Research Report

An Investigation of the Reliability and Validity of Posteroanterior Spinal Stiffness Judgments Made Using a Reference-Based Protocol

Background and Purpose. The reliability and criterion-related validity of ratings of posteroanterior (PA) spinal stiffness made using reference values for comparison have not been investigated. In this study, mechanical reference stimuli for points on an 11-point rating scale were used to determine whether using a reference scale may be feasible. **Subjects.** Five different raters took part in 2 studies in which they rated 40 subjects who were asymptomatic for low back pain. **Methods.** The interrater reliability of ratings was evaluated with intraclass correlation coefficients (ICCs) and standard errors of the measurement (SEMs). Criterion-related validity was evaluated by correlating judgments of PA spinal stiffness assessed manually with measurements of PA spinal stiffness provided by a mechanical device, the "Stiffness Assessment Machine" (SAM). **Results.** Although the reliability indices were generally high, with ICCs reaching .77 and with SEMs as low as 0.72 points, the evidence for criterion-related validity (ie, the ability of the examiner to judge spinal stiffness levels) was not strong, with correlations reaching only .56. **Conclusion and Discussion.** The reference-based protocol allows for more reliable measures of PA stiffness judgments than previous protocols have; however, the human ratings are not highly correlated with the SAM measures. The protocol will have clinical value if judgments made using it are shown to be reliable in clinically relevant subjects and to have validity for clinical management of patients. [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.]

Key Words: *Lumbar spine, Palpation skills, Tests and measurements.*

Christopher G Maher

Jane Latimer

Roger Adams

The motion of the human spine is sometimes assessed with the posteroanterior (PA) pressure test.[1] This test is performed by applying an anteriorly directed force over the spinous process of the prone patient.[1] During this test, the clinician assesses the movement produced and notes any pain reported by the patient. One common method of describing the movement is in terms of its stiffness or the slope of the force-displacement curve.[2(p520)] Jull[2(p522)] suggested that the results from the test can be used by therapists to assist in formulating clinical diagnoses and for identifying which spinal level or levels to treat. Recently, it has been demonstrated that patients with nonspecific low back pain (LBP) have increased PA spinal stiffness, as compared with when they have little or no pain,[3] a finding that provides support for the relationship between LBP and PA spinal stiffness proposed by several manual therapy authors.[1,4,5]

Despite the widespread use of the PA pressure test by physical therapists, studies[6–9] have demonstrated that the stiffness judgments made with this test have poor interrater reliability. One potential cause of this poor reliability is that the current protocol for PA spinal stiffness testing does not specify the manner in which the test should be performed. Thus, it is possible that differences in the way the test is performed cause this poor reliability. One approach to maximizing the reliability of stiffness assessments would be to standardize the factors that have been shown to affect the measured PA stiffness of the spine as well as those factors that have been shown to affect the therapist's perception of stiffness.

Lee and colleagues[10] have used instruments to measure the PA stiffness of the spine and have noted that such measurements are affected by the position of the subject during stiffness testing; the subject's breathing pattern; the presence of spinal muscle activity; the plinth surface on which the measurement is performed; and the frequency, magnitude, and number of loading cycles used. Maher and Adams[11,12] have shown that the grip type adopted, and whether the test is performed with the eyes open or closed, will affect the perceived magnitude of stiffness stimuli provided by a spring, although the actual stiffness remains unchanged.

Another explanation for the low reliability for stiffness judgments may relate to the fact that the scales used to assess PA spinal stiffness typically require the rater to make a judgment as to what the therapist believes is "normal" or "average." In addition, definitions for stiffness are typically not provided, so raters may be attending to characteristics other than stiffness such as friction or viscosity,[13] both of which could provide resistance to movement but are distinct from stiffness. Providing a reference stiffness stimulus that defines both the characteristic to be measured (stiffness) and the magnitude of stiffness that corresponds to a given point on the rating scale could improve reliability and accuracy.

The aims of the 2 studies we conducted were (1) to evaluate the interrater reliability of measurements obtained with 2 new PA spinal stiffness rating methods and (2) to evaluate the criterion-related validity of these ratings. These protocols differed from the existing protocol by our attempt to standardize a range of factors known to affect measured and perceived stiffness and to provide reference stimuli for points on the PA spinal stiffness rating scale. Because subjects without symptoms were examined, the studies did not set out to demonstrate the reliability or validity of the stiffness measures as they are used in clinical practice. Rather, the studies evaluated the potential of this approach to improve the quality of the measures.

