Maintenance and Improvement of Interobserver Reliability of Osteopathic Palpatory Tests Over a 4-Month Period

Brian F. Degenhardt, DO; Jane C. Johnson, MA; Karen T. Snider, DO; and Eric J. Snider, DO

Context: Few studies have shown that diagnostic palpation is reliable. No studies have shown that the reliability of diagnostic palpatory skills can be maintained and improved over time.

Objective: To investigate whether the reliability of selected palpatory tests used to identify lumbar somatic dysfunction was maintained during a 4-month period as part of a clinical observational study.

Methods: Participants with low back pain and participants without low back pain, recruited from a rural Midwestern community, were examined during 6 separate sessions over a 4-month period. During each data collection session, two blinded examiners, who had previously completed comprehensive consensus training, evaluated the lumbar region with four tests: static segmental positional asymmetry of the transverse processes in the horizontal plane, tissue texture abnormalities, resistance to anterior springing on the spinous processes, and tenderness induced by pressure on the spinous processes. Detailed protocols for each test were defined during a previous comprehensive consensus training period and were not revised during the current study. To verify that established interobserver reliability was maintained throughout the clinical study, quality control sampling was performed on all data. When findings were inconsistent between the two examiners, focused consensus training was performed as a means of recalibration to understand why assessments were inconsistent. Interobserver reliability for determining the presence or absence of somatic dysfunction was assessed using kappa coefficients.

Results: The study enrolled 64 participants, and 14 to 33 participants were examined per session. All four tests had acceptable interobserver reliability by the final data collection session. The test for static segmental positional asymmetry of the transverse processes in the horizontal plane had moderate to substantial reliability in all 6 sessions. The test for tissue texture abnormalities had moderate reliability in 5 of the 6 sessions. The test for resistance to anterior springing on the spinous processes had moderate reliability for 3 of the 6 sessions. The test for tenderness had substantial to almost perfect reliability for all 6 sessions. In general, interobserver reliability improved over time.

Conclusions: Examiners were able to maintain and improve interobserver reliability of four lumbar diagnostic palpatory tests over a 4-month period.


Practitioners from multiple manual-therapy professions, such as chiropractic, massage therapy, osteopathic medicine, and physical therapy, use palpation to diagnose and treat patients with disorders of the spine. Despite the widespread use of palpation, scientific assessment of the value of diagnostic palpation and of the effectiveness of manual treatment has been limited, and there is a lack of fundamental evidence demonstrating the reliability and validity of the numerous tests and techniques used within each profession. Although variation exists in the techniques used by individual professions, tests evaluating the spine can be organized into categories that assess the following four characteristics: symmetry of spinal landmarks, soft tissue abnormalities, segmental and regional motion, and pain.

For more than 30 years, studies have investigated the reliability of spinal diagnostic palpatory tests, but systematic reviewers have criticized most of these studies for having poor methodologic designs. Despite variations in the inclusion criteria used in recent systematic reviews of research on diagnostic palpatory tests, the conclusions made by those reviewers have been similar: pain provocation tests are reliable both between and within observers; motion and landmark location tests are more reliable within the same observer (ie, intraobserver reliability) than between observers (ie, interobserver reliability); and, overall, tests other than pain provocation have poor interobserver reliability.
A previous study by the present authors9 showed that improvement in reliability of several commonly used palpatory tests for the lumbar spine could be achieved after a period of consensus training. These tests, performed at the L1 through L4 vertebral levels, evaluated static segmental positional asymmetry of the transverse processes in the horizontal plane, tissue texture abnormalities, resistance to induced motion on the transverse processes, resistance to anterior springing on the spinous processes, and tenderness induced by pressure on the spinous processes. In that study,9 three phases were part of the training process: pretraining interobserver reliability assessment, consensus training, and posttraining interobserver reliability assessment.

The acceptable reliability of tests in the previous study9 was limited to the examiners in that study at that time. For the current study, we hypothesized that the reliability of these tests could be maintained and improved upon over an extended period. A review of the literature identified only one report10 on the evaluation of interobserver agreement over time. In that study,10 thoracic vertebrae were evaluated by two examiners in two sessions within a 10-day period. Interobserver reliability for the presence vs absence of motion restriction was fair in both sessions (Cohen’s kappa coefficient $k = 0.22$ and 0.24).

