Research Report

Reliability of Physical Examination Items Used for Classification of Patients With Low Back Pain

Background and Purpose. The purpose of this study was to examine the interrater reliability of measurements obtained by examiners administering tests proposed to be important for classifying low back pain (LBP) problems. **Subjects.** Ninety-five subjects with LBP (41 men, 54 women) and 43 subjects without LBP (17 men, 26 women) were examined by 5 therapists trained in the techniques used. **Methods.** A manual was developed by the first author that described the clinical examination procedures. The therapists were trained by the first author in the test procedures and definitions. The training included instruction through videotapes, practice, and a written examination. Each examination was conducted by a pair of therapists. Within a pair, a therapist was the primary examiner for half of the subjects and an observer was the primary examiner for half of the subjects. Examination findings were recorded independently, without discussion. **Results.** Percentage of agreement and generalized kappa coefficients were used to analyze the data. Kappa values were ≥.75 for all 28 items related to the symptoms elicited and ≥.40 for 72% of the 25 items related to alignment and movement. **Conclusion and Discussion.** The results suggest that experienced therapists who had trained together were able to agree on the results of examinations and obtain an acceptable level of reliability. Future work should focus on testing of reliability when more than one therapist performs the examination and when therapists not trained by the test developer to administer the examination perform the tests. [Van Dillen LR, Sahrmann SA, Norton BJ, et al. Reliability of physical examination items used for classification of patients with low back pain. *Phys Ther.* 1998;78:979–988.]

Key Words: *Classification, Low back pain, Reliability, Spinal disorders.*

Linda R Van Dillen
Shirley A Sahrmann
Barbara J Norton
Cheryl A Caldwell
Deborah A Fleming
Mary Kate McDonnell
Nancy B Woolsey

Data from recent studies of practice indicate that mechanical low back pain (LBP) is the most common diagnosis for which patients are treated in outpatient physical therapy settings.[1] Due to the paucity of research, however, there is little consensus regarding what constitutes appropriate care.[2] In the absence of evidence for the effectiveness of treatment, clinicians opt for other rational approaches to treatment. One approach has been to use a pathoanatomical diagnosis to guide treatment. Unfortunately, a number of studies[3–9] have demonstrated that there are several problems associated with determining a specific pathoanatomical diagnosis in many cases of LBP. As a result, the formulation of a management strategy guided by a pathoanatomically based diagnosis is not possible for a large number of the patients with LBP.

As a second alternative for guiding treatment decisions, some authors[10–12] have proposed that the development of a classification system based on clusters of signs and symptoms relevant to physical therapy is needed. Several classification systems for LBP have been proposed,[13–24] but only 2 systems meet the criteria of having been designed specifically to direct physical therapy treatment and having been evaluated scientifically to some extent. The 2 systems that meet the criteria are the McKenzie system[20] and the Erhard and Bowling system.[23] The research examining the measurement properties of these systems, however, is limited[25–29]; there is minimal evidence in support of treatment effects for these systems[28,30–36]; and not all aspects of the existing systems are necessarily optimal. Thus, in our opinion, these classification systems do not eliminate the need for the development of alternative classification systems. In the remainder of this section, we will discuss 4 points of concern regarding the McKenzie system and the Erhard and Bowling system and provide a rationale for the alternatives we have incorporated into the system we are developing and testing.

The first issue is related to the names of the impairment-level categories that are used in the Erhard and Bowling classification system. The category names of the Erhard and Bowling system reflect the type of treatment to be administered (eg, "traction syndrome"). As noted by Rothstein,[12] however, the use of treatment descriptors for category names may be limiting. For example, patients who display signs and symptoms consistent with a particular Erhard and Bowling treatment-based category may be treated differently by 2 groups of clinicians. If the treatment strategy is used as the basis for naming categories, the same clinical entities (groups of patients with similar signs and symptoms) could be given different names. Additionally, different, more effective treatments may be developed for a particular category at a later point in time.

Because movement dysfunction is a unique focus of physical therapy, we believe category names that specify the primary movement problem may be more helpful in directing treatment by a physical therapist than treatment-based names. The classification system we are developing is impairment-based and defines 5 mutually exclusive categories of LBP problems. The LBP problems are named for the movements and postures that appear to be associated with the patient's LBP symptoms. For example, one of the category names is "lumbar rotation with extension." Although our system is focused on

L Van Dillen, PhD, PT, is Instructor, Program in Physical Therapy, Washington University School of Medicine, Campus Box 8502, St Louis, MO 63110 (USA) (vandille@medicine.wustl.edu). Address all correspondence to Dr Van Dillen.

SA Sahrmann, PhD, PT, FAPTA, is Professor and Associate Director for Doctoral Studies, Program in Physical Therapy, Washington University School of Medicine.

BJ Norton, PhD, PT, is Assistant Professor and Associate Director of Post-Professional Studies, Program in Physical Therapy, Washington University School of Medicine.

CA Caldwell, PT, CHT, is Instructor, Program in Physical Therapy, Washington University School of Medicine.