Method

Overview

The 2 studies (A and B) were conducted sequentially to examine the reliability and criterion-related validity (ie, whether the examiners' measurements reflected actual stiffness) of 2 new protocols for manual PA spinal stiffness testing. In study A, we used a protocol in which a standard mechanical stiffness target, designated as normal stiffness on the stiffness rating scale, was pro-

CG Maher, PhD, PT, MMPAA, is Lecturer, School of Physiotherapy, Faculty of Health Sciences, The University of Sydney, East Street, PO Box 170, Lidcombe, New South Wales, Australia 2141 (c.maher@cchs.usyd.edu.au). Address all correspondence to Dr Maher.

J Latimer, GradDipAppSc(Manip Phty), PT, MMPAA, is Lecturer, School of Physiotherapy, Faculty of Health Sciences, The University of Sydney.

R Adams, PhD, is Senior Lecturer, School of Physiotherapy, Faculty of Health Sciences, The University of Sydney.

The data collection protocol was approved by The University of Sydney Human Ethics Committee.

This project was supported, in part, by a grant from the Physiotherapy Research Foundation.

This article was submitted February 24, 1997, and was accepted April 13, 1998.

vided to raters. In study B, we provided raters with a standard stiffness target for each point on the rating scale and a more rigorously controlled protocol. In both studies, raters' estimates of stiffness were compared with their peers' estimates of stiffness (reliability analysis) and with instrumented measurements of PA spinal stiffness (validity analysis).

Instrumentation

The stiffness values of both the human spines and the mechanical reference stimuli were determined using the "Stiffness Assessment Machine" (SAM), an instrument designed to mimic the performance of the PA pressure test by the physical therapist.[13] This instrument consists of a rigid test bed, a small metal indenter that applies force to the subject's spine, and a mechanical head that both controls the movement of the indenter and collects data on the applied force and resultant displacement. The SAM has been described in detail elsewhere[14] and has been shown to provide reliable measurements of lumbar PA spinal stiffness in humans and accurate measurements of the stiffness of aluminum beams. From the force-displacement curve generated using the instrument, 2 measurements of PA spinal stiffness are obtained. The first measurement (K) is obtained by calculating the slope of the linear portion of the force-displacement curve above 30 N. The second measure (D30) characterizes the toe region of the force-displacement curve and is obtained by measuring the displacement to 30 N.

The reference stiffness stimuli were generated by a mechanical device[13] that has been used to provide precisely controlled stiffness stimuli in previous perceptual studies.[11,12] The resistance to movement is provided by metal compression springs, and the stiffness of the movement can be changed by adjusting the position of the existing spring (which affects the length of the lever arm the subject uses) or by fitting a different spring. Although this mechanical device is not ideally suited for clinical practice, in that it is somewhat cumbersome to move and expensive to make, we envisage that it would be relatively easy to develop a smaller, cheaper version of the device for use in the clinic. Accordingly, it could then be feasible to use the new stiffness testing protocols that are evaluated in this study in clinical contexts, but the current study did not use a clinical context.

Raters

The 3 raters who took part in study A were graduate physical therapist students who were completing a master's degree in manipulative physical therapy. The raters were selected by random ballot from the group of 16 students enrolled in a course devoted to examination and treatment of the lumbar spine. At the time the study was undertaken, the raters had completed both the academic and clinical education components of the course. The first 2 authors (CGM and JL) were the 2 raters who took part in study B. Both authors are university academicians and are involved in academic teaching, clinical education, and research in the area of manipulative treatment of persons with LBP. A description of the raters is given in Table 1.

Table 1.
Description of the Raters Who Took Part in the Studies

	Study A			Study B	
Variable	**Rater 1**	**Rater 2**	**Rater 3**	**Rater 1 (CGM)**	**Rater 2 (JL)**
Years of clinical experience using PA[a] pressure	5	5	7	13	13
Years as MPAA[b] member	0	0	0	6	6
Frequency of using PA pressure	Daily	Daily	Daily	Weekly	Weekly

[a] PA=posteroanterior.

[b] MPAA=Manipulative Physiotherapists Association of Australia.

Stiffness Stimuli

The volunteers (N=40) who provided the stiffness stimuli in these studies were currently asymptomatic for LBP. The K values of the subjects ranged from 6.83 to 22.99 N/mm, with a mean value of 13.77. The D30 values ranged from 3.1 to 8.46 mm, with a mean value of 5.44 mm. These ranges would be likely to incorporate the PA spinal stiffness of people with nonspecific LBP.[14] A full description of the subjects is given in Table 2. All subjects gave written informed consent to participate.

