DAVID WALTON, PT, PhD[1] • **JOY MACDERMID,** PT, PhD[2] • **WARREN NIELSON,** PhD[3]
ROBERT TEASELL, MD[4] • **MARCO CHIASSON**[5] • **LAUREN BROWN**[5]

Reliability, Standard Error, and Minimum Detectable Change of Clinical Pressure Pain Threshold Testing in People With and Without Acute Neck Pain

STUDY DESIGN: Clinical measurement.

OBJECTIVES: To evaluate the intrarater, interrater, and test-retest reliability of an accessible digital algometer, and to determine the minimum detectable change in normal healthy individuals and a clinical population with neck pain.

BACKGROUND: Pressure pain threshold testing may be a valuable assessment and prognostic indicator for people with neck pain. To date, most of this research has been completed using algometers that are too resource intensive for routine clinical use.

METHODS: Novice raters (physiotherapy students or clinical physiotherapists) were trained to perform algometry testing over 2 clinically relevant sites: the angle of the upper trapezius and the belly of the tibialis anterior. A convenience sample of normal healthy individuals and a clinical sample of people with neck pain were tested by 2 different raters (all participants) and on 2 different days (healthy participants only). Intraclass correlation coefficient (ICC), standard error of measurement, and minimum detectable change were calculated.

RESULTS: A total of 60 healthy volunteers and 40 people with neck pain were recruited. Intrarater reliability was almost perfect (ICC = 0.94-0.97), interrater reliability was substantial to near perfect (ICC = 0.79-0.90), and test-retest reliability was substantial (ICC = 0.76-0.79). Smaller change was detectable in the trapezius compared to the tibialis anterior.

CONCLUSIONS: This study provides evidence that novice raters can perform digital algometry with adequate reliability for research and clinical use in people with and without neck pain. *J Orthop Sports Phys Ther 2011;41(9):644-650. doi:10.2519/jospt.2011.3666*

KEY WORDS: *algometer, cervical spine, PPT, tibialis anterior*

Measurement of pain is an important component of clinical practice. Its importance is evident in the frequency with which it drives healthcare utilization and its impact on quality of life.[1,5] Pain measurement can be useful as an outcome measure to be evaluated over time, or a prognostic measure that can predict future outcomes.[22] In addition, pain is a primary outcome measure in most musculoskeletal intervention studies. The most common approach to pain measurement is through patient self-report, using visual analog or numeric pain rating scales. While self-reported pain intensity is important, it is a composite of physiological and psychological features of the person and the health problem, mediated by social aspects[10,19] that can render responses difficult to interpret. There is value in having pain measures that are not self-reported, as they reflect different aspects of the health condition.

Recently, thermal and pressure pain threshold (PPT) have demonstrated predictive ability as useful prognostic indicators for individuals with acute whiplash-associated disorder (WAD). Specifically, Sterling and colleagues[15,16] have found that low PPT or elevated cold pain threshold, both of which indicate nociceptive sensitivity, are predictive of poor short- and long-term outcomes in a subset of people with acute WAD. Interestingly, the predictive relationship exists

[1]Assistant Professor, School of Physical Therapy, The University of Western Ontario, London, Ontario, Canada. [2]Adjunct Faculty, School of Physical Therapy, The University of Western Ontario, London, Ontario, Canada; Professor, School of Physical Therapy, McMaster University, Hamilton, Ontario, Canada; Director, Clinical Research Lab, Hand and Upper Limb Centre, St Joseph's Hospital, London Ontario, Canada. [3]Clinical Director, Rheumatology Day Programs, St Joseph's Hospital, London Ontario, Canada. [4]Professor, Schulich School of Medicine, Department of Physical Medicine and Rehabilitation, University of Western Ontario and Parkwood Hospital, St Joseph's Health Care, London, Ontario, Canada. [5]Student (at time of study), School of Physical Therapy, The University of Western Ontario, London, Ontario, Canada. This study was approved by the Health Sciences Research Ethics Board at the University of Western Ontario and funded in part by the Physiotherapy Foundation of Canada and the Ontario Rehabilitation Research Advisory Network. The primary and secondary authors are funded through the Canadian Institutes of Health Research. The authors have received no funding or compensation from Wagner Instruments beyond a small discount on the retail price of each unit for bulk purchasing. The authors are not affiliated in any way with Wagner Instruments, and disclose no conflict of interest. Address correspondence to Dr David Walton, School of Physical Therapy, Room EC1443, The University of Western Ontario, 1201 Western Road, London, ON N6H 5E5, Canada. E-mail: Dwalton5@uwo.ca

even when the stimulus is applied at sites distal to the neck, such as the tibialis anterior. While both appear to have some predictive value, cold pain threshold may be the better predictor, as shown by the exclusion of PPT in the final regression model. However, PPT is currently the more accessible of the 2 and more likely to be implemented in routine clinical practice.