DA Fleming, PT, is Lead Physical Therapist, BJC Health System, St Louis, Mo.

MK McDonnell, PT, OCS, is Instructor, Program in Physical Therapy, Washington University School of Medicine.

NB Woolsey, OT, PT, is Instructor, Program in Physical Therapy, Washington University School of Medicine.

This study was approved by the Human Studies Committee of Washington University School of Medicine.

This work was funded, in part, by Grant 94R-03-N0R-02 from the Foundation for Physical Therapy Inc.

This article was submitted November 21, 1997, and was accepted March 5, 1998.

impairments, this does not mean that the use of other levels of classification to assist in the management of LBP problems is inappropriate (eg, the use of a functional status measure to classify at the level of functional limitation).[23,37,38]

A second issue relates to the emphasis on different lumbar spine movements within the McKenzie examination and the Erhard and Bowling examination. Although lumbar flexion and extension are examined in several different positions in both systems, lumbar rotation is assessed primarily through side gliding in a standing position in both systems, as well as through side bending in a standing position in the Erhard and Bowling system. Pearcy,[39] however, found that more lumbar rotation motion occurs in flexed positions, particularly sitting, than occurs in a standing position. Gordon et al[40] reported that small amounts of repeated rotation can produce microscopic injury to spinal structures. In addition, we believe that patients repeatedly use rotation to perform basic activities of daily living (eg, getting in and out of bed, getting in and out of a car). Considering that people appear to be performing small amounts of lumbar rotation repeatedly throughout their day in several different positions, we believe that a detailed assessment of lumbar spine rotation within the physical examination is warranted. Within our system, lumbar rotation, as well as flexion and extension, is examined in the following positions: (1) standing, (2) sitting, (3) supine, (4) prone, and (5) quadruped kneeling.

A third issue is the lack of attention given to assessment of the amount of lumbar spine movement that occurs with limb movements and the effect of limb movements on the patient's LBP symptoms in the McKenzie physical examination and the Erhard and Bowling physical examination. One of the assumptions underlying our classification system is that patients develop a predisposition to move the spine in a specific direction. For example, with upper-extremity flexion during a reaching movement, a person also might perform a lateral bending movement in the lumbar spine. We propose that if the lateral bending movement strategy is used repeatedly, such as in daily activities, the spine becomes more flexible in the direction of the repeated spine movement. In our physical examination, we examine the effect of limb movement on movement of the lumbar spine and on the production of LBP symptoms.

A final issue is the use of repeated spinal movements for assessment and treatment in the McKenzie system and the Erhard and Bowling system. McKenzie developed a method of eliciting symptoms in which patients perform repeated spinal movements in various positions. In general, repeated spinal movements that improve the patient's symptoms in the physical examination are prescribed as the exercise to manage the LBP problem.[20] The McKenzie method of eliciting symptoms has been adopted by the developers of the Erhard and Bowling system, and repeated spinal movements are used as treatment for 3 categories of the Erhard and Bowling system: extension, flexion, and lateral shift.[23]

Although repeated movement testing is part of both systems, we question the use of this approach for all patients with LBP. Neither the conditions under which nor the types of patients with which the McKenzie-based symptom tests can be administered reliably have been demonstrated clearly. The number of studies examining the reliability of data obtained by therapists performing the McKenzie-based symptom tests is small, and the findings from the studies are mixed.[25–27,29] Additionally, Delitto et al[23] have described 3 LBP categories in which the patient's symptoms are unaffected or are worsened with McKenzie-based repeated movement testing. They also noted that it is not uncommon to see patients whose symptoms initially may improve with a repeated spinal movement treatment regimen, then plateau, reverse, or worsen over time when the spinal exercise is continued.[23] In our view, the use of repeated movement testing during the examination may aggravate the patient's symptoms and confound the results of testing.

In our movement impairment system, we do not elicit symptoms with repeated spinal motions. Instead, we seek to confirm whether movements and positions may contribute to the patient's LBP through the use of several different, but related, tests. For example, the effect of rotation of the lumbar spine on LBP symptoms is assessed with side bending in a standing position, knee extension in a sitting position, hip rotation in a prone position, and rocking backward in the quadruped position rather than with repeated rotation movements in the same position. In our system, tests that produce symptoms may be repeated, but only for the purpose of determining how the movement can be modified to alleviate the symptoms.

Having provided a rationale for proposing an alternative classification system, the purpose of the remainder of this report is to describe the results of interrater reliability testing we have conducted using the physical examination items that we presume are necessary to categorize a patient's condition.