In the current study, two blinded examiners evaluated spinal dysfunctions from L1 through L4 using four common palpatory tests over a 4-month period within a clinical observational study. Participants with low back pain and with no low back pain were included in the study. The purpose of the current study was to verify that established interobserver reliability was maintained throughout the clinical study. Quality control sampling was performed on all data, and focused consensus training was used as an efficient recalibration procedure when disagreements in findings were identified.

**Methods**

The current study sought to determine whether established interobserver reliability of four palpatory diagnostic tests for evaluating lumbar somatic dysfunction could be maintained during a clinical observational study. There were 6 data collection sessions over the 4-month period of the study, quality control sampling was performed on all data, and occasional recalibration procedures were performed. Study participants were recruited from the faculty, staff, and students of Kirksville College of Osteopathic Medicine–A.T. Still University (KCOM) in Missouri, as well as from the surrounding rural community.

Individuals of both sexes and those with and without low back pain were included in the study. Individuals were excluded if they had known congenital vertebral anomalies or other conditions (eg, fractures) that could potentially alter lumbar bony anatomic features or if they had received spinal manipulation within the 8 weeks preceding the study. The KCOM Institutional Review Board reviewed and approved the study design. All participants signed institutional review board–approved informed consent forms prior to participation in the study.

Two examiners who are American Osteopathic Association board certified in neuromusculoskeletal medicine—with 10 years (B.F.D.) and 3 years (K.T.S.) of clinical experience—performed the tests in the study. Pretraining levels of interobserver reliability were determined for the following four segmental vertebral palpatory diagnostic tests: positional asymmetry of the transverse processes in the horizontal plane, tissue texture abnormalities, segmental rotational motion asymmetry, and pain provocation to assess tenderness.9

As described in the previous study,9 consensus training was performed following pretraining interobserver reliability assessment. After consensus training was completed, the interobserver reliability was again tested, revealing improved reliability for positional asymmetry, tissue texture abnormalities, and tenderness. However, because of insignificant improvement and overall poor reliability of the method for measuring rotational motion asymmetry during consensus training,9 that test was changed for the current study to resistance to anterior springing on the spinous processes. Results for pretraining and posttraining periods for the two examiners in the current study were obtained from our previous study.9

After assessment of posttraining interobserver reliability, a 2-month interval occurred before the maintenance of interobserver reliability was studied for another 4-month period. During these 4 months, the protocol remained consistent with the method developed during the consensus training study.9 The time between the 6 data collection sessions of the current study ranged from 1 to 8 weeks, depending on the availability of examiners and participants.

Participants were examined in the prone position. The spinous processes of L1 through L4 were identified by palpation. L5 was not included because of its high frequency of anatomic variability, which complicates palpation. A mark was placed on the skin overlying and bisecting the vertical mid portion of the spinous processes. Both examining osteopathic physicians agreed on the placement of these marks, which were used to ensure consistent localization of the vertebrae being tested.

Throughout the study, the examiners alternated in regard to which one evaluated the participants first. In each of the 6 data collection sessions, both examiners performed the tests in the same sequence (Figure), and each test (other than pain provocation to assess tenderness) was performed 2 or 3 times to determine a consistent outcome (ie, at least 2 of 3 tests in agreement). Pain provocation was performed 1 or 2 times to avoid causing hypersensitivity.

The presence or absence of physical findings was recorded in a blinded manner after each test was completed. In addition, the examiners were blinded regarding participants’ histories of low back pain. After both examiners completed their testing and recording of findings, they compared results. If lack of agreement for a specific test was found, that test was repeated
by both examiners in an unblinded manner. The repeated testing allowed each examiner to observe the other’s techniques (eg, location of digits, amount and direction of forces), leading to an understanding of why the findings were different and to a consensus on findings. This process functioned as an efficient and meaningful method of recalibration. Only data collected in a blinded manner were used for the assessment of interobserver reliability.