The identification of low pain threshold at anatomically distant sites may be indicative of generalized nociceptive hypersensitivity.[16] Because pain is the chief presenting complaint in WAD and most other musculoskeletal conditions, a simple clinical tool that provides reliable, quantitative assessment of pain threshold at different sites has potential to augment practice for a broad spectrum of patients, as long as scores can be interpreted against normative values and the significance of change scores is known. To date, however, the bulk of research on pain algometry in whiplash has been performed in lab-based settings using high-precision algometers,[17] which, though scientifically sound, are cost prohibitive for routine clinical use. Others have shown PPT testing to be reliable in otherwise healthy samples,[3,12] with questionable generalizability to clinical samples.

An important step toward integrating PPT testing into routine clinical practice is to establish of the reliability of tools that could be viable for clinical practice. Such tools must be commercially available, meet measurement standards under healthy, ideal conditions, and, when used in a clinical setting, not be cost prohibitive. The clinimetrics of PPT testing with devices available in the clinical setting is largely unknown for people with acute neck pain.

The objective of this study was to establish the interrater, intrarater, and test-retest reliability of a clinically available digital algometer in normal healthy individuals and in people with neck pain, and to determine the minimum detectable change (MDC) and standard error of measurement (SEM) of testing to facilitate clinical interpretation of change over time.

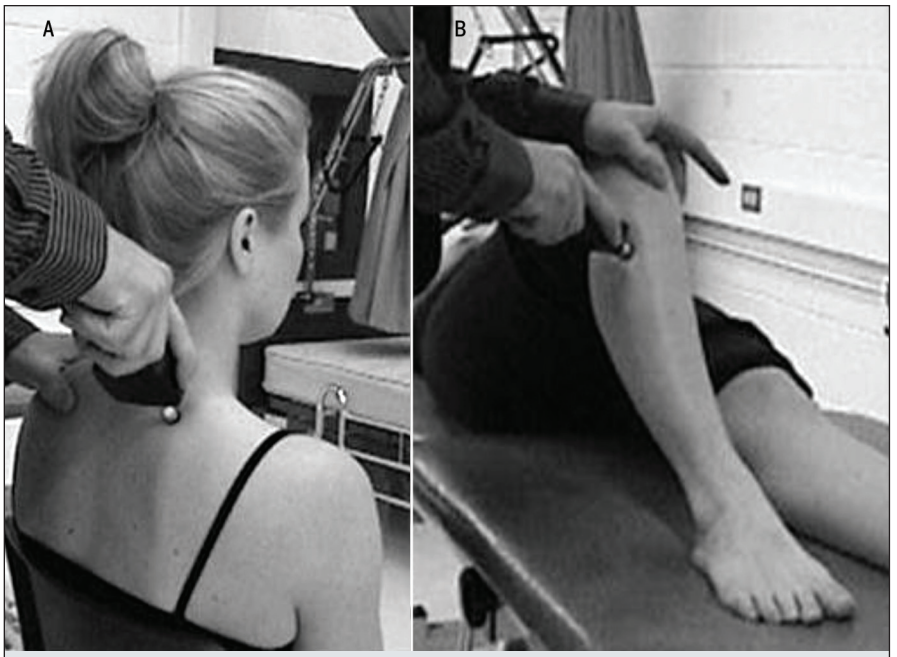

FIGURE 1. Locations for testing of the upper fibers of the trapezius (A) and tibialis anterior (B) sites.

METHODS

Pressure Pain Threshold

The following protocol was used to collect PPT for both phases of this study. In preparation for testing, raters practiced applying force until they were able to do so consistently at a rate of approximately 50 kPa/s, which corresponds to 5 N/s. This was accomplished by applying increasing pressure on a firm surface, while a colleague counted out 5 seconds. Consistent ability to achieve a rating of 250 ± 25 kPa after 5 seconds was evidence that the rate of application was satisfactory. When evaluating PPT of the participants, the following standardized instructions were used: "I'm going to begin applying pressure to your muscle. I want you to tell me the moment the sensation changes from comfortable pressure to slightly unpleasant pain." The descriptor "unpleasant" was chosen as it agrees with the definition of pain established by the International Association for the Study of Pain,[11] and Gracely[6] has suggested that "slightly unpleasant" represents an approximate intensity of 5 on a 21-point numeric pain rating scale, satisfying our intention to capture pain threshold rather than pain tolerance.