Method

Development of the Physical Examination

The first author (LVD) assumed primary responsibility for developing the physical examination. The examination was developed in collaboration with another author

Table 1.
Symptom Behavior Items: Measurement Scale, Kappa (κ) Values, and Percentages of Agreement

Physical Examination Item	Unit of Measurement	κ	Percentage of Agreement (%)
Standing			
Symptoms in standing	Yes, no	1.00	100
Forward bending	Same, decreased, increased	.97	99
Return from forward bending	Same, decreased, increased	.99	99
Corrected forward bending	Same, decreased, increased	.98	99
Corrected return from forward bending	Same, decreased, increased	.98	99
Side bending	Same, decreased, increased	.98	99
Posterior pelvic tilt against wall	Same, decreased, increased	.98	99
Sitting			
Sitting with lumbar spine flat	Same, decreased, increased	.99	99
Sitting with lumbar spine flexed	Same, decreased, increased	1.00	100
Sitting with lumbar spine extended	Same, decreased, increased	1.00	100
Active knee extension	Same, decreased, increased	1.00	100
Corrected active knee extension	Same, decreased, increased	.92	98
Supine			
Hips and knees flexed	Same, decreased, increased	.99	99
Hips and knees extended	Same, decreased, increased	.97	99
Passive straight leg raising	Positive, negative	.93	99
Hook lying			
Active hip abduction and lateral rotation (without pelvis stabilized)	Same, decreased, increased	.98	99
Active hip abduction and lateral rotation (with pelvis stabilized)	Same, decreased, increased	.97	99
Prone			
Prone (without abdominal support)	Same, decreased, increased	.97	99
Prone (with abdominal support)	Same, decreased, increased	.96	98
Active knee flexion	Same, decreased, increased	.87	98
Active hip rotation	Same, decreased, increased	.95	98
Active hip extension	Same, decreased, increased	.97	98
Quadruped kneeling			
Natural alignment	Same, decreased, increased	.97	99
Corrected alignment	Same, decreased, increased	.99	100
Arm lifting	Same, decreased, increased	.89	98
Rocking backward	Same, decreased, increased	.89	100
Corrected rocking backward	Same, decreased, increased	.93	98
Rocking forward	Same, decreased, increased	.97	99

(SAS), a group of 4 orthopedic physical therapists (CAC, DAF, MKM, and NBW) experienced in the treatment of persons with LBP problems, and a physical therapist (BJN) with expertise in measurement theory and research methods.[41]

The initial step in development of the physical examination entailed the identification of categories to be included in the classification scheme. Five mutually exclusive LBP categories were proposed by the second author (SAS) based on her clinical observations. She specified the physical examination findings characteristic of each LBP category. Finally, operational definitions and procedures for each test item were developed and summarized in a reference manual by the first author.

The physical examination included items related to (1) reports of symptoms associated with various positions and movements and (2) judgments of alignment and movement (signs) in different patient positions. For each of the items related to symptoms, the subjects either assumed a posture or performed a movement (spinal or lower extremity) and then reported the status of their LBP symptoms with the current test activity relative to a specified prior test position or movement. For example, subjects were asked to indicate the status of their LBP symptoms in their natural standing position. Subjects then moved into trunk flexion. On completion of the test movement, the subjects were asked to report the status of their symptoms during the trunk flexion movement and to compare it with the status of their symptoms in the natural standing position. Response options were (1) symptoms increased, (2) symptoms decreased, and (3) symptoms remained the same. The Appendix contains the operational definitions for the 3 response options. Table 1 lists the symptom behavior items included in the physical examination.

Table 2.
Alignment and Movement Items (Signs): Measurement Scale, Kappa (κ) Values, and Percentages of Agreement

Physical Examination Item	Unit of Measurement	κ	Percentage of Agreement (%)
Standing			
Shape of the lumbar curve (without flexible ruler)	Increased flexion, normal, increased extension	.49	80
Shape of the lumbar curve (with flexible ruler)	Increased flexion, normal, increased extension	.66	92
Asymmetry of the lumbar region	Yes, no	.27	84
Regularity of the lumbar curve (without flexible ruler)	Yes, no	.32	67
Swayback	Yes, no	.50	80
Regularity of the lumbar curve (with flexible ruler)	Yes, no	.52	87
Lumbar flexion with forward bending	Yes, no	.00	100
Lumbar extension with forward bending	Yes, no	.00	100
Relative flexibility with forward bending	Yes, no	.51	76
Hip extension with return from forward bending	Yes, no	.48	91
Lumbar extension with return from forward bending	Yes, no	.54	92
Pelvic and shoulder sway with return from forward bending	Yes, no	.39	74
Asymmetry with side bending	Yes, no	.26	65
Sitting			
Lumbar spine or pelvic rotation with knee extension	Yes, no	.58	86
Hook lying			
Relative flexibility with active hip abduction and lateral rotation	Yes, no	.60	88
Prone			
Relative flexibility with active knee flexion	Yes, no	.76	90
Asymmetrical pelvic rotation with active knee flexion	Yes, no	.43	90
Relative flexibility with active hip rotation	Yes, no	.56	83
Asymmetrical pelvic rotation with active hip rotation	Yes, no	.52	74
Quadruped kneeling			
Lumbar spine alignment	Flexed, flat, extended	.58	74
Asymmetry of the lumbar region	Yes, no	.42	83
Alignment of the hip joint	<90°, 90°, >90°	.61	88
Asymmetrical rotation with arm lifting	Yes, no	.21	55
Relative flexibility with rocking backward	Yes, no	.78	95
Pelvic rotation/tilt with rocking backward	Yes, no	.51	82

