Intrarater and Interrater Reliability of Assessment of Lumbar Multifidus Muscle Thickness Using Rehabilitative Ultrasound Imaging

Ultrasound imaging has been used previously to characterize the dimensions of the multifidus muscle in healthy populations. Measurements have most commonly been performed in the transverse plane, reporting cross-sectional area (CSA) and linear values for the muscle at various lumbar vertebral levels. Reliability of performing these measures has been previously demonstrated. Between-side symmetry in size has also been demonstrated in healthy populations, and, when linear measurements (anteroposterior distance \([\text{thickness}] \times \text{mediolateral distance [width]}\)) are multiplied, the results correlated well with CSA at the L4 and L5 vertebral levels (range, \(r = 0.92-0.98\)). Validity of CSA measurements has been established by comparing measurements of the lumbar multifidus muscle established using ultrasound imaging with those obtained by magnetic resonance imaging (MRI).

An alternative to imaging the multifidus muscle in transverse section is to adopt a parasagittal (longitudinal) orientation of the transducer. In this plane, the zygapophyseal joints, the overlying multifidus muscle bulk at 2 to 3 vertebral levels, and the thoracolumbar fascia can be visualized. Apart from the advantage of being able to visualize more than 1 vertebral level at a time, this orientation allows measurement of the thickness of the muscle, and muscle contraction can be observed more easily than in the transverse plane. Muscle contraction is seen on the ultrasound image as an increase in thickness of the muscle as it shortens along its length. The parasagittal view...

**STUDY DESIGN:** Within-session intrarater and interrater reliability study.

**OBJECTIVE:** To establish the intrarater and interrater reliability of thickness measurements of the multifidus muscle in a parasagittal plane, conducted by an experienced ultrasound operator and a novice assessor.

**BACKGROUND:** There is considerable evidence for the importance of the multifidus muscle in segmental stabilization of the lumbar spine. The cross-sectional area of the multifidus muscle has been assessed in healthy subjects and patients with low back pain using real-time ultrasound imaging. However, few studies have measured the thickness of the multifidus muscle using a parasagittal view.

**METHODS AND MEASURES:** The thickness of the multifidus muscle was measured at rest, using real-time ultrasound imaging, in 10 subjects without a history of low back pain, at the levels of the L2-3 and L4-5 zygapophyseal joints. The measure was carried out 3 times at each level by 2 assessors (1 experienced, 1 novice). Intrarater (model 3) and interrater (model 2) reliability was assessed by calculation of an F statistic (analysis of variance), the intraclass correlation coefficient (ICC), and the standard error of measurement (SEM).

**RESULTS:** On the basis of an average of 3 trials, the 2 operators showed very high interrater agreement on the measurement of thicknesses at the L2-3 level (ICC = 0.96; 95% CI: 0.84 to 0.99) and the L4-5 vertebral level (ICC = 0.97; 95% CI: 0.87 to 0.99), with no systematic differences in muscle size across operators \((P > .05)\). Intrarater reliability was relatively lower for the L2-3 level (ICC = 0.85; 95% CI: 0.51 to 0.96) than the L4-5 level (ICC = 0.87; 95% CI: 0.52 to 0.97) when a single trial per rater was used, but these values still indicated a high level of agreement. In addition, the novice and experienced operator produced reliable intrarater measurements at L2-3 (ICC = 0.89; 95% CI: 0.72 to 0.97 and 0.94; 95% CI: 0.86 to 0.99) and at L4-5 (ICC = 0.88; 95% CI: 0.68 to 0.97 and 0.95; 95% CI: 0.86 to 0.99), with no systematic differences in muscle size across trials \((P > .05)\). The consistently low SEM values also indicate low measurement error.

**CONCLUSION:** A novice and an experienced assessor were both able to reliably perform this measure at rest for 2 vertebral levels using real-time ultrasound imaging. An average of 3 trials produced higher intrarater reliability scores, though using a single trial per rater was also reliable. J Orthop Sports Phys Ther 2007;37(10):608-612. doi:10.2519/jospt.2007.2418

**KEY WORDS:** back muscles, lumbar spine, muscle assessment, repeatability, ultrasonography
has already been successfully used to provide feedback of multifidus muscle recruitment during 2 randomized controlled trials. Hides et al showed that subjects with first-episode low back pain had difficulty recruiting the multifidus muscle and first described the use of ultrasound imaging of the multifidus in the parasagittal plane to provide direct feedback to the subjects. In a separate study using the parasagittal view as a form of visual feedback, a group of subjects without low back pain were able to demonstrate greater improvements in both the performance and retention of a voluntary isometric contraction of the multifidus muscle, when compared with a group who did not receive the ultrasound biofeedback training.

