The Effect of Averaging Multiple Trials on Measurement Error During Ultrasound Imaging of Transversus Abdominis and Lumbar Multifidus Muscles in Individuals With Low Back Pain

Ultrasound imaging is increasingly used in rehabilitation to evaluate muscle morphology and function in patients with neuromusculoskeletal disorders such as low back pain (LBP). Some aspects of the reliability of TrA and LM muscle thickness measurements using RUSI have been established. While most studies report high reliability for muscle thickness and cross-sectional area measurements (intraclass correlation coefficients [ICCs] of greater than 0.85), standard error of measurement (SEM) and minimal detectible change (MDC) are usually large in comparison to mean measurements.

A few researchers have reported improvements in the reliability and precision of RUSI measurements by using the mean of 3 thickness measurements instead of a single measure of the TrA and LM. While both studies found good reliability (ICCs, >0.85) and precision (TrA SEM, <0.45 mm; LM SEM, <1.3 mm) with single measures, MDC for the TrA was relatively large compared to its mean thickness. For example, using the MDC for resting thickness, Teysen et al calculated a single measurement of...
resting TrA thickness would require a 41% change for one to be 95% confident that a change occurred. However, an average of 3 measurements resulted in a reduction in the MDC value such that a change occurred with only a 17% change in resting muscle thickness. This improvement in measurement precision is likely even more important when using percent change in muscle thickness ([thickness_{contracted} – thickness_{rest}]/thickness_{rest}), as these more clinically relevant measures incorporate the measurement error from both resting and contracted conditions.

While it is intuitive that the mean of multiple measures would reduce inter-repetition variability (thereby improving the reliability and precision of the measurements), the benefit must be weighed against the additional time and resources it takes to acquire more images. Shrout and Fleiss discussed the advantages of using a mean of ratings when the reliability of a single measure was deemed unacceptable. They purported that the number of ratings used to form the mean should be determined by a reliability study. To our knowledge, no such study has been performed investigating the effect on reliability of using a mean of other than 3 RUSI measures.

The primary purpose of this analysis was to investigate the improvements in measurement precision when using the mean value of percent change in muscle thickness of the TrA and LM muscles from multiple repetitions of a standardized task. We hypothesized that measurement precision would increase using the mean of measures, and that there would be a threshold, where averaging additional measurements would not result in large enough gains in precision to warrant the additional effort. The secondary purpose was to assess for a systematic order effect (eg, learning and/or fatigue) when using sequential measurements of percent change in TrA and LM muscle thickness during standardized tasks.

**METHODS**

**Participants**

Thirty volunteers, aged 18 to 60 years, with current LBP, with or without lower extremity symptoms, were recruited for this study. Exclusion criteria were prior history of lumbar surgery and presence of medical “red flags” of conditions such as cauda equina syndrome, major or rapidly progressing neurological deficit, fracture, cancer, infection, or systemic disease. Participants provided consent on forms approved by the University of Utah and Intermountain Healthcare Institutional Review Boards, and the rights of all participants were protected.

**Examiner**

A chiropractor (J.H.) that had not previously used RUSI in clinical practice served as the examiner for this reliability analysis. Prior to testing, he underwent 16 hours of practical training with a co-investigator (D.T.) experienced with the specific RUSI protocol.

**Procedures**

After signing consent forms, participants completed self-report measures, including demographic and historical information, and questionnaires on pain and disability. An 11-point numeric pain rating scale was used to assess current pain intensity and best and worst pain intensity in the past 24 hours. The modified Oswestry questionnaire was used to quantify self-reported disability. A standard physical examination was then performed to assess for study inclusion and exclusion criteria and to determine the participant’s symptomatic side, which was used for all subsequent RUSI images. If pain was evenly distributed, the imaged side was determined randomly.

All images were acquired with a Sonosite Titan ultrasound machine (Sonosite Inc, Bothell, WA) and a 60-mm, 2- to 5-MHz curvilinear array used in B-mode at 5 MHz. Image acquisition for each condition was performed 6 times. The examiner positioned the transducer and optimized the quality of each image, while an assistant captured and saved the image. To control for the influence of respiration, all images were captured at the end of normal exhalation. The order in which the muscles (TrA and LM) were imaged was counterbalanced.