For alignment items, examiners generally had to judge whether the lumbar spine was primarily flexed, extended, or flat. Alignment was judged in various positions, such as standing, sitting, and quadruped kneeling. Items related to movement focused on the type of lumbar spine motion associated with either trunk movement or movement of the extremities. For example, as a subject performed a trunk flexion movement from a standing position, the therapist had to decide whether the early part of the movement was performed with flexion or with extension of the lumbar spine, and whether the lumbar spine flexed at a faster rate than the hips in the first 50% of the forward bending movement. Table 2 lists the alignment and movement items included in the physical examination.

Subjects

Ninety-five subjects with LBP problems were tested. Two different methods were used to recruit subjects with LBP. First, subjects were recruited from patients referred for treatment to 1 of 6 different outpatient physical therapy centers in the St Louis (Mo) metropolitan area. A process was implemented at each center for distributing a written description of the study to every new patient with the diagnosis of LBP. Patients were contacted by telephone, and the study was described in more detail. If an individual indicated that he or she was interested in participating in the study, that person was asked a series of questions based on the inclusion and exclusion criteria. Second, in addition to obtaining subjects through the 6 local clinical facilities, subjects also were recruited from families and friends of patients with LBP participating in the study and through advertisements and posters distributed throughout the medical center and campus of the university. Individuals recruited in this manner called a contact person in the department where the study was conducted.

Subjects between 18 and 75 years of age who had symptoms (pain or paresthesia) related to a low back problem in either the region of the lower back, proximal lower extremity, or distal lower extremity[17] were eligible

Table 3.
Characteristics of Study Sample

Variable	Subjects With LBP	Subjects Without LBP
Sex		
Male	41	17
Female	54	26
Age (y)		
$\bar{X}$	44.07	39.38
SD	13.29	13.05
Duration of current LBP symptoms[a]		
Acute: <7 days	6	
Subacute: 7 days–7 weeks	18	
Chronic: >7 weeks	71	
Location of current symptoms[b]		
Low back only	56	
Low back/proximal LE	12	
Low back/distal LE	5	
Low back/proximal LE/distal LE	22	
History of previous episode of LBP		
Yes	79	
No	16	
Oswestry Questionnaire[c] Disability Scores[d]		
$\bar{X}$	24%	
SD	15%	

[a] Classification defined by Quebec Task Force on Spinal Disorders based on duration of current low back pain (LBP) symptoms.[17]

[b] Definitions from Quebec Task Force on Spinal Disorders based on current LBP symptoms.[17] Low back: area extending from T12 to gluteal fold; proximal lower extremity (LE): area extending from gluteal fold to knee; distal LE: area extending from knee to foot.

[c] Fairbank JCT, Couper J, Davies JB, O'Brien JP. The Oswestry Low Back Pain Questionnaire. *Physiotherapy.* 1980;66:271–273.

[d] Disease-specific disability questionnaire that represents the degree of disability as a percentage score. 0%=no disability, 100%=maximal disability.

for the study. Subjects were excluded in the case of pregnancy, severe kyphosis or scoliosis, spinal stenosis, a history of spinal surgery in the last 3 months, more than one surgical procedure on the spine, pending spinal surgery, cancer, rheumatoid arthritis, ankylosing spondylitis, neurological disease, or the inability to stand and walk without an assistive device. The information related to the exclusion criteria was obtained from (1) the recruitment form each subject completed, which included the patient's diagnosis and any other notes thought to be relevant by the therapist treating the patient, and (2) the subject report during the telephone screening interview. All subjects read and signed an informed consent statement approved by the Washington University School of Medicine Human Studies Committee before participating in the study.

In the interest of testing the sensitivity and specificity of the clinical examination items, 43 age- and sex-matched subjects without LBP also were examined. There was no more than a 5-year difference in age between each subject with LBP and the matched subject without LBP. Subjects without LBP were recruited from the same sources as the subjects with LBP. Subjects without LBP were excluded if they had (1) an episode of LBP in the 12 months prior to testing that affected their ability to perform activities of daily living (ADL) for more than 3 days, (2) low back, hip, or leg pain in the 12 months prior to testing that required medical attention, caused missed work, or resulted in a noticeable adjustment or restriction of their ADL, or (3) any low back, hip, or leg pain in the week before testing.[24] Descriptive information about the subjects is presented in Table 3.

Examiners

Five orthopedic physical therapists (CAC, DAF, MKM, SAS, and NBW) participated as testers. All of the therapists practiced part-time in the same university-based outpatient orthopedic clinic. The 5 therapists varied with respect to their background and application of the many evaluation and treatment approaches proposed for use with patients with LBP. All of the therapists would include various items examined in this study as part of their evaluation of patients with LBP. The therapists ranged in age from 36 to 57 years ($\bar{X}$=43.2, SD=7.02), and the number of years of clinical experience ranged from 5 to 35 years ($\bar{X}$=16.8, SD=8.73). The clinical expert (SAS) developed the LBP classification scheme. Each of the other 4 therapists had participated in various aspects of the development of the operational definitions and procedures included in the clinical examination.