Device The Wagner FDX-25 (Wagner Instruments, Greenwich, CT) is a handheld digital algometer with a linear response to force application between 0 and 1300 kPa (approximately 0-14 kg). It has a 1-cm^2 round rubber tip, and values are displayed as the maximum force applied before the individual verbally states that pain threshold has been reached. The algometers were calibrated using a known-weight technique prior to commencing the study (**FIGURE 1**).

Testing Location and Position The 2 sites tested for all participants were the angle of the upper fibers of the trapezius (UFT) muscle, approximately 5 to 8 cm superomedial to the superior angle of the scapula, and the muscle belly of the tibialis anterior (TA), approximately 2.5 cm lateral and 5 cm inferior to the tibial tubercle (**FIGURE 1**). Palpation during resisted active dorsiflexion ensured that the TA muscle belly had been found. We selected the UFT point because many people, following traumatic neck injury,

report that their symptoms are in the area of the base of the neck and because this is an easily identifiable landmark for testing local pain threshold. The TA site selection was based on its previous use by other researchers as a distal comparison site for people with neck pain.[13,16] For testing of the UFT participants were seated, and for testing of the TA participants were in supine hook-lying on an examination table, with the knees flexed to 90° and feet flat. The skin over the area to be tested was exposed. Three measurements of pain threshold were recorded for each site and on each side of the body, starting at the TA, then moving to the UFT. An interval of approximately 1 minute separated each measurement. For consistency, the right side was tested first, followed by the left, in all instances. All participants gave informed consent prior to testing, and all elements of the study were approved by The Health Sciences Research Ethics Board at the University of Western Ontario, prior to commencing data collection.

Reliability

Recruitment and Environment A convenience sample of healthy, noninjured participants were recruited from the student body and faculty at the University of Western Ontario. Participants were included if they were between the ages of 18 and 55 years, were able to read and understand English, reported no injuries to the neck or lower leg within the past year, did not visit a healthcare provider for issues of neck or lower leg pain within the past 6 months, and reported no current neck or lower leg pain. All participants provided informed, written consent prior to participation.

For PPT testing, participants were visually isolated from each other by screens or curtains. Temperature, noise, and ambient lighting were kept as constant as possible between testing sessions. Three physical therapy students acted as raters. All 3 raters were male, and were instructed to act in a friendly manner but to limit conversation with the participant as much as possible. Rater order was randomized using a coin flip.

Patients with neck pain were recruited from 8 physiotherapy clinics across Canada. Participants were included if they presented for physiotherapy rehabilitation within 12 weeks of the onset of their neck pain, were able to speak French or English, and were 18 years of age or older. Rater pairs at each clinic were composed of 1 therapist who was currently enrolled in an advanced manual and manipulative therapy training program, and 1 physiotherapist colleague. All raters involved in this study (students and clinicians) underwent a 1-hour training session prior to initiating data collection. None of the raters had previous experience performing PPT measurements.

Testing Healthy participants were tested on 2 different occasions, separated by 3 to 5 days. On the first occasion, participants completed a general survey of demographics (age, sex), neck and/or leg pain history, caffeine intake within the past 6 hours or vigorous exercise within the past 24 hours, and 2 numeric rating scales developed specifically for this study—sleep quality the previous night (0 to 10, with 0 as terrible and 10 as very good) and state of anxiety (0 to 10, with 0 as not at all and 10 as extremely anxious). Each of these items was captured to evaluate its role as a potential moderator of PPT.

Participants with neck pain were presented with an intake survey that inquired about the nature of the disorder causing their neck pain, their age and sex, whether or not they were currently involved in litigation, and their current neck pain intensity using a standard numeric rating scale (0 to 10, with 0 as no pain and 10 as terrible pain). In both samples, raters were blinded to the responses in the survey.

Following the first rater's assessment of PPT described above, participants were allowed to rest comfortably for 5 minutes prior to the second rater entering the room and repeating the same protocol. The entire protocol took approximately 25 minutes per participant. Healthy participants returned 3 to 5 days later and again completed the brief intake survey. Then, the same PPT protocol performed on the first occasion was completed, but this time by only 1 rater.

Statistical Analysis Estimates of reliability were calculated using relative ($ICC_{2,1}$ for absolute agreement[14]) and absolute (SEM, MDC) estimates. SEM, as an indication of expected measurement error in a single individual's score using the same units as the algometer (kPa), was calculated using the following formula: $SD_{pooled} \times \sqrt{1-ICC}$.[7] MDC was calculated

TABLE 1 Sample Characteristics*

Variable	Uninjured (n = 60)	Symptomatic (n = 40)
Sex, percent female	65%	66%
Age, y	25.4 (22-55)	40.4 (20-68)
Duration of symptoms, d	...	38 (7-90)
Pain intensity (0-10)	...	4 (0-8)
Cause of injury		
Traffic accident	...	33%
Other trauma	...	4%
Repetitive strain/postural	...	33%
Unknown	...	21%
Other	...	8%
Active legal/compensation case	...	25%

Values are mean (range) unless otherwise indicated.