**Transversus Abdominis** Images of the TrA muscle were acquired in the following order: (1) supine at rest, with lower extremities extended, (2) during an active straight-leg raise (ASLR) maneuver (ipsilateral lower extremity extended), (3) hook lying at rest, and (4) during an abdominal drawing-in maneuver (ADI) performed in hook lying. Ultrasound images of the TrA muscle were obtained following the techniques outlined by Teyhen et al, with the transducer positioned just superior to the iliac crest along the midaxillary line and the middle of the muscle belly centered within the field of view.

The ASLR maneuver (FIGURE 1) has been advocated as an assessment for lumbopelvic dysfunction and was used in this study to assess automatic (ie, nonvolitional) changes in muscle thickness of the TrA muscle. The lower extremity con-
tralateral to the imaged side was lifted, as it seemed to elicit the same or larger TrA response as the ipsilateral lower extremity during pilot work and was consistent with raising the contralateral arm when testing the LM. Participants were instructed to “raise your leg off of the table approximately 8 inches (20 centimeters), without bending your knee” and were given a single practice of the ASLR maneuver before image acquisition. The ADIM (FIGURE 2) is a motor control exercise for the TrA muscle and was utilized in this study to assess changes in muscle thickness associated with a volitional activation. Participants were instructed to “take a relaxed breath in and out, hold the breath out, and then draw-in your lower abdomen without moving your spine.”24,29 Alternate cues of “cut off the flow of urine” or “close your rear passage” were sometimes given in an attempt to maximize a preferential TrA contraction. The cue resulting in the largest preferential change in muscle thickness of the TrA during an ADIM was used, as the subject practiced the ADIM until a ceiling effect occurred in performance (approximately 5 times).

Lumbar Multifidus Images of the LM muscle were taken (1) during prone rest and (2) during a contralateral arm raise (CAL) maneuver (FIGURE 3), following techniques outlined by Kiesel et al.15 The transducer was placed longitudinally (parasagittal view), then angled slightly medially until the facet joints were visualized. After identifying the superior border of the sacrum, the transducer was moved superiorly until the L4-5 facet was centered on the screen. The CAL maneuver was performed prone with the elbows flexed 90°, shoulders abducted 120°, and holding a hand weight. The weight used depended on the participant’s body weight and has been shown to elicit approximately 30% of the maximal voluntary isometric contraction of the LM muscle.15 Participants weighing less than 68 kg used 0.68 kg, those between 68 and 90 kg used 0.91 kg, and those greater than 91 kg used 1.36 kg. Participants were instructed to “lift your arm approximately 2 inches (5 centimeters) off the table” and were given a single practice CAL trial before image acquisition.

Measurements All images were measured offline using National Institutes of Health (Bethesda, MD) Image J, Version 1.38t. TrA thickness measurements were made between the superficial and deep borders of the TrA muscle, as visualized by the hyperechoic fascial lines (FIGURE 4). LM thickness measurements were made between the posteriormost portion of the L4-5 facet joint and the plane between the superficial muscle and subcutaneous tissue (FIGURE 5). By using Image J’s automatic measurement function (Ctrl+M) and concealing the measurement output...
on the computer screen, the examiner was blinded during measurement to the current and previous thickness values.

**Data Analysis**

Statistical analysis was performed using the Statistical Package for the Social Sciences, Version 16.0 software (SPSS, Chicago, IL). The dependent measure of percent change in TrA and LM muscle thickness from rest was calculated as $(\text{thickness}_{\text{contracted}} - \text{thickness}_{\text{rest}})/\text{thickness}_{\text{rest}}$. Each sequential measurement was arranged in SPSS as a separate variable, and intraexaminer reliability with 95% confidence intervals (CIs) was estimated using 2-way mixed-model, consistency-type intraclass correlation coefficients (ICCs). The reliability when using a single measurement was estimated using the first 2 measurement variables and the “single measures” output from SPSS (model 3,1). The reliability when using a mean of 2 measurements was estimated using the first 2 measurement variables and the “average measures” output from SPSS (model 3,2). The reliability, when using a mean of 3, 4, 5, and 6 measurements, was similarly estimated using the first 3, 4, 5, and 6 measurement variables and the “average measures” output from SPSS (model 3,k).

Using the resulting ICC values and the equation $SD_{\text{pooled}} \times \sqrt{(1 - ICC)}$, SEM was calculated to assess measurement precision. The 95% CIs were estimated for SEM using sum-of-squares error divided by the chi-square value for probability level $\alpha = .025$ for the lower limit and $.975$ for the upper limit. We visually assessed for the presence of a systematic order effect (eg, learning and/or fatigue) by graphing the mean value of the first, second, third, fourth, fifth, and sixth measurements of percent change in TrA and LM muscle thickness. Repeated-measures analysis of variance (ANOVA) tests were then performed to assess for stability of sequential measurements using $\alpha = .05$ for each comparison.