While it appears that imaging and measuring the multifidus muscle in parasagittal section may be useful in clinical practice, it is necessary to examine the validity of the measure and the reliability of assessors performing the thickness measure. With regard to validity, the relationship between multifidus muscle thickness change on ultrasound imaging and electromyographic (EMG) signal amplitude has been examined in healthy subjects. The authors showed that muscle thickness change was associated with EMG signal amplitude at the L4 vertebral level (r = .79, \(P<.001\)), for a range between 19% to 34% of maximal voluntary isometric contraction. Two studies have previously reported on the reliability of raters performing thickness measurements of the multifidus muscles in parasagittal section, and a similar reliability study has been performed measuring the thickness of the lumbar erector spinae muscles. Keisel et al conducted on-screen measurements of 8 healthy subjects and reported intrarater reliability values using intraclass correlation coefficients (ICC) of 0.85. Images of the multifidus were also saved for later measurement by 2 raters, and high interrater reliability was also demonstrated (ICC = 0.95) when measuring the same image from the display unit. Van et al also reported high reliability for 2 raters who measured 6 healthy subjects (intrarater ICC = 0.97 to 0.98, interrater ICC = 0.98). For the lumbar erector spinae muscles, Watanabe et al reported high intrarater (r = 0.938-0.962) and interrater (r = 0.9-0.962) reliability estimates.

The aim of this investigation was to examine intrarater and interrater reliability of 2 raters (1 experienced and 1 novice tester) performing assessment of lumbar multifidus muscle thickness in parasagittal section in healthy subjects at rest, using real-time ultrasound imaging. The measurements were conducted at rest to allow establishment of standardized measurement protocols in relation to variables, such as location of anatomical landmarks, subject positioning, cursor location, and transducer pressure, without the added variable of muscle contraction. To avoid the confounding influences of pain and muscle inhibition, a healthy subject population was selected.

**METHODS**

**Subjects**

Ten healthy subjects (2 males, 8 females) participated in the study. Their mean ± SD age was 30.8 ± 18.1 years, height was 165 ± 6.1 cm, and body mass was 58.2 ± 7.4 kg. The exclusion criteria for the study were pregnancy, a history of low back pain, spinal abnormality, spinal or abdominal surgery or severe trauma, neuromuscular or joint disease, and training involving the back muscles within the previous 3 months. This study was approved by the Medical Research Ethics Committee at The University of Queensland, Australia. Informed consent was obtained and the rights of human subjects were protected.

**Procedures**

Ultrasound imaging assessment of multifidus muscle thickness was conducted using a Diasonics Synergy ultrasound imaging apparatus equipped with a 5-MHz curvilinear transducer (GE Diasonics Ultrasound Inc, Santa Clara, CA). The ultrasound measurements were performed by 1 experienced and 1 novice assessor. The experienced operator had 17 years experience in assessing muscles using ultrasound imaging. Training of the novice rater included basic education in ultrasound imaging techniques (3 hours) and observation of the experienced operator with instruction in the measurement technique (3 hours). In addition, the novice rater and the experienced operator measured 5 subjects prior to commencing the study. Both raters were experienced physical therapists familiar with clinical assessment of the lumbar spine and location of bony landmarks in the region.

For assessment by ultrasound imaging, the subject was positioned in prone lying, with a pillow placed under the abdomen to minimize the lumbar lordosis (Figure 1). Each rater manually palpated the lumbar vertebral levels and marked the location of the spinous processes on the skin with a pen. Raters were allowed to confirm their own skin markings using ultrasound imaging in parasagittal section (verified by identification of the sacrum and counting spinous processes on progression in a cephalad direction). All skin markings were removed with alcohol after the first rater conducted measurements, and the subject stood up and moved around. This ensured that each rater independently established anatom-
Raters were not allowed to observe each other conducting measurements.