Training

The training of the therapists involved 3 steps. First, each therapist was given a reference manual that included (1) the findings characteristic of each LBP problem, (2) operational definitions and procedures for administering the clinical examination, and (3) a sample form for recording the clinical examination data. The description of the procedure for each item included (1) a description of the movement or position to be tested, (2) the position or movement the patient used as a comparison for the effect of a particular test on symptoms, (3) examiner and patient positioning, (4) instructions to be given to the patient, (5) the kinds of information used by the examiner to make a judgment about the patient's response to a particular test item (eg, vision, patient report, palpation), and (6) details regarding how to deal with common problems encountered when conducting a particular test item.

Therapists were required to study the manual and take a written examination covering the information. The examination was written and scored by the first author. A score of 90% was required before the therapist could begin formal testing. Second, a videotape was made to facilitate learning the material and to facilitate calibrating the therapists' responses based on vision. The videotape provided examples of tests using several subjects that demonstrated different values for their responses. Before viewing the videotape of each subject, the therapists were told the correct response for each test they were about to view to allow them to associate the performance with the correct response. Each therapist was required to view the videotape at least one time. Finally, the first author met individually with each of the therapists for one 45-minute session to explain the process for reliability testing, to review portions of the manual, and to answer any questions.

Design

One of the primary assumptions underlying reliability studies is that the phenomenon being measured is stable.[42] Because many of the variables included in the physical examination were likely to be affected by repeated testing, we chose to have 2 therapists examine subjects simultaneously. One therapist served as the primary examiner, and the other therapist served as the observer. Initially, each of 3 testers (CAC, DAF, MKM) was paired with each of the other 2 testers, resulting in a total of 3 pairs (ie, CAC and DAF, CAC and MKM, and DAF and MKM). After 1 month of data collection, each of the 2 other therapists was paired with 1 of the original 3 testers (ie, SAS was paired with MKM and NBW was paired with DAF). Therapists within each pair were assigned as either the primary examiner or the observer. A pseudorandom assignment process was used to ensure that each therapist performed an equal number of examinations in both roles. This process entailed random assignment of the 2 therapists to the examiner and observer roles in the majority of instances. If, however, a therapist of the pair was nearing their quota of total patient examinations and had not performed an equal number of examinations in both roles, the therapists were assigned to specific roles to eliminate the inequity. Blocks of times for testing were reserved with each pair of therapists on a weekly basis. Subjects were scheduled for testing at their convenience within the blocks of time available for the examiners.

Procedure

The primary examiner was responsible for conducting the clinical examination. The therapist assigned to the observer role listened and observed the examination. For instances in which the observer needed to perform palpation to make a judgment about the response on a test item, the observer would perform the test on the subject after the primary examiner had completed the test. The therapists were not allowed to discuss their interpretation of the subjects' responses with each other. All items of the clinical examination were administered in the same order to every subject. Each therapist of a pair recorded the examination findings on separate data forms at the time of testing. The data forms were collected at the completion of each examination.

Data Analysis

Generalized kappa and percentage of agreement were used to analyze the agreement between therapists for each of the dichotomous scale items of the examination.[43,44] The generalized kappa statistic provides an index of chance-corrected agreement.[44] Weighted generalized kappa and weighted percentage of agreement were used to analyze the agreement between therapists for each of the ordinal scale items of the examination.[43,44] Both weighted percentage of agreement and weighted kappa are indexes of agreement that take into account partial agreement. Each ordinal scale item included in this study had 3 response categories. The weights assigned to the 3 levels of agreement were as follows: (1) maximum agreement=1.0, (2) partial agreement=0.5, (3) maximum disagreement=0.0.

Results

The percentages of agreement for the symptom behavior items ranged from 98% to 100%. The kappa values ranged from .87 to 1.00 for the symptom behavior items. The percentages of agreement for the signs related to alignment and movement ranged from 65% to 100%. The kappa values for the alignment and movement items ranged from .00 to .78. Table 1 contains the kappa values and percentages of agreement for the symptom behavior items. Table 2 contains the kappa values and percentages of agreement for the alignment and movement items.

Discussion

Overall, the reliability of administering and interpreting the responses to the 28 symptom behavior items was very good.[45,46] The kappa values for all of the symptom behavior items were .75 or above.

The kappa coefficients for the alignment and movement items were not as large as those for the symptom behavior items. Of the 25 items in the alignment and movement category, 2 items had kappa values of .75 or above (what we would consider excellent reliability), 3 items had kappa values ranging between .60 and .74 (what we would consider good reliability), and 13 items had kappa ranging between .40 and .59 (what we would consider fair reliability).[45,46] Seven of the alignment and movement items had kappa values below .40, indicating what we would consider poor reliability.[45,46] Only 3 of these 7 items, however, had percentages of agreement below 70% (Tab. 2).

