: no painful symptoms are produced with segmental testing.
2. Pain: segmental testing provokes pain either locally or distally. Local pain refers to pain produced directly under the examiner's hand, whereas distal pain refers to provocation at an anatomic area not directly under the examiner's hand.

Special Tests

Two different special tests were performed: the posterior shear test and the prone instability test.

Posterior shear test.[2] The subject stands with his/her arms crossed over the lower abdomen. The examiner stands at 1 side of the subject and places 1 arm around the subject's abdomen, over the subject's crossed hands. The heel of the examiner's opposite hand is placed on the subject's pelvis for stabilization while the index or middle finger palpates the L5-S1 interspace. The examiner produces a posterior shear force through the subject's abdomen and an anterior stabilizing force with the opposite hand. The test is repeated at each lumber level. A positive test occurs when symptoms are provoked and is not based on the amount of intersegmental motion detected.

Prone instability test.[19,20] The subject is prone with the torso on the examining table and legs over the edge with the feet resting on the floor. While the subject rests in this position, the examiner performs passive intervertebral motion testing as previously described. The patient is asked to report any provocation of pain. The subject then lifts the legs off the floor (hand-holding to the table may be used to maintain position), and the passive intervertebral motion testing is reapplied to any segments that were identified as painful. A positive test occurs when pain is provoked during the first part of the test but disappears when the test is repeated with the legs off the floor.

References

1. Engel G. The need for a new medical model: a challenge for biomedicine. Science 1977;196:129-35.
2. Delitto A, Erhard RE, Bowling RW. A treatment based classification approach to low back syndrome: identifying and staging patients for conservative treatment. Phys Ther 1995;75:470-89.
3. McKenzie RA. The lumbar spine: mechanical diagnosis and therapy. Waikanae (New Zealand): Spinal Publications; 1989.

4. Borkan JM, Koes B, Reis S, Cherkin DC. A report from the second international forum for primary care research on low back pain: reexamining priorities. Spine 1998;23:1992-6.
5. Farfan HF, Gracovetsky S. The nature of instability. Spine 1984; 9:714-9.
6. Frymoyer JW, Akeson W, Brandt K, et al. Clinical perspectives. In: Frymoyer JW, Gordon SL, editors. New perspectives on low back pain. Park Ridge (IL): American Academy of Orthopedic Surgeons; 1989. p 222-30.
7. Panjabi MM. The stabilizing system of the spine. Part I. Function, dysfunction, adaptation, and enhancement. J Spinal Disord 1992; 5:383-9.
8. Panjabi MM. The stabilizing system of the spine. Part II. Neutral zone and instability hypothesis. J Spinal Disord 1992;5:390-7.
9. Pope MH, Frymoyer JW, Krag MH. Diagnosing instability. Clin Orthop 1992;Jun(279):60-7.
10. Kirkaldy-Willis WH, Farfan HF. Instability of the lumbar spine. Clin Orthop 1982;May(165):110-23.
11. Panjabi MM. Low back pain and spinal instability. In: Weinstein JN, Gordon SL, editors. Low back pain: a scientific and clinical overview. Rosemont (IL): American Academy of Orthopedic Surgeons; 1996. p 367-84.
12. Dupuis PR, Yong-Hing K, Kassidy JD, Kirkaldy-Willis WH. Radiographic diagnosis of degenerative lumbar spinal instability. Spine 1985;10:262-76.
13. Posner I, White AA III, Edwards WT, Hayes WC. A biomechanical analysis of the clinical stability of the lumbar and lumbosacral spine. Spine 1982;7:374-89.
14. Hayes MA, Howard TC, Gruel CR, Kopta JA. Roentgenographic evaluation of the lumbar spine flexion-extension in asymptomatic individuals. Spine 1989;14:327-31.
15. Panjabi MM, Lydon C, Vasavada A, Grob D, Crisco JJ, Dvorak J. On the understanding of clinical instability. Spine 1994;19:2642-50.
16. Boden SD, Frymoyer JW. Segmental instability: overview and classification. In: Frymoyer JW, editor. The adult spine: principles and practice. 2nd ed. Philadelphia: Lippincott-Raven; 1997. p 2137-55.
17. Maitland GD. Vertebral manipulation. 5th ed. Oxford: Butterworth Heinemann; 1986. p 74-6.
18. Paris SV. Physical signs of instability. Spine 1985;10:277-9.
19. Magee DJ. Orthopaedic physical assessment. 3rd ed. Philadelphia: WB Saunders; 1997. p 399.
20. Wadsworth CT, DiFabio R, Johnson D. The spine. In: Wadsworth CT, editor. Manual examination and treatment of the spine and extremities. Baltimore: Williams & Wilkins; 1988. p 70-1.
21. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
22. Cohen J. Weighted kappa. Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968; 70:213-20.
23. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-6.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
25. Al-Rawi Z, Nessan AH. Joint hypermobility in patients with chondromalacia patellae. Br J Rheumatol 1997;36:1324-7.
26. Krivickas L, Feinberg J. Lower extremity injuries in college athletes: relation between ligamentous laxity and lower extremity muscle tightness. Arch Phys Med Rehabil 1996;77:1139-43.
27. Binkley JM, Stratford PW, Gill C. Interrater reliability of lumbar accessory motion mobility testing. Phys Ther 1995;75:786-95.
28. Gonnella C, Paris SV, Kutner M. Reliability in evaluating passive intervertebral motion. Phys Ther 1982;62:436-44.
29. Maher C, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. Phys Ther 1994;74: 801-11.
30. Strender LE, Sjoblom A, Sundell K, Ludwig R, Taube A. Interexaminer reliability in physical examination of patients with low back pain. Spine 1997;22:814-20.
31. Shields RK. Invited commentary. Phys Ther 1994;74:809-10.
32. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine 2nd ed. Boston: Little, Brown; 1992.
33. Cyriax JH. Diagnosis of soft tissue lesions. In: Cyriax JH. Textbook of orthopaedic medicine. Volume 1: Diagnosis of soft tissue lesions 6th ed. Baltimore: Williams & Wilkins; 1976. p 389.
34. Nachemson A. Lumbar spine instability: a critical update and symposium summary. Spine 1985;10:290-1.
35. Ogon M, Bender BR, Hooper DM, et al. A dynamic approach to spinal instability. Part II. Hesitation and giving-way during interspinal motion. Spine 1997;22:2859-66.
36. MacNab I, McCulloch J. Backache. 2nd ed. Baltimore: Williams & Wilkins; 1990. p 159.
37. Beighton P, Horan F. Orthopedic aspects of Ehlers-Danlos syndrome. J Bone Joint Surg Am 1969;51:444-53.
38. Beighton P, Solomon L, Soskolne CL. Articular mobility in an African population. Ann Rheum Dis 1973;32:413-8.
39. Fairbank JC, Couper J, Davies JB, O'Brien JP. The Oswestry low back pain disability questionnaire. Physiotherapy 1980;66: 271-3.