Rating Protocol

The points on the 11-point stiffness rating scale were anchored to discrete stiffness stimuli. This practice differed from typical clinical practice and from the earlier stiffness rating protocol we have used[7] in which stiffness magnitude was judged relative to each rater's memory value for "normal" or "average" stiffness. In study A, only one anchor stimulus was used (ie, 13.19 N/mm), and this stimulus was designated as the zero, or normal, stiffness point on the rating scale (Tab. 3). The −5 point on the scale was described as corresponding to "markedly reduced stiffness," and the +5 point on the scale was described as "markedly increased stiffness." The reference anchor stimulus was generated by a mechanical device[13] that was positioned on a 73-cm-high trolley at the end of the plinth on which the subject was lying. Before rating each subject, the rater would press on the mechanical reference stimulus, located at the foot of the plinth, and then walk to the middle of the plinth to press the lumbar spine. Thirteen subjects, whose spines were to be rated, were each positioned prone on a height-

Table 2.
Characteristics of the Subjects Who Were Rated (N=40)

	$\bar{X}$		SD		Range	
Variable	**Study A**	**Study B**	**Study A**	**Study B**	**Study A**	**Study B**
Age (y)	30	22.55	4.18	5.86	26–41	18–43
Height (cm)	169.15	171.52	10.29	7.97	155–182	157–192
Weight (kg)	63.23	66.02	9.19	11.27	52–84	47–88
K value (N/mm)	14.43	13.46	4.13	4.37	8.82–22.7	6.83–22.99
D30 value (mm)	5.58	5.37	1.29	1.54	3.94–8.22	3.1–8.46
Gender						
Male	8	10				
Female	5	17				

Table 3.
Reference Stimuli Used in Studies A and B[a]

Point on Rating Scale	Reference Stimulus (N/mm)	
	Study A	**Study B**
−5		4.52
−4		6.11
−3		7.76
−2		8.83
−1		10.40
0	13.19	12.13
1		13.84
2		15.74
3		17.43
4		19.45
5		23.25

[a] In study A, point −5 was accompanied by the descriptor "markedly reduced stiffness" and point 5 was accompanied by the descriptor "markedly increased stiffness."

adjustable plinth. The subjects' lumbar spines were exposed by lowering their pants to the natal cleft and raising their shirt to the midthoracic spine, and the L3 spinous process was palpated and marked. The raters were asked to rate the PA stiffness of the L3 level using their preferred method (all raters chose the pisiform-grip method) on 2 occasions, a week apart, with the order of testing the same on both occasions.

To ensure consistent location of the L3 level within a session, the spinous process was marked with an indelible pen after all raters had agreed on its location, and all raters subsequently rated PA spinal stiffness with respect to this marked level. This procedure may artificially increase reliability estimates, because in clinical practice each examiner needs to find this location independently. To ensure consistent location from week to week, the distance between this mark and the top of the natal cleft was recorded and this information was used when relocating L3 for the second rating occasion. On the second occasion, the subjects also were measured with the SAM after all manual testing had been completed.

The rating protocol in study B was further controlled, in that mechanical reference stimuli were provided for each of the 11 points on the scale (Tab. 3), and factors that subsequently had been found to affect the magnitude of perceived stiffness (ie, vision, grip type) or PA spinal stiffness as measured with the SAM (ie, subject's posture and breathing, plinth surface on which the test was performed) were all controlled. The anchor stimuli were provided by the same mechanical device used in study A. In study B, however, the mechanical device was clamped to the frame of a height-adjustable plinth placed immediately adjacent to the plinth on which the subjects lay. This method allowed the raters to position the device at the same height as the subjects and then to swivel between the device and the subject. Prior to the study, the raters practiced generating the 11 anchor stimuli using the mechanical device, although both raters were already quite familiar with the device, having used it in data collection in earlier studies.

In study B, a thin layer of foam was attached to the contact pad of the spring device because raters had commented that this modification made the contact feel more like a human spine. During testing, the subjects undressed to their underwear and wore a hospital gown. The subjects were positioned in the standard SAM testing position,[14] and all testing was done at functional residual capacity. The raters were allowed to press on the spine as often as they wanted using the pisiform grip, but they were required to look at their hands as they performed the test. To ensure that the raters and the SAM operator were assessing the same level, the first measurer (either a rater or the SAM operator) would identify and mark the L3 level with indelible ink, and all subsequent measurements were made with respect to this mark.

Table 4.
Reliability Results

	Study A			
Reliability Index[a]	**Occasion 1**	**Occasion 2**	**Occasions 1 and 2 Combined**	**Study B**
ICC (2,1)	.62	.50	.55	.77
ICC (2,1) 95% CI	.27–.85	.18–.78	.32–.79	.57–.89
SEM (average)	1.35	1.58	1.49	0.72

[a] ICC=intraclass correlation coefficient, CI=confidence interval, SEM=standard error of the measurement.