The large reliability coefficients for the items related to symptom behavior may be a result of a number of features of the study design. First, the questions to be asked by the therapists were designed to yield structured responses, the wording of the questions was established before testing began, and the response choices were defined clearly. Second, because we chose to focus on the variability associated with the clinician's administration of the test items and interpretation of a subject's responses, both therapists were present at the time of the testing. The variability due to repeated testing, therefore, was eliminated.

A number of factors could have contributed to the attenuation of the kappa values obtained for the alignment and movement items. First, many of these items required the therapist to make judgments of the alignment of anatomical regions (eg, the lumbar spine) or of aspects of the movement of the lumbar spine (eg, type, timing, and relative amounts of movement) with different trunk and extremity motions. The judgments were based on visual—and, in some cases, tactile—information. As evidenced by the findings from other studies of the reliability of examiners administering LBP impairment measures, judgments based on visual and tactile information are often difficult to make reliably.[47–50]

Many of the previous studies of reliability that obtained poor reliability of examiners, however, varied with regard to the amount of attention given to the detail of the definitions and procedures, as well as training of the examiners. The results of our study suggest that—with explicitly defined procedures, operational definitions that provide quantifiable threshold values, and standardized training—at least a fair level of agreement between therapists can be attained for many of these types of judgments. Our therapists, however, also were test developers and worked together extensively to increase reliability. The question of whether a different type or amount of training would yield a different level of reliability has not been tested.

Second, a review of anecdotal comments made by the therapists at completion of the study suggests that the responses from the alignment and movement tests were often very close to the threshold values provided for deciding on the presence or absence of various signs. As a result, small differences in a therapist's perception of a patient's behavior could have contributed to the disagreement between the therapists. Patients in the acute stage of an episode of LBP may have more pronounced alignment and movement impairments than patients assessed at a later point in their episode of LBP. The majority of our patients, however, were no longer in the acute phase of their episode of LBP (Tab. 3). Thus, the responses displayed by the patients in our sample may have been especially difficult to judge.

Although kappa values for the majority of the items related to alignment and movement were below the value of .75, a large proportion of these items had percentages of agreement or weighted percentages of agreement of 70% or larger. Discrepancies between kappa values and percentages of agreement of the type described can occur when the distribution of the item responses across options is skewed. Examination of items from our protocol with small kappa values but relatively large percentages of agreement revealed that the distribution of responses for a number of the items was skewed. The skewness in the response distributions may be related, in part, to the characteristics of the study sample. The majority of subjects in our study reported having an episode of LBP with a duration of greater than 7 weeks, and most subjects had a history of recurrent LBP problems (Tab. 3). Future studies will need to include patients with different characteristics than those of the patients in the current study (eg, more acute LBP injuries). Studies of patients with different characteristics may provide a better estimate of the reliability of individual items and may assist in determining whether the clinical findings identified and defined by our expert clinician actually exist in the population of persons with LBP.

The generalizability of our findings is limited due to certain aspects of the study's design and the nature of the examiners. First, consider the design of the study, specifically the schedule of testing. In our study, we chose to have 2 therapists examine each subject at the same time rather than at different times because (1) our clinical examination included items for which responses potentially could be affected by prior testing and (2) previous investigators[29] have suggested that poor reliability for items related to the symptoms elicited may have resulted from using a repeated testing (test-retest) design. Thus, we chose the simultaneous testing design so that we could focus on the variability associated with a clinician's administration of the test items and interpretation of a patient's responses, rather than on the variability due to changes in a patient's condition. The stability of a patient's status over time remains to be tested.

The method of simultaneous examination also may limit the generalizability of our findings to a clinical setting because patients seldom are seen by 2 therapists simultaneously. Nonetheless, we judged the simultaneous examination method to be acceptable for this stage of testing because (1) the majority of the clinical examination items were dependent on either visual judgments of alignment and movement or verbal reports of symptom behavior with movement testing and (2) given the nature of the judgments, there was no reason to suspect that they would be greatly affected by the therapists observing the patient together. Furthermore, items that required a combination of visual and tactile information to make a judgment about a patient's response were administered first by the primary examiner and then by the observing therapist. Administering the items sequentially could have influenced the judgments made by the second therapist, but (1) therapists were instructed to perform and record the results of their testing independently, (2) examinations were observed regularly by the first author (LVD) to monitor the therapist's behavior, and (3) examiners were admonished for any behaviors that potentially could introduce bias into the testing. Although the possibility existed that the therapists could influence each other's judgment, we judged the risk to be small relative to the risk of changes in a patient's status potentially associated with a test-retest design. The reliability for examinations conducted separately remains to be tested.