TABLE 2 Reliability Indicators of Pressure Pain Threshold Testing for Intrarater, Interrater, and Test-Retest Reliability at Each Site

Group/Measure	$ICC_{2,1}$ (95% CI)	SEM, kPa*	MDC, kPa†
Uninjured			
Intrarater			
UFT	0.97 (0.94, 0.98)	18.2	42.7
TA	0.94 (0.91, 0.97)	37.4	86.3
Interrater			
UFT	0.79 (0.66, 0.87)	52.5	121.9
TA	0.84 (0.75, 0.90)	59.2	137.0
Test-retest			
UFT	0.76 (0.61, 0.85)	48.9	113.4
TA	0.79 (0.66, 0.87)	66.7	154.4
Symptomatic			
Intrarater			
UFT	0.96 (0.91, 0.98)	20.5	47.2
TA	0.97 (0.94, 0.99)	42.3	97.9
Interrater			
UFT	0.81 (0.67, 0.90)	50.3	117.0
TA	0.90 (0.81, 0.94)	73.8	171.3

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; MDC, minimum detectable change; SEM, standard error of measurement; TA, tibialis anterior; UFT, upper fibers of trapezius.
**Using pooled SD.*
†At the 90% confidence level.

at the 90% level, which is appropriate for assessing change for routine clinical use.[4] The formula $MDC_{90} = SEM \times \sqrt{2} \times 1.64$ [18] was used to provide the threshold amount of change in scores required for the rater to be 90% confident that true change beyond that of measurement error had occurred. This calculation has been used for situations in which the same rater assessed the same person within a single session (intrarater), 2 independent raters assessed the same person within a single session (interrater), and when the same rater assessed the same person across 2 independent sessions (test-retest).

The data used for calculation of intrarater reliability were the second and third PPT measurements made at each site, on each participant, by the first rater. Interrater reliability was estimated using the mean of the 3 values obtained from each participant at each site, averaged across the 2 sides (6 total measurements per site). Test-retest reliability was estimated using the mean of the 3 measures at each site, by the same rater across the 2 testing occasions. Reliability coefficients were interpreted with the subjective categories of Landis and Koch,[9] in which values of less than 0.40 were considered unacceptable, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, and 0.81 to 1.00 almost perfect agreement. Bland-Altman plots[2] were constructed for each condition and inspected visually for consistency of agreement across all levels of the scale. All analyses were done using SPSS Version 17 (Chicago, IL), and the *P* value for significance was set at .05.

RESULTS

Reliability in Healthy Participants

Sixty healthy individuals volunteered for the study. Females with a mean age of 25 years comprised 65% of the sample (**TABLE 1**). The mean value for PPT at the UFT site was 251.8 kPa (2.57 kgf), with a standard deviation of 102.3 kPa and a 95% confidence interval (CI) of 225.5 to 278.0 kPa. The mean value at the TA site was 334.1 kPa (3.41 kgf), with a standard deviation of 157.0 kPa and a 95% CI of 293.6 to 374.5 kPa.

Intrarater reliability was excellent at both sites (0.94-0.97), and interrater and test-retest reliability were substantial (0.76-0.84) (**TABLE 2**). The SEM and MDC were smaller for the UFT as compared to the TA. **FIGURE 2** presents the Bland-Altman plots for each of the reliability conditions. Visual inspection revealed that all mean differences were close to zero, and that the difference between raters tended to increase with higher levels of the scale, indicating low but proportional measurement error across the spectrum of the scale.

Reliability in Participants With Neck Pain

Patients (n = 40) were predominantly female (66%), and had a mean symptom duration of 38 days (**TABLE 1**). The modal mechanisms of injury were traffic accident (33%) and postural/repetitive strain (33%); 25% of the sample were receiving compensation or were actively involved in litigation. The mean PPT at the UFT site was 238.9 kPa (2.44 kgf), with a standard deviation of 121.4 kPa and a 95% CI of 201.1 to 276.2 kPa. The mean value at the TA site was 401.7 kPa (4.10 kgf), with a standard deviation of 240.6 kPa and 95% CI of 330.5 to 472.8 kPa.

Intrarater reliability was excellent at both sites: 0.96 at the UFT and 0.97 at the TA (**TABLE 2**). Interrater reliability for