In study B, the raters matched the stiffness of the human spine to that of the spring device by adjusting the spring stiffness value and recording the number on the scale that corresponded to the matching stiffness stimulus. In this study, the SAM testing and human rater testing were done in the same session, with the order of testing counterbalanced. All SAM measurements were made by a research assistant who had been trained in its operation by the second author (the developer of the SAM).

In both studies, the raters were blinded to each other's ratings until data collection was completed. In study B, where 2 of the authors participated as raters, blinding was achieved by hiring a second research assistant who coordinated data collection and kept all rating sheets until the study was completed.

Data Analysis

These 2 studies were different from previous research that has evaluated judgments of PA spinal stiffness because we were able to examine criterion-related validity as well as the reliability of raters' judgments. To provide a comprehensive analysis of reliability, intraclass correlation coefficients (ICC[2,1]) with 95% confidence intervals (95% CI)[15] and standard errors of the measurement (SEMs) were calculated. The 95% CIs for the ICCs provide an estimate of the interval within which the population reliability would fall and thus can be used for hypothesis testing.

In study A, where there were 2 rating sessions, indexes were calculated separately for each session and then for the pooled data from both sessions. The point estimates and 95% CIs for the ICCs were calculated using software developed at The University of Sydney. Standard errors of the measurement were estimated using the formula provided by Ottenbacher and colleagues.[16]

To examine validity, we correlated each rater's manual stiffness judgments with the K measurements obtained with the SAM. Pearson product-moment correlation coefficients were calculated using QuattroPro for Windows 5.0.* All raters were familiar with the K measure and were aware that the validity of their judgments would be evaluated by comparison with it. In study A, this evaluation entailed pooling the manual ratings for both rating occasions. Pearson correlation coefficients were corrected for an attenuation due to low reliability, using the formula provided by Fleiss.[17] Estimates of the reliability of SAM measurements were taken from an earlier study,[14] whereas this study provided the reliability coefficients for the manual ratings. This form of correction is common in research where the real interest is in the relationship between the true scores rather than the observed scores, which contain error. The effect of this error is to reduce the magnitude of a correlation coefficient.[18] Ninety-five percent CIs were calculated for both the uncorrected and Fleiss-corrected Pearson correlation coefficients to evaluate whether the obtained Pearson correlation coefficients were different from zero.

Results

Reliability

The reliability results are shown in Table 4. In study A, the reliability of ratings was fair to good, with ICC (2,1) values ranging from .50 to .62 and SEMs ranging from 1.35 to 1.58 scale points. In study B, where the protocol was more standardized, the reliability was higher, with an ICC of .77 and a SEM of .72 scale points.

Validity

The validity results from both studies are shown in Table 5. The correlations of the manual ratings with the SAM K values ranged from .26 to .56. When corrected for an attenuation due to low reliability, the values ranged from .35 to .65. Because the correlations with the SAM K values were not high, the manual stiffness judgments were correlated with the SAM D30 values to determine whether there was evidence that raters' stiffness judgments were being influenced by the initial feel. Correlations with the SAM D30 values were lower, rang-

*Borland International Inc, 1800 Green Hills Rd, PO Box 660001, Scotts Valley, CA 95067-0001.

Table 5.
Validity Results[a]

Validity Index Point Estimate and 95% CI	Study A			Study B	
	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2
Correlation K	.26 (−.14 to .59)	.42 (.04 to .69)	.37 (−.02 to .66)	.41[b] (.04 to .68)	.56[b] (.23 to .78)
Fleiss-corrected correlation K	.35 (−.04 to .65)	.57[b] (.16 to .78)	.50[b] (.14 to .74)	.48[b] (.12 to .73)	.65[b] (.36 to .83)
Correlation D30	.00 (−.39 to .39)	−.19 (−.21 to .54)	−.20 (−.20 to .55)	−.13 (−.26 to .49)	−.19 (−.20 to .53)
Fleiss-corrected correlation D30	.00 (−.39 to .39)	−.27 (−.13 to .60)	−.28 (−.12 to .60)	−.15 (−.24 to .50)	−.22 (−.17 to .55)

[a] Results of validity analysis where manual posteroanterior (PA) stiffness ratings were correlated with 2 instrumented measures of PA spinal stiffness: K and D30. K is the slope of the linear portion of the force-displacement curve above 30 N, whereas D30 is the displacement to 30 N. Point estimates and 95% confidence intervals (95% CI) are presented for the uncorrected Pearson product-moment correlations and for the Fleiss-corrected Pearson product-moment correlations.
[b] 95% CI does not include zero.

ing from −.20 to .00. With the Fleiss correction, the correlations ranged from −.28 to .00.