Because our movement impairment classification system contains elements we thought were different from those traditionally included in a clinical examination for LBP problems, we believed a development process was required to develop operational definitions for the examination items and responses and to define procedures for administration. The therapists who participated in discussions during the development process also participated as examiners in the reliability study. Although the nature of the examiners in our study limits the generalizability of our findings, we thought it was important first to determine whether clinicians with some background in the theoretical basis for the examination process were able to administer the examination reliably. Our examiners, therefore, were uniquely prepared, and the agreement they showed could have been a function of their working together. The reliability of data obtained by other examiners remains to be tested.

We are now attempting to determine the rules necessary to put patients into the movement impairment categories. In addition, we are currently implementing an examination in a clinical setting that includes the physical examination items tested in this study. The examination includes only those items for which our therapists attained at least a fair level of reliability. The purposes of this phase of our work are (1) to determine the feasibility of the use of the instrument in the clinic and (2) to expand the range of patients assessed with the instrument.

Conclusions

Overall, the examiners in this study demonstrated acceptable reliability in administering the majority of the clinical examination items. The reliability of the data obtained by the examiners was particularly good for the items related to the subjects' symptoms. Examiners were not as likely to agree on responses for items related to judgments of alignment and movement as they were for items related to the symptoms elicited. The examiners administering the majority of the alignment and movement items, however, had fair kappa values and observed agreement values above 70%.

References

1 Jette AM, Davis KD. A comparison of hospital-based and private outpatient physical therapy practices. *Phys Ther.* 1991;74:366–375.

2 Battié MC, Cherkin DC, Dunn R, et al. Managing low back pain: attitudes and treatment preferences of physical therapists. *Phys Ther.* 1994;74:219–226.

3 White AA 3rd, Gordon SL. Synopsis: workshop on idiopathic low-back pain. *Spine.* 1982;7:141–149.

4 Deyo RA. Early diagnostic evaluation of low back pain: clinical review. *J Gen Intern Med.* 1986;1:328–338.

5 Waddell G. A new clinical model for the treatment of low-back pain. *Spine.* 1987;12:632–644.

6 Waddell G, Main CJ, Morris EW, et al. Chronic low-back pain, psychologic distress, and illness behavior. *Spine.* 1984;9:209–213.

7 Boden SD, Davis DO, Dina TS, et al. Abnormal magnetic-resonance scans of the lumbar spine in asymptomatic subjects: a prospective investigation. *J Bone Joint Surg Am.* 1990;72:403–408.

8 Jensen MC, Brant-Zawadzki MN, Obuchowski N, et al. Magnetic resonance imaging of the lumbar spine in people without back pain. *N Engl J Med.* 1994;331:69–73.

9 Wiesel SW, Tsourmas N, Feffer HL, et al. A study of computer-assisted tomography I: the incidence of positive CAT scans in an asymptomatic group of patients. *Spine.* 1984;9:549–551.

10 Rose SJ. Editorial: Musing on diagnosis. *Phys Ther.* 1988;68:1665.

11 Sahrmann SA. Diagnosis by the physical therapist—a prerequisite for treatment: a special communication. *Phys Ther.* 1988;68:1703–1706.

12 Rothstein JM. Editor's note: Patient classification. *Phys Ther.* 1993; 73:214–215.

13 Saunders HD. Classification of musculoskeletal spinal conditions. *J Orthop Sports Phys Ther.* 1979;1:3–15.

14 Mooney V. The syndromes of low back disease. *Orthop Clin North Am.* 1983;14:505–515.

15 Sikorski JM. A rationalized approach to physiotherapy for low back pain. *Physiotherapy.* 1985;10:571–579.

16 Bernard TN Jr, Kirkaldy-Willis WH. Recognizing specific characteristics of nonspecific low back pain. *Clin Orthop.* 1987;217:266–280.

17 Spitzer WO, LeBlanc FE, Dupuis M. Scientific approach to the assessment and management of activity-related spinal disorders. In: Monograph for Clinicians: Report of the Quebec Task Force on Spinal Disorders. *Spine.* 1987;12:S16–S21.

18 Sypert GW. Low back pain disorders. *Trans Assoc Life Insur Med Dir Am.* 1988;71:174–197.

19 Porter RW. Mechanical disorders of the lumbar spine. *Ann Med.* 1989;21:361–366.

20 McKenzie RZ. *The Lumbar Spine: Mechanical Diagnosis and Therapy.* Waikanae, New Zealand: Spinal Publications Ltd; 1989.

21 DeRosa CP, Porterfield JA. A physical therapy model for the treatment of low back pain. *Phys Ther.* 1992;72:261–269.

22 Binkley JM, Finch E, Hall J, et al. Diagnostic classification of patients with low back pain: report on a survey of physical therapy experts. *Phys Ther.* 1993;73:138–150.

23 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;75:470–485.

24 Moffroid MT, Haugh LD, Henry SM, Short B. Distinguishable groups of musculoskeletal low back pain patients and asymptomatic control subjects based on physical measures of the NIOSH Low Back Atlas. *Spine.* 1994;19:1350–1358.

25 Kilby J, Stigant M, Roberts A. The reliability of back pain assessment by physiotherapists using a "McKenzie algorithm." *Physiotherapy.* 1990; 76:579–583.