The multifidus muscle was imaged in parasagittal (longitudinal) section, allowing visualization of the zygapophyseal joints, multifidus muscle bulk, and thoracolumbar fascia (Figure 2). Examiners were careful to maintain the transducer in the plane of the zygapophyseal joints. With this technique, if the transducer is placed too far laterally, the appearance is similar but in fact it is the lumbar erector spinae muscles and the transverse processes that are imaged rather than the multifidus muscle and the zygapophyseal joints. The thickness of the multifidus muscle was measured at the levels of L2-3 and L4-5 zygaphophyseal joints using on-screen calipers. Two vertebral levels were selected, as the morphology of the multifidus muscle varies according to vertebral level. Linear measurements were conducted from the tip of the target zygapophyseal joint to the inside edge of the superior border of the multifidus muscle (Figure 2). The measure was carried out 3 times by each assessor for each of the 2 vertebral levels assessed.

Reliability analyses were conducted using the SPSS statistical package Version 14 (SPSS Inc, Chicago, IL). This analysis produces an ICC that represents consistency in the rank order of scores and an F statistic, which represents systematic change in size of scores. Separate analyses were conducted for data at the L2-3 and L4-5 vertebral levels to assess consistency between the 2 raters for 1 trial (ICC2,1), and for 3 trials (ICC2,3), and across the 3 trials for each rater (ICC3,1). In addition, consistency of measurement was examined across operators and trials by calculation of the standard error of the measurement (SEM). SEM is the difference between the actual measured score across trials and an estimated “true” score, and was calculated as $SEM = \text{pooled SD} \times \left(1 - ICC^{\frac{1}{2}}\right)$.

### RESULTS

Descriptive data (mean ± SD) of the multifidus muscle thickness at rest for both raters at the 2 vertebral levels measured, averaged across the 3 trials, is shown in Table 1. The results of the interrater and intrarater reliability analyses are shown in Tables 2 to 4. The novice rater and experienced rater demonstrated a high level of agreement between them for measurement of multifidus thickness from a single trial (Table 2) and from the average of 3 trials (Table 3). In addition, each rater was reliable across trials (Table 4). Nonsignificant F statistics indicate no systematic difference across rater and trial. SEM values suggest stability of measuring thickness of the multifidus muscle in a parasagittal plane.

### DISCUSSION

Results from this study showed that a novice and an experienced assessor were both able to reliably measure lumbar multifidus muscle thickness at 2 vertebral levels using real-time ultrasound imaging. Using an average of 3 trials per rater produced high interrater reliability scores (ICC2,3 = 0.96 for L2-3 and ICC2,3 = 0.97 for L4-5). Actual differences between mean values obtained by the 2 raters were very small in terms of the actual thickness of the muscle (0.07 cm for L2-3 and 0.04 cm for L4-5).
In addition, reliability of measurements for both vertebral levels was similar. The novice and experienced operators exhibited high intrarater reliability (ICC$_{3,1}$ = 0.89 and 0.94 for L2-3 and ICC$_{3,1}$ = 0.88 and 0.95 for L4-5) and high interrater reliability (ICC$_{2,3}$ = 0.85 for L2-3 and ICC$_{2,3}$ = 0.87 for L4-5) when a single trial per rater was used. The overall results are comparable with those of Watanabe et al$^{12}$ for similar thickness measurements performed on the lumbar erector spinae muscles. In their study, interoperator reliability ranged from 0.9 to 0.96 for measurements made in 3 different trunk positions, and intraobserver repeatability ranged from 0.94 to 0.96.$^{13}$ The results of the current study are also comparable with the results obtained by Van et al$^{10}$ and Kiesel et al$^{7}$ for intrarater reliability (ICC$_{3,1}$ = 0.97-0.98 and ICC$_{3,1}$ = 0.85, respectively) and interrater reliability (ICC$_{2,3}$ = 0.98 and ICC$_{3,1}$ = 0.95, respectively). However, there are differences in the methodologies of these studies. Van et al$^{10}$ used a similar protocol to that of the current study, while Kiesel et al$^{7}$ saved ultrasound images of the multifidus for later measurement by 2 raters to establish interrater reliability of making measurements off the same image. In the current study and that of Van et al$^{10}$ each subject was imaged 3 times (requiring accurate repositioning of the transducer and reproducible pressure on the transducer), and each rater performed his/her own images after positioning the subject and identifying the bony landmarks. A difference is that in the current study 2 vertebral levels were measured.