An inspection of the results from study B revealed that there was a large disagreement between the therapists' ratings and the SAM values for 2 subjects, and this disagreement would have substantially reduced the correlations between manual ratings and SAM measurements. These 2 subjects initially had SAM K values of 19.81 and 20.18 N/mm, respectively, whereas they were rated by both therapists at points on the scale that would equate to stiffness values of the order of 10 N/mm lower. To check for SAM error, both subjects' spinal stiffness was remeasured 1 day later, and the SAM values were found to be essentially the same as the initial values (ie, 20.49 and 20.54 N/mm, respectively). Thus, instrument error does not seem to be a likely explanation for the lack of agreement between the SAM ratings indicating high stiffness and the manual judgments of low stiffness.

Discussion

Both revised stiffness rating protocols using mechanical reference stimuli achieved much higher reliability values than have been reported with previous PA pressure protocols that have rated stiffness or mobility. Matyas and Bach[6] reported poor reliability for stiffness assessments, with Pearson correlation coefficients ranging from .09 to .38 and kappa values ranging from .08 to .34. Binkley and colleagues[8] similarly noted low reliability, with an ICC (1,1) value of .25, a generalized kappa value of .09, and a SEM of 1.2 points on a 9-point scale. Finally, in our previous study,[7] we found ICC (1,1) values ranging from .03 to .37 when the ratings of 3 pairs of therapists were pooled. In this earlier study, we also calculated the reliability of ratings for individual rating pairs at each of the 5 lumbar levels (total of 15 ICC[2,1] values), and these calculations produced only one situation with better reliability than that observed in the current study. The reliability of the ratings, however, typically was much lower, with one rating pair not achieving any ICC values above zero.

It is reasonable to consider whether other factors, apart from the revised rating protocol, could have been responsible for the higher reliability observed in this study. Although the raters in study B had extensive clinical experience and postgraduate training in manipulative therapy, we do not believe that this background could satisfactorily explain the difference in reliability between the current study and our earlier study[7] because the raters in the earlier study had the same training and either equal or greater clinical experience. With regard to the issue of clinical experience, it is interesting to note that no effect of experience was noted in the series of studies, reported in the article by Matyas and Bach,[6] that examined the reliability of stiffness judgments made with the PA pressure test.

An alternative explanation for the higher reliability is that the subjects in this study were asymptomatic for LBP and thus easier to rate than the subjects in our earlier study who were symptomatic for LBP. Again, we do not believe that the difference in subjects can satisfactorily explain our results. First, instruments can measure the PA stiffness of both subjects who are asymptomatic for LBP[19] and subjects who are symptomatic for LBP[14] with high reliability. Additionally, previous studies that have evaluated PA pressure have demonstrated low reliability, regardless of whether the subjects were symptomatic or asymptomatic for LBP.[7]