**RESULTS**

Demographic and baseline characteristics of the patient sample are provided in Table 1. Images from 1 participant for the LM muscle were excluded because the examiner was unable to identify muscle boundaries, leaving 30 subjects with full TrA images and 29 with full set of LM images. Reliability coefficients and SEM (with 95% CIs), when using a single measurement and when using a mean of 2, 3, 4, 5, and 6 measurements, are presented in Table 2. SEM decreased between single measurements,
a mean of 2 measurements, and a mean of 3 measurements for both the TrA and LM muscles during all standardized tasks (Figure 6). Averaging more than 3 measurements resulted in little to no further decrease in SEM.

The mean (SD) change in TrA muscle thickness across all 6 trials during the ASLR was 17.1% (36.0%) and 82.5% (41.9%) during the ADIM. The mean (SD) change in LM muscle thickness across all 6 trials of the CAL was 10.5% (8.7%). No significant differences were found between mean changes in muscle thickness of each sequential measurement for any muscle or standardized task (P<.05). Furthermore, no obvious upward (learning) or downward (fatigue) “trends” were apparent from visual inspection of mean percent change in muscle thickness data during any standardized task (Figure 7).

**DISCUSSION**

The results from this analysis suggest that, when using percent change in muscle thickness as an indirect assessment of muscle activation, using the mean of multiple measures substantially improves intraexaminer measurement precision. Our results suggest that measurement precision is optimized by using an average of 3 consecutive measurements of the TrA and LM during standardized tasks. Regardless of muscle (TrA, LM) or standardized task (ADIM, ASLR, CAL), SEM decreased approximately 25% when using a mean of 2 measures, and by almost 50% when using the mean of 3 measures, when compared to a single measure. Averaging more than 3 measures did not result in consistent additional improvements in measurement precision.

When deciding how many measurements to average, practitioners need to weigh the benefits of increased precision with the costs of spending more time and resources to take each additional measurement. Unfortunately no studies to date have established minimal clinically important differences (MCIDs) of percent change in TrA or LM muscle thickness. However, the improvements found in SEM when averaging 2 measurements as compared to a single measurement, and 3 measurements as compared to 2 measurements, are both at least moderate in relation to mean percent thickness change values. Because the cost of taking each additional measurement is minimal (perhaps 1-2 minutes to capture and measure each additional image), using the mean of 3 measurements seems to provide the optimal cost-benefit ratio.

Although improvements in measurement precision cannot be compared to MCIDs, they can be compared to differences reported between symptomatic and asymptomatic individuals. The mean percent change in muscle thickness in this symptomatic sample was 17.1% for the TrA during the ASLR, 82.5% for the TrA during the ADIM, and 10.5% for the LM during the CAL. In asymptomatic individuals, Kiesel et al 16 reported a mean TrA percent change in muscle thickness of 99.1% during an ADIM and a mean LM percent change in L4-5 muscle thickness of 24.4% during a CAL. Comparison of the symptomatic and asymptomatic values between the current study and that of Kiesel et al 16 suggests that patients with LBP (compared to persons who are asymptomatic) demonstrate a diminished change in muscle thickness of the LM, at a level that is greater than measurement error (SEM), even when using only a single measure. In terms of percent change in the TrA muscle during the ADIM, the difference between persons who are symptomatic and those who are asymptomatic is greater than measurement er-
the need for identifying more reliable contraction strategies of the TrA, as well as methods to further reduce error during such measurements.

In the primary reliability analysis of this study, we focused on the reliability of thickness measures as opposed to percent change in muscle thickness measures. We decided to report our primary results using the mean of 2 measures, as there seemed to be very little additional benefit to averaging in a third measure. Single thickness measures may be important if used to assess within person changes over time, when one expects muscle atrophy or hypertrophy. Although some studies have found LM muscle atrophy in patients with LBP, most of the relevant literature focuses on functional deficits of the TrA and LM muscles. Because thickness measures likely relate more to a person’s size than their muscle function, percent change in muscle thickness is arguably a more clinically relevant measure than that of thickness. However, as expected, measurement of percent change is less reliable than single thickness measures, as it incorporates measurement error from both resting and contracted states. This is evident from our analysis and is consistent with the recommendation of using the mean of at least 3 measurements.