In addition to the findings of high intrarater and interrater reliability for both the experienced and the novice assessors, the reported SEM values were low when compared to the resting thickness value of the muscle. This is important, as the most likely clinical application of the measurement is to assess changes in thickness due to muscle contraction. A low SEM value relative to the resting value would suggest that the ability to detect a real change (exceeding measurement error) would be likely. The smallest real difference (SRD) can be calculated to indicate the magnitude of change that would exceed the expected trial to trial variability.$^6$ Based on the SEM of 0.06 cm for the experienced operator at L4-5, the SRD can be calculated by the following formula: $SEM \times \sqrt{2} \times 2.26 \, cm$, where 2.26 represents the value of the $t$ distribution for a 95% CI ($df = 9$). If this value is divided by the thickness of the muscle at the L4-5 vertebral level (2.86 cm), a change in thickness of 6.6 % would be required to be 95% confident that a real change has occurred. The study of Van et al$^{10}$ reported a mean increase of 11.6% in thickness of the multifidus due to voluntary isometric contractions in a group that received ultrasound biofeedback. Kiesel et al$^7$ measured multifidus thickness in subjects who performed 4 levels of graded resistance, and percentage increases in muscle thickness ranged from 32% to 48%. Furthermore, similar to the results reported by Springer et al$^{10}$ who measured the thickness of the lateral abdominal muscles, the SEM was lower when an average of 3 measurements was used (TABLES 2 and 3). Although the SEM would need to be assessed in each experimental situation, these results would suggest that thickness measurements of multifidus may be clinically useful, with potential to detect changes that exceed measurement error.

Before the measurement of multifidus

### TABLE 2

<table>
<thead>
<tr>
<th>Vertebral Level</th>
<th>ICC$_{3,1}$</th>
<th>95% CI</th>
<th>SEM (cm)</th>
<th>F (P)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2-3</td>
<td>0.85</td>
<td>(0.51-0.96)</td>
<td>0.13</td>
<td>1.12 (.32)</td>
</tr>
<tr>
<td>L4-5</td>
<td>0.87</td>
<td>(0.52-0.97)</td>
<td>0.10</td>
<td>0.84 (.39)</td>
</tr>
</tbody>
</table>

* Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; SEM, standard error of measurement.

* P >.05 indicates no systematic difference in muscle thickness values between raters.

### TABLE 3

<table>
<thead>
<tr>
<th>Vertebral Level</th>
<th>ICC$_{2,3}$</th>
<th>95% CI</th>
<th>SEM (cm)</th>
<th>F (P)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2-3</td>
<td>0.96</td>
<td>(0.84-0.99)</td>
<td>0.06</td>
<td>2.47 (.15)</td>
</tr>
<tr>
<td>L4-5</td>
<td>0.97</td>
<td>(0.87-0.99)</td>
<td>0.05</td>
<td>1.27 (.29)</td>
</tr>
</tbody>
</table>

* Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; SEM, standard error of measurement.

* P >.05 indicates no systematic difference in muscle thickness values between raters.

### TABLE 4

<table>
<thead>
<tr>
<th>Vertebral Level/Tester</th>
<th>ICC$_{2,3}$</th>
<th>95% CI</th>
<th>SEM (cm)</th>
<th>F (P)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2-3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novice</td>
<td>0.94</td>
<td>(0.98-0.96)</td>
<td>0.09</td>
<td>0.91 (.87)</td>
</tr>
<tr>
<td>Experienced</td>
<td>0.88</td>
<td>(0.86-0.96)</td>
<td>0.09</td>
<td>0.37 (.70)</td>
</tr>
<tr>
<td>L4-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Novice</td>
<td>0.88</td>
<td>(0.68-0.97)</td>
<td>0.09</td>
<td>0.37 (.70)</td>
</tr>
<tr>
<td>Experienced</td>
<td>0.95</td>
<td>(0.86-0.99)</td>
<td>0.06</td>
<td>2.05 (.16)</td>
</tr>
</tbody>
</table>

* Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; SEM, standard error of measurement.

* P >.05 indicates no systematic difference in muscle thickness values across trials.
CONCLUSION

The thickness of the multifidus muscle was measured in parasagittal section by an experienced assessor and a novice assessor. When a standardized protocol was followed, high interrater reliability, as well as high intrarater reliability, for both raters was demonstrated. This thickness measure may be useful in physiotherapy practice and research, as other studies have shown that changes in the thickness of the multifidus muscle occur when the muscle contracts. Future studies could use a similar measurement protocol and investigate the reliability of performing the thickness measurement in the contracted state, in different subject populations, and across days.

ACKNOWLEDGEMENTS

The authors wish to thank the subjects studied, the staff at the UQ/Mater Back Stability Clinic, and its Director, Ms Linda Blackwell, for their assistance.

REFERENCES