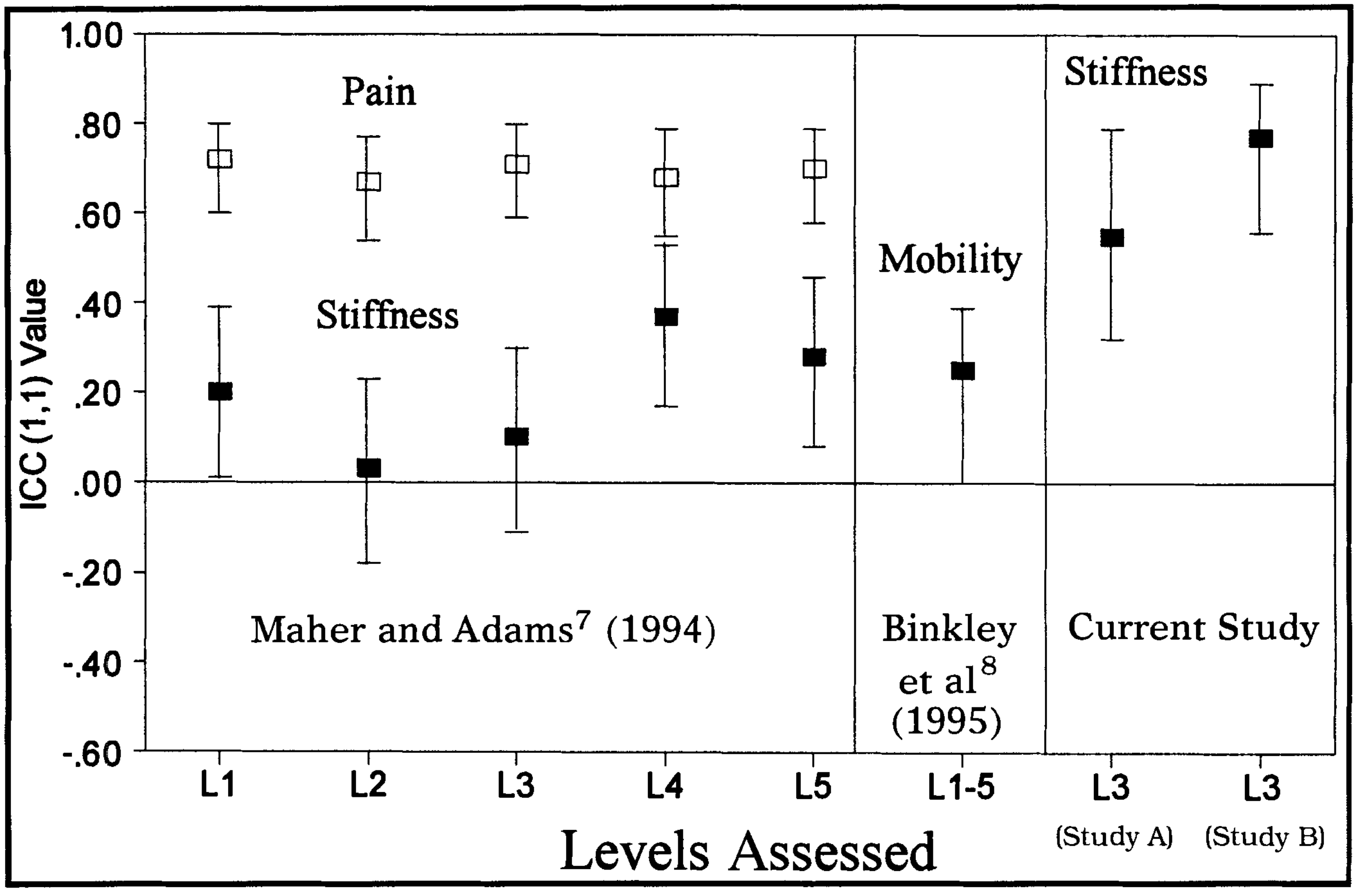

Figure.
Comparison of reliability achieved in current study with that achieved in 2 previous studies. Reliability is expressed using intraclass correlation coefficients (ICC[1,1]), with 95% confidence intervals. The open squares represent pain judgments obtained with the posteroanterior pressure, and the closed squares represent stiffness judgments or mobility judgments.

The results of the current study, with those of the study by Binkley and colleagues[8] and our original study,[7] are shown in the Figure. To allow for comparisons among the 3 studies, the ICC (2,1) reliability coefficients for this study were converted to ICC (1,1) values. The Figure shows that the lower limit of the confidence interval for the ICC (1,1) for study B does not overlap with the upper limit of the confidence intervals for our previous study or for the study by Binkley and colleagues. The lower limits of the confidence intervals for studies A and B, in which PA spinal stiffness at L3 was assessed, also do not overlap with the upper limit of the confidence intervals for L3 in our previous study. These results support the argument that there are real differences in reliability between the previous protocol and the revised versions we evaluated here. The Figure also shows that the 2 new reference-based protocols used here to rate PA spinal stiffness have reliability values that are of the same magnitude as the reliability of pain judgments. This finding is in contrast to our earlier study where there was no overlap of the ICC confidence intervals for the pain and stiffness judgments.

Although there are no agreed-on benchmarks for acceptable reliability, the ICC (2,1) value of .77 observed with the final revised rating protocol would be considered by Fleiss[17] to represent "excellent" reliability. The reliability values obtained in the current study are also of the same magnitude as those obtained with several well-accepted lumbar assessment procedures, such as the modified-modified Schöber test[20] and the Short-Form McGill Pain Questionnaire.[21] We must be cautious, however, because many aspects of this study do not relate to what may be seen in clinical settings. For example, the subjects were asymptomatic for LBP and the raters took advanced courses in manual therapy.

If the quality of the assessments obtained with the 2 new protocols had been evaluated solely on the basis of reliability, it would have been possible to conclude that the protocols produce somewhat reliable measurements of spinal stiffness in persons without LBP under ideal conditions and to have progressed to testing the new protocols on a patient sample. The results of the validity testing, however, caution against taking such a step.