When using the mean of multiple measures to best estimate an individual’s true measure, it is important not to have a systematic order effect. Because muscle performance measures are naturally prone to be influenced by fatigue and/or learning effects, the second part of this analysis was aimed at identifying any systematic bias related to measurement order. Regardless of muscle (TrA, LM) or standardized task (ADIM, ASLR, CAL), we found that 6 consecutive measures, when taken within approximately 5 to 10 minutes, are relatively stable. The lack of an identifiable upward or downward trend in sequential means suggests that neither fatigue nor learning systematically affected percent change in

![Graph](image_url)

**FIGURE 7.** Mean percent change in transversus abdominis (TrA, left axis) and lumbar multifidus (LM, right axis) muscle thickness during an abdominal drawing-in maneuver (ADIM), an active straight-leg raise (ASLR), and a contralateral arm raise (CAL) during sequential measurements. Error bars are standard error of the mean.
TrA or LM muscle thickness during any standardized task.

In addition to the limitations discussed in our primary reliability report,17 this analysis is limited by the inclusion of only intraexaminer data obtained during the same day. While we believe that this analysis yields important information for RUSI practitioners and researchers, we cannot generalize the results to measurement error when images are taken during different days or by different examiners. Because making between-day and/or between-examiner comparisons add additional error, it is possible that such comparisons would benefit from more than 3 measurements. Moreover, quantitative interpretations of measurement error would ideally be made by comparing changes in SEM to established MCIDs. Because MCIDs have not been established for any RUSI measurements, we cannot comment on the importance of the reported improvements in measurement precision. Lastly, error during measurement can be due to multiple sources, including the instrument, the examiner (eg, transducer motion), and intra-subject variability. Identifying the amount of measurement error that comes from each of these different sources requires a much larger sample and a different analysis (generalizability theory) and cannot be inferred from this study.

Future studies should continue to investigate measurement error associated with RUSI, especially using the more clinically relevant measure of percent change in muscle thickness during muscle activation. Because absolute measurement error (SEM) is relatively large in comparison to mean TrA measures, studies should focus on identifying strategies to improve the reliability of both acquiring and measuring TrA images.

CONCLUSION

When using RUSI to determine percent change in TrA and LM muscle thickness, intraexaminer measurement precision is substantially improved by using the mean of multiple measures. Using an average of 3 measurements of the TrA and LM muscles appears to optimize measurement precision and is recommended.

KEY POINTS

FINDINGS: When using RUSI to determine percent change in TrA and LM muscle thickness, intraexaminer measurement precision was optimized by using an average of 3 consecutive measurements. Little precision was gained by averaging more than 3 measurements.

IMPLICATION: Using an average of 3 measurements when assessing percent thickness change of the TrA and LM muscles substantially improves measurement precision compared to fewer trials and is recommended for both research and clinical use.

CAUTION: This analysis is limited by the inclusion of only intraexaminer data obtained during the same day. Comparisons using measurements from different days and from different examiners include additional error and, therefore, may benefit from more than 3 measurements. Moreover, because MCIDs have not been established for RUSI measurements, this analysis cannot evaluate the “importance” of these improvements in measurement precision or allow us to identify specific sources of measurement error.

REFERENCES

19. Mens JM, Vleeming A, Snijders CJ, Koes BW, Stam HJ. Validity of the active straight leg raise test for measuring disease severity in patients...
with posterior pelvic pain after pregnancy.

EARN CEUs With JOSPT’s Read for Credit Program

JOSPT’s Read for Credit (RFC) program invites Journal readers to study and analyze selected JOSPT articles and successfully complete online quizzes about them for continuing education credit. To participate in the program:

1. Go to www.jospt.org and click on “Read for Credit” in the left-hand navigation column that runs throughout the site or on the link in the “Read for Credit” box in the right-hand column of the home page.
2. Choose an article to study and when ready, click “Take Exam” for that article.
3. Login and pay for the quiz by credit card.
4. Take the quiz.
5. Evaluate the RFC experience and receive a personalized certificate of continuing education credits.

The RFC program offers you 2 opportunities to pass the quiz. You may review all of your answers—including the questions you missed. You receive 0.2 CEUs, or 2 contact hours, for each quiz passed. The Journal website maintains a history of the quizzes you have taken and the credits and certificates you have been awarded in the “My CEUs” section of your “My JOSPT” account.