**Statistical Analysis**

Data used in the statistical analysis were collected in a blinded manner at pretraining and posttraining and at the 6 data collection sessions of the clinical observational study. The data were analyzed for interobserver reliability by using Cohen’s κ coefficient, 95% confidence intervals for κ, and percent agreement. The κ values were interpreted using the scale established by Landis and Koch,11 as follows: 0.81-1.00 indicates almost perfect reliability; 0.61-0.80, substantial reliability; 0.41-0.60, moderate reliability; 0.21-0.40, fair reliability; 0-0.20, slight reliability; and <0, poor reliability. Acceptable reliability was defined as κ ≥ 0.40.12 Analyses were conducting using SAS 9.2 software (SAS Institute Inc, Cary, North Carolina).

Prevalence of specific somatic dysfunctions was calculated as the percent of examinations in which the physical finding was present. To test for changes in probability of agreement between the examiners from the pretraining period to the data collection sessions, P values were calculated using logistic regression analysis. For resistance to anterior springing on the spinous processes, change in probability of agreement was analyzed in comparison to the postraining reliability session because of the change in this palpatory diagnostic test between the pretraining and the postraining periods.

Logistic regression analysis was also used to test for the effect of participant demographic characteristics (ie, age, body mass index [BMI], sex, and presence of low back pain) on the probability of agreement between the two examiners.

**Results**

Sixty-four participants were included in the current study. Thirty-three participants were included in session 1, 31 in session 2, 14 in session 4, and 17 each in sessions 3, 5, and 6. Participants were aged between 20 and 40 years, with a mean (SD) age of 30 (6) years. Of the 64 participants, 17 (25%) had low back pain and 48 (75%) did not. Forty-nine (77%) of the study participants were women.

Over the 6 data collection sessions during the 4-month study period, improvement in interobserver reliability was obtained for all tests. All four palpatory tests demonstrated acceptable interobserver reliability, as defined by κ ≥ 0.40. For static segmental positional asymmetry of the transverse processes in the horizontal plane, agreement between examiners significantly improved from the pretraining session to the 6 data collection sessions (Table 1). The κ values in the 6 sessions ranged from 0.56 to 0.72, with moderate to substantial reliability achieved in all sessions.

For tissue texture abnormalities, agreement between examiners also significantly improved from the pretraining session to the 6 data collection sessions (Table 2). The κ values ranged from 0.60 to 0.72, indicating moderate to substantial reliability.
from 0.23 to 0.55, with moderate reliability achieved in 5 of the 6 sessions.

For resistance to anterior springing on the spinous processes, no pretraining data were available because, as previously mentioned, the test used for this category of palpatory diagnosis was changed between the pretraining and posttraining sessions. Between the post consensus training session and the 6 data collection sessions, improvement in agreement between examiners was not significant (Table 3). The $\kappa$ values ranged from 0.30 to 0.50, with moderate reliability achieved in 3 of the 6 sessions.

For pain provocation to access tenderness, agreement between examiners significantly improved from the pretraining session to 5 of the 6 data collection sessions (Table 4). The $\kappa$ values in the 6 sessions ranged from 0.61 to 0.88, with substantial to almost perfect reliability achieved in all sessions.

Demographic characteristics of study participants (Table 5) had little impact on the reliability of the tests during the current study. The BMI of participants was not significantly related to interobserver reliability for static segmental positional asymmetry of the transverse processes in the horizontal plane ($P=.17$), tissue texture abnormalities ($P=.75$), resistance to anterior springing on the spinous processes ($P>.99$), or tenderness ($P=.15$). Interobserver reliability was significantly greater when examining women than men for static segmental positional asymmetry of the transverse processes in the horizontal plane ($P=.04$), but not for tissue texture abnormalities ($P=.36$), resistance to anterior springing on the spinous processes ($P=.11$), or tenderness ($P=.22$).

Participant age was significantly related to interobserver
mental positional asymmetry of the transverse processes in the horizontal plane had similar \( \kappa \) values to pain provocation to assess tenderness, showing a level of reliability in the moderate to substantial range. Such a high level of reliability for static segmental positional asymmetry of the transverse processes in the horizontal plane, as far as we are aware, has not previously been reported.