The correlations between the manual stiffness judgments and the values obtained with the 2 SAM spinal stiffness measures (K and D30) were not high enough, even with the Fleiss correction, to constitute convincing evidence of the validity of the raters' spinal stiffness judgments. Interpretation of the magnitude of the Pearson correlation coefficients, however, is difficult, because, in the social and educational psychology field, a validity coefficient of the order of .50 would be considered acceptably large, but in psychophysics much higher Pearson correlation coefficients are typically expected.[22] Because the raters agreed with each other, it can be argued that they were tracking something with their ratings (ie, they did not simply represent noise or error), but the low validity coefficient suggests that the raters were measuring something different from what is measured by the SAM. The values for the 2 SAM measures were related to the manual ratings, but, at best, manual ratings predict only about 40% of the variance in SAM K measurements and less than 10% of SAM D30 measurements.

A consideration of the SAM measures may explain the disassociation between the manual stiffness ratings and the 2 SAM stiffness measures. Our current protocol for obtaining the SAM measurements is to precondition the spine with at least 5 testing cycles, and then collect data for 5 cycles at 0.5 Hz with a maximum force of 105 N and with the indenter angled 5.5 degrees cephalad when testing L3. The 2 SAM measurements are then obtained from the average of the loading curves for cycles 2 through 5. Although this measurement protocol produces reliable measurements of PA spinal stiffness, research that has been completed subsequent to the development of the SAM has shown that the protocol is different from the manual strategies adopted by therapists when pushing on the spine, and these differences may explain the disassociation between SAM stiffness measurements and judgments of PA spinal stiffness assessed manually.

One of the differences evident from a biomechanical perspective is that, with the SAM protocol, the angle of force application is fixed for all subjects, whereas therapists may vary the direction of applied manual force depending on the lordosis of the patient.[23] Additionally, therapists may apply larger forces[24] and load the spine at a higher frequency[25] than is the case with the current SAM testing protocol. Lastly, the SAM measurements were obtained on a rigid bed surface, whereas the manual ratings were obtained on a padded treatment couch. Each of these 4 factors—the angle of force application, the force magnitude, the frequency of loading, and the testing surface—has been shown to affect PA spinal stiffness.[10] Thus, it could be hypothesized that both the SAM stiffness measurements and the therapists' ratings represent the same stiffness measurements, although, in this study, they did not agree because they were obtained under different testing conditions. This hypothesis could be evaluated by repeating the study B but adjusting the SAM testing protocol so that it mimics the conditions of testing used by the therapists.

A psychophysical perspective also provides potential explanations for the disassociation in stiffness measures. In contrast to the earlier hypothesis, these explanations support the hypothesis that the SAM stiffness measurements and the manual stiffness ratings may be different. For example, it is not clear whether raters only consider the loading curve (as the SAM does) or are also considering the unloading curve when they make a stiffness judgment. It is also possible that therapists incorporate other mechanical variables that provide resistance to movement into their stiffness judgments (eg, viscosity, friction). Finally, it is possible that totally unrelated factors such as body type and sex may influence therapists' judgments of spinal stiffness. This last hypothesis presumes that stiffness judgments may be affected in a similar way to heaviness judgments, which have been shown to be influenced by non-weight cues such as the surface texture,[26] color, and volume of the lifted object.[27]

The hypothesis that judgments of PA spinal stiffness assessed manually are influenced by factors other than K and D30 may be difficult to evaluate because it would first be necessary to identify the mechanical variables (eg, viscosity, friction) and the nonmechanical variables (eg, age, sex, height, weight) that influence manual stiffness judgments. Given that heaviness judgments are influenced by such unlikely factors as the color of the lifted object, the identification of the critical features physical therapists attend to when they make a judgment of PA spinal stiffness may not be straightforward.

The identification of the factors that influence physical therapists' judgments of PA spinal stiffness may have implications for the management of LBP. If it is presumed that the SAM measures are the relevant markers of pathology and dysfunction in persons with LBP, it would make sense to attempt to train raters to attend to K and D30 and disregard the irrelevant cues they currently also attend to. Alternatively, it may be that the current SAM measures are not the best markers of pathology or dysfunction in persons with LBP. For example, there is a range of variables that could be extracted from the complex force-displacement data collected by the SAM, but, for simplicity, only 2 variables (K and D30) have been extracted. It may be that there are other mechanical factors that could be extracted from the SAM data that could prove to be more valid markers of the spinal level to direct treatment toward.

Identifying the mechanical cues that physical therapists attend to when making a stiffness judgment and studying their diagnostic value in epidemiological studies could provide better tests to direct manipulative care of persons with LBP.