26 Riddle DL, Rothstein JM. Intertester reliability of McKenzie's classifications of the syndrome types present in patients with low back pain. *Spine.* 1993;18:1333–1344.

27 Nelson RM, Nestor DE. Standardized assessment of industrial low-back injuries: development of the NIOSH Low Back Atlas. *Topics in Acute Care Trauma Rehabilitation.* 1988;2:16–30.

28 Cibulka MT, Delitto A, Koldehoff RM. Changes in innominate tilt after manipulation of the sacroiliac joint in patients with low back pain: an experimental study. *Phys Ther.* 1988;68:1359–1363.

29 Delitto A, Shulman AD, Rose SJ, et al. Reliability of a clinical examination to classify patients with low back syndrome. *Physical Therapy Practice.* 1992;1:1–9.

30 Ponte DJ, Jensen GJ, Kent BE. A preliminary report on the use of the McKenzie protocol versus Williams protocol in the treatment of low back pain. *J Orthop Sports Phys Ther.* 1984;29:130–139.

31 Nwuga G, Nwuga V. Relative therapeutic efficacy of the Williams and McKenzie protocols in back pain management. *Physiotherapy Practice.* 1985;6:130–139.

32 Stankovic R, Johnell O. Conservative treatment of acute low-back pain—a prospective randomized trial: McKenzie method of treatment versus patient education in "mini back school." *Spine.* 1990;15: 120–123.

33 Williams MM, Hawley JA, McKenzie RA, van Wijmen PM. A comparison of the effects of two sitting postures on back and referred pain. *Spine.* 1991;16:1185–1191.

34 Stankovic R, Johnell O. Conservative treatment of acute low back pain: a 5-year follow-up study of two methods of treatment. *Spine.* 1995;20:469–472.

35 Delitto A, Cibulka MT, Erhard RE, et al. Evidence for use of an extension-mobilization category in acute low back syndrome: a prescriptive validation pilot study. *Phys Ther.* 1993;73:216–222.

36 Erhard RE, Delitto A, Cibulka MT. Relative effectiveness of an extension program and a combined program of manipulation and flexion and extension exercises in patients with acute low back syndrome. *Phys Ther.* 1994;74:1093–1100.

37 Jette AM. Diagnosis and classification by physical therapists: a special communication. *Phys Ther.* 1989;69:967–969.

38 Guccione AA. Physical therapy diagnosis and the relationship between impairments and function. *Phys Ther.* 1991;71:499–503.

39 Pearcy MJ. Twisting mobility of the human back in flexed postures. *Spine.* 1993;18:114–119.

40 Gordon SJ, Yang KH, Mayer PJ, et al. Mechanism of disc rupture: a preliminary report. *Spine.* 1991;16:450–456.

41 Van Dillen LR, Sahrmann SA, Norton BJ, et al. Development and standardization of a clinical approach for classification of patients with low back pain [abstract]. *Phys Ther.* 1993;73:S16. Abstract PO-S042-M.

42 Cohen RJ, Montague P, Nathanson LS, Swerdlik ME. *Psychological Testing: An Introduction to Tests and Measurement.* Mountain View, Calif: Mayfield Publishing Co; 1988: chap 5.

43 Kramer MS, Feinstein AR. Clinical biostatistics, LIV: the biostatistics of concordance. *Clin Pharmacol Ther.* 1981;29:111–123.

44 Uebersax JS. A generalized kappa coefficient. *Educational and Psychological Measurement.* 1982;42:181–183.

45 Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Ment Defic.* 1981;86:127–137.

46 Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics.* 1977;33:363–374.

47 Gonnella C, Paris SV, Kutner M. Reliability in evaluating passive intervertebral motion. *Phys Ther.* 1982;62:436–444.

48 Potter NA, Rothstein JM. Intertester reliability for selected clinical tests of the sacroiliac joint. *Phys Ther.* 1985;65:1671–1675.

49 Maher CG, Adams R. Reliability of pain and stiffness assessments in clinical manual lumbar spine examination. *Phys Ther.* 1994;74: 801–808.

50 Strender LE, Sjöblom A, Sundell K, et al. Interexaminer reliability in physical examination of patients with low back pain. *Spine.* 1997;7:814–820.

Appendix.

Operational Definitions for Responses for Symptom Behavior Items of the Physical Examination

Symptoms increased:

The subject's symptoms (pain or paresthesia) are produced, the symptoms present at initiation of a particular test are increased in intensity, or the symptoms have moved distally from the lumbar spine, with assumption of a test position or performance of a test movement, as compared with the referent symptom status.

Symptoms decreased:

The subject's symptoms are diminished or absent, or the symptoms have moved more proximally toward the lumbar spine with assumption of a test position or performance of a test movement, as compared with the referent symptom status.

Symptoms remained the same:

Assumption of a test position or performance of a test movement has no effect on the intensity or location of the subject's symptoms, as compared with the referent symptom status.