Only one session for tissue texture abnormalities (session 4) resulted in a \( \kappa \) value lower than the posttraining level. Overall, the test for tissue texture abnormalities showed moderate reliability, and the test for resistance to anterior springing on the spinous processes showed fair to moderate reliability.

A comparison of the session 4 cohort vs the cohorts of the other sessions may explain the reduction in the \( \kappa \) value for tissue texture abnormalities for that session. Session 4 had the reliability for resistance to anterior springing on the spinous processes \( (P=.01) \), with reliability being poorer in older participants. Participant age was not significantly related to interobserver reliability for static segmental positional asymmetry of the transverse processes in the horizontal plane \( (P=.82) \), tissue texture abnormalities \( (P=.82) \), or tenderness \( (P=.54) \).

**Comment**

Results of the current study indicate that acceptable interobserver reliability can be maintained and improved over 4 months with occasional recalibration consisting of focused consensus training. In a comparison of individual palpatory diagnostic tests, pain provocation to assess tenderness had the highest \( \kappa \) values for almost all assessments. This result is consistent with previously reported findings.\(^3,4\) Static segmental positional asymmetry of the transverse processes in the horizontal plane had similar \( \kappa \) values to pain provocation to assess tenderness, showing a level of reliability in the moderate to substantial range. Such a high level of reliability for static segmental positional asymmetry of the transverse processes in the horizontal plane, as far as we are aware, has not previously been reported.

Only one session for tissue texture abnormalities (session 4) resulted in a \( \kappa \) value lower than the posttraining level. Overall, the test for tissue texture abnormalities showed moderate reliability, and the test for resistance to anterior springing on the spinous processes showed fair to moderate reliability.

A comparison of the session 4 cohort vs the cohorts of the other sessions may explain the reduction in the \( \kappa \) value for tissue texture abnormalities for that session. Session 4 had the

**Table 3**

Interobserver Reliability and Maintenance of Reliability for Resistance to Anterior Springing on the Spinous Processes (N=64)

<table>
<thead>
<tr>
<th>Testing Session (n)</th>
<th>Study Week</th>
<th>Total Vertebrae Examined, No.</th>
<th>Agreement, %</th>
<th>( \kappa ) (95% CI)</th>
<th>Prevalence, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posttraining (14)</td>
<td>-8</td>
<td>60</td>
<td>70</td>
<td>0.29 (0.03-0.55)</td>
<td>73</td>
</tr>
<tr>
<td>Session 1 (33)</td>
<td>0</td>
<td>132</td>
<td>70</td>
<td>0.30 (0.12-0.47)</td>
<td>72</td>
</tr>
<tr>
<td>Session 2 (31)</td>
<td>8</td>
<td>124</td>
<td>73</td>
<td>0.46 (0.30-0.62)</td>
<td>55</td>
</tr>
<tr>
<td>Session 3 (17)</td>
<td>9</td>
<td>68</td>
<td>76</td>
<td>0.50 (0.29-0.71)</td>
<td>50</td>
</tr>
<tr>
<td>Session 4 (14)</td>
<td>10</td>
<td>56</td>
<td>75</td>
<td>0.31 (0.04-0.58)</td>
<td>52</td>
</tr>
<tr>
<td>Session 5 (17)</td>
<td>13</td>
<td>68</td>
<td>71</td>
<td>0.37 (0.15-0.60)</td>
<td>57</td>
</tr>
<tr>
<td>Session 6 (17)</td>
<td>17</td>
<td>68</td>
<td>76</td>
<td>0.44 (0.21-0.67)</td>
<td>61</td>
</tr>
</tbody>
</table>

*Because of a change in testing procedures between the pretraining and posttraining sessions, comparisons for resistance to anterior springing are made to the posttraining session.