Conclusion

The 2 PA spinal stiffness rating protocols evaluated in this study suggest that physical therapists may have a means of developing a clinically useful protocol for the assessment of PA spinal stiffness. Although the therapists in this study agreed among themselves, however, their stiffness judgments were somewhat disassociated from values obtained with 2 instrumented measures of spinal stiffness. Thus, the criterion-related validity of these measures is unclear, as is the clinical usefulness of judgments of PA spinal stiffness.

References

1 Maitland G. *Vertebral Manipulation.* 2nd ed. London, England: Butterworth & Co (Publishers) Ltd; 1968.

2 Jull G. Examination of the articular system. In: Boyling J, Palastanga N, eds. *Grieve's Modern Manual Therapy: The Vertebral Column.* Edinburgh, Scotland: Churchill Livingstone; 1995:520, 522.

3 Latimer J, Lee M, Adams R, Moran CC. An investigation of the relationship between low back pain and lumbar posteroanterior stiffness. *J Manipulative Physiol Ther.* 1996;19:587-591.

4 Mennell JM. *Back Pain: Diagnosis and Treatment Using Manipulative Techniques.* Boston, Mass: Little, Brown and Co; 1960.

5 Grieve G. *Mobilisation of the Spine: Notes on Examination, Assessment, and Clinical Method.* 4th ed. Edinburgh, Scotland: Churchill Livingstone Ltd; 1984.

6 Matyas T, Bach T. The reliability of selected techniques in clinical arthrometrics. *Australian Journal of Physiotherapy.* 1985;31:175–199.

7 Maher CG, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Phys Ther.* 1994;74: 801–809.

8 Binkley JM, Stratford PW, Gill C. Interrater reliability of lumbar accessory motion mobility testing. *Phys Ther.* 1995;75:786–792.

9 Phillips D, Twomey L. A comparison of manual diagnosis with a diagnosis established by a uni-level lumbar spinal block procedure. *Manual Therapy.* 1996;2:82–87.

10 Lee M, Steven G, Crosbie J, Higgs J. Towards a theory of lumbar mobilisation: the relationship between applied force and movements of the spine. *Manual Therapy.* 1996;2:67–75.

11 Maher CG, Adams R. Stiffness judgments are affected by visual occlusion. *J Manipulative Physiol Ther.* 1996;19:250–256.

12 Maher CG, Adams R. A comparison of pisiform and thumb grips in stiffness assessment. *Phys Ther.* 1996;76:41–48.

13 Maher CG, Adams R. Is the clinical concept of spinal stiffness multidimensional? *Phys Ther.* 1995;75:854–860.

14 Latimer J, Goodsell MM, Lee M, et al. Evaluation of a new device for measuring responses to posteroanterior forces in a patient population, part 1: reliability testing. *Phys Ther.* 1996;76:158–165.

15 Shrout P, Fleiss J. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86:420–428.

16 Ottenbacher K, Johnson M, Hojem M. The significance of clinical change and clinical change of significance: issues and methods. *Am J Occup Ther.* 1988;42:156–163.

17 Fleiss J. *The Design and Analysis of Clinical Experiments.* New York, NY: John Wiley & Sons Inc; 1986:4.

18 Schmidt F, Hunter J. Measurement error in psychological research: lessons from 26 research scenarios. *Psychological Methods.* 1996;1:199–223.

19 Lee M, Svensson N. Measurement of stiffness during simulated spinal physiotherapy. *Clin Phys Physiol Meas.* 1990;11:201–207.

20 Williams R, Binkley JM, Bloch R, et al. Reliability of the modified-modified Schöber and double inclinometer methods for measuring lumbar flexion and extension. *Phys Ther.* 1993;73:33–44.

21 Burckhardt C, Bjelle A. A Swedish version of the Short-Form McGill Pain Questionnaire. *Scand J Rheumatol.* 1994;23:77–81.

22 Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates Publishing Co; 1988.

23 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.

24 Simmonds MJ, Kumar S, Lechelt E. Use of a spinal model to quantify the forces and motion that occur during therapists' tests of spinal motion. *Phys Ther.* 1995;75:212–222.

25 Petty N, Messenger N. Can the force platform be used to measure the forces applied during a PA mobilisation of the lumbar spine? *Journal of Manual and Manipulative Therapy.* 1996;4:70–76.

26 Flanagan J, Wing A, Allison S, Spenceley A. Effects of surface texture on weight perception when lifting objects with a precision grip. *Perception and Psychophysics.* 1995;57:282–290.

27 Jones LA. Perception of force and weight: theory and research. *Psychol Bull.* 1986;100:29–42.