**Abbreviations:** CI, confidence interval; \( \kappa \), kappa coefficient.

**Table 4**

Interobserver Reliability and Maintenance of Reliability for Pain Provocation to Assess Tenderness Over the Spinous Processes (N=64)

<table>
<thead>
<tr>
<th>Testing Session (n)</th>
<th>Study Week</th>
<th>Total Vertebrae Examined, No.</th>
<th>Agreement, %</th>
<th>( \kappa ) (95% CI)</th>
<th>Prevalence, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretraining (42)</td>
<td>-38</td>
<td>168</td>
<td>75</td>
<td>0.32 (0.16-0.49)</td>
<td>24</td>
</tr>
<tr>
<td>Posttraining (3)</td>
<td>-26</td>
<td>12</td>
<td>83</td>
<td>0.56 (0.01-1.00)</td>
<td>25</td>
</tr>
<tr>
<td>Session 1 (33)</td>
<td>0</td>
<td>132</td>
<td>91*</td>
<td>0.70 (0.54-0.86)</td>
<td>18</td>
</tr>
<tr>
<td>Session 2 (31)</td>
<td>8</td>
<td>124</td>
<td>91*</td>
<td>0.68 (0.50-0.86)</td>
<td>17</td>
</tr>
<tr>
<td>Session 3 (17)</td>
<td>9</td>
<td>68</td>
<td>97*</td>
<td>0.88 (0.72-1.00)</td>
<td>15</td>
</tr>
<tr>
<td>Session 4 (14)</td>
<td>10</td>
<td>56</td>
<td>82</td>
<td>0.61 (0.40-0.83)</td>
<td>36</td>
</tr>
<tr>
<td>Session 5 (17)</td>
<td>13</td>
<td>68</td>
<td>94*</td>
<td>0.81 (0.63-0.99)</td>
<td>19</td>
</tr>
<tr>
<td>Session 6 (17)</td>
<td>17</td>
<td>68</td>
<td>87*</td>
<td>0.65 (0.45-0.85)</td>
<td>24</td>
</tr>
</tbody>
</table>

*Significant difference in examiner agreement compared to pretraining session \( (P<.001) \).

**Abbreviations:** CI, confidence interval; \( \kappa \), kappa coefficient.
The diagnostic procedures may have irritated the back, causing tissues to react to the palpatory pressures and resulting in changes in baseline pain characteristics. In future studies, participants could complete preexamination and postexamination questionnaires to assess their perceptions of any symptom change as a result of the examination. An objective measure, such as algometry, could also be used, though use of such a measure would defeat the purpose of evaluating the usefulness of diagnostic palpation. Further refinement of the descriptors used for evaluating nuances of tissue texture abnormalities may be helpful during consensus training, especially for symptomatic patients.

In contrast, static segmental positional asymmetry of the transverse processes in the horizontal plane had the highest level of reliability for session 4 compared to the other sessions. Perhaps positional asymmetry is related to the cause of low back pain and is more prominent in individuals with low back pain as a result of tissue hypertonicity, a condition commonly found in individuals with low back pain.

When assessing static segmental positional asymmetry of the transverse processes in the horizontal plane, interobserver reliability was significantly higher for women than for men. This result may be caused by commonly found hypertrophy and hypertonicity of the erector spinae musculature in men, making localization of the transverse process more difficult. Further study is required to determine if the higher interobserver reliability of positional asymmetry in women is a consistent finding and, if so, why it occurs.

Although the assumption that heavier individuals are more difficult to palpate than lighter individuals may seem to be valid, data from the current study’s predominantly overweight population did not substantiate this assumption. The mean (SD) BMI for the 6 sessions ranged from 25.8 (4.1) to 26.5 (5.7). No statistically significant effect of BMI on the reliability of any one type of test was observed in this study. Because of the narrow range of BMIs of study participants, additional studies on individuals with a wider range of BMIs are needed to determine whether the thickness of soft tissue impedes the reliability of these tests.

Similar to previous findings for motion testing, resistance to anterior springing on the spinous processes in the current study had the poorest level of interobserver reliability. Assessment of reliability assumes, by definition, that the characteristics being observed are stable. During recalibration, motion characteristics observed by the initial examiner often changed, indicating that the finding for resistance to anterior springing was not stable with repetitive testing. Although this change was not quantified in the current study, the fair to moderate reliability of this test reinforces findings from previous studies and calls into question the reproducibility of vertebral motion testing on a segmental basis.

Alternatively, we may not know the best way to interpret findings of segmental motion tests. Statistical analysis suggested that reliability of the resistance to anterior springing motion test was significantly related to participant age. However, closer analysis of the data indicated that this difference, though statistically significant, may not be clinically significant. This difference in reliability was observed between participants aged 21 to 25 years and those older than 25 years. Differences in the prevalence of findings could account for the statistically significant relationship of age to reliability.

### Table 5

Demographic Characteristics of Participants in Study of Interobserver Reliability of Osteopathic Palpatory Tests (N=64)

<table>
<thead>
<tr>
<th>Testing Session (n)</th>
<th>Study Week</th>
<th>Sex</th>
<th>Presence of Low Back Pain, No. (%)</th>
<th>Age, mean (SD)</th>
<th>Body Mass Index, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretraining (42)</td>
<td>-38</td>
<td>Men, No. (%)</td>
<td>Women, No. (%)</td>
<td>...</td>
<td>26 (4)</td>
</tr>
<tr>
<td>Posttraining* (3)</td>
<td>-26</td>
<td>2 (67)</td>
<td>1 (33)</td>
<td>...</td>
<td>27 (1)</td>
</tr>
<tr>
<td>Posttraining* (30)</td>
<td>-20</td>
<td>14 (47)</td>
<td>16 (53)</td>
<td>...</td>
<td>29 (6)</td>
</tr>
<tr>
<td>Posttraining* (14)</td>
<td>-8</td>
<td>7 (50)</td>
<td>7 (50)</td>
<td>...</td>
<td>36 (13)</td>
</tr>
<tr>
<td>Session 1 (33)</td>
<td>0</td>
<td>8 (24)</td>
<td>25 (76)</td>
<td>2 (6)</td>
<td>30 (6)</td>
</tr>
<tr>
<td>Session 2 (31)</td>
<td>8</td>
<td>8 (26)</td>
<td>23 (74)</td>
<td>2 (6)</td>
<td>31 (6)</td>
</tr>
<tr>
<td>Session 3 (17)</td>
<td>9</td>
<td>2 (12)</td>
<td>15 (88)</td>
<td>0</td>
<td>27 (7)</td>
</tr>
<tr>
<td>Session 4 (14)</td>
<td>10</td>
<td>5 (36)</td>
<td>9 (64)</td>
<td>14 (100)</td>
<td>29 (6)</td>
</tr>
<tr>
<td>Session 5 (17)</td>
<td>13</td>
<td>2 (12)</td>
<td>15 (88)</td>
<td>0</td>
<td>27 (7)</td>
</tr>
<tr>
<td>Session 6 (17)</td>
<td>17</td>
<td>2 (12)</td>
<td>15 (88)</td>
<td>0</td>
<td>27 (7)</td>
</tr>
</tbody>
</table>

* Posttraining for the four palpatory tests occurred during different testing sessions.
... indicates data not collected.
In other studies,\textsuperscript{13,14} investigators have preconditioned vertebral segments by repetitively inducing motion before the actual testing to improve the likelihood of stable findings. Because preconditioning is not commonly performed in practice or taught in osteopathic medical training programs, it was not used in the current study.

Although consensus training has been shown to be useful in establishing interobserver reliability, such findings are meaningless unless adequate interobserver reliability can be demonstrated throughout clinical research studies that use palpation as a diagnostic tool or as an outcome measure. Consequently, development of quality control and recalibration methods that can be effectively performed in clinical research is crucial.

In our previous study on the effect of consensus training on interobserver reliability,\textsuperscript{9} comprehensive consensus training involved three phases. The first phase was pretraining interobserver reliability assessment and involved defining and standardizing testing procedures and interpretation. The second phase was consensus training, which involved simultaneous evaluation of numerous individuals by all study examiners to clarify the testing procedures and to refine examiner skills to the consensus standard. The third phase was post-training interobserver reliability assessment and involved reassessment of examiner skills. This training established calibration parameters for specific palpatory tests for that group of examiners. The interobserver reliability assessment in post-training demonstrated the effectiveness of the other two phases of comprehensive consensus training.\textsuperscript{9} In the current study, quality control procedures with recalibration, consisting of focused consensus training, were shown to be successful in maintaining and improving the interobserver reliability of the two examiners.

The recalibration procedures of the current study are distinct from the comprehensive consensus training procedures previously published.\textsuperscript{9} Recalibration in the current study does not involve the simultaneous evaluation of every participant by every examiner, as occurs in comprehensive consensus training, in which the definition, performance, and interpretation of tests are refined. During recalibration, no changes or refinements are made in the definition, intended performance, or interpretation of the techniques. Furthermore, recalibration is performed only when a disagreement exists in findings. As a result, feedback from recalibration is efficient and can be performed without derailing clinical research studies that use diagnostic palpation.

In the current study, percent agreement ranged from 71% to 97% for each data collection session. Thus, for 80% of the examinations, no interaction occurred between examiners. If recalibration with focused consensus training is required for more than 25% of the tests performed, we propose that comprehensive consensus training needs to be repeated.

Acceptable \( \kappa \) values were achieved and reliability was improved over time for several palpatory tests in the lumbar region in the current study. However, the study’s confidence intervals were large, with widths ranging from 0.23 to 0.66. The lower bound of these confidence intervals periodically extended below the range of acceptable reliability (\( \kappa < 0.40 \)).\textsuperscript{12} Therefore, widths of the confidence intervals should be considered when interpreting results in this type of study. The accuracy of estimated \( \kappa \) values, as measured by the margin of error (ie, half the width of the confidence interval), is affected by the prevalence rates of test findings, the sample size, and the true (ie, actual) level of reliability. For example, a \( \kappa \) value estimated to be 0.44—classified as acceptable within the current standard—may actually be as low as 0.21, which is only fair reliability.

To help assess the feasibility of future reliability studies, the margins of error for 95% confidence intervals with various prevalence values and true \( \kappa \) values were determined for different sample sizes (Table 6). As shown in Table 6, a palpatory test with a true \( \kappa \) of 0.60 and a prevalence rate of the test findings of 80% would require 120 observations to establish 95% confidence that the palpatory test has a \( \kappa \) of at least 0.40. However, if the test had a true \( \kappa \) of 0.70 and a prevalence rate of 70%, the same level of confidence could be achieved using only 30 observations. If estimating reliability with sufficient accuracy (ie, narrow confidence intervals) requires 120 observations,
server reliability during clinical studies that involve diagnostic palpation in either the evaluation of study participants or as an outcome measure can be accomplished with a two-step process: (1) establishing adequate interobserver reliability ($\kappa > 0.40$) using comprehensive consensus training and (2) maintaining interobserver reliability, initially using quality control sampling of all data and, when examiners do not agree on palpatory findings, using focused consensus training. After reliability has been sustained at an appropriate level over several data collection sessions, issues related to determining the frequency of quality control sampling (eg, daily, weekly, or monthly) should be considered in light of various factors, such as the number of landmarks being assessed per study participant, the prevalence of positive findings, the number of examiners, and the number of data collection sites.

Further study is needed to determine the optimal frequency for quality control assessments and the conditions necessary to enable a reduction in the quality control sampling rate. It is likely that such standards may vary depending on the examiners performing the research.

The current study has several strengths and limitations. The study was conducted in a busy patient care setting, demonstrating the feasibility of practicing physicians performing this type of research. The difference in clinical experience between the two examiners supports generalizability of the findings, because maintenance and improvement of interobserver reliability was established for those with levels of clinical experience ranging from 3 years to 10 years. The participant demographics of the current study are a limitation as a result of the narrow age and BMI ranges, the predominance of women, and high prevalence rates of certain somatic dysfunctions.

Conclusions
The current study indicates that quality control sampling on all data, with occasional recalibration consisting of focused consensus training, can maintain and improve the interobserver reliability of commonly used palpatory diagnostic tests of the lumbar spine in an efficient and meaningful manner. The study design may be particularly useful in manual therapy research and educational programs. Future studies should consider the prevalence rates of test findings when determining sample size and study feasibility, as well as the confidence intervals when interpreting the reliability of test findings.

Acknowledgments
The authors thank Patty Lyons for her technical support and Deborah Goggin, MA, for her editorial support. This study was supported by grants from the National Institutes of Health—National Center for Complementary and Alternative Medicine, grant no. 1R01AT00305, and the American Osteopathic Association, grant no. 00-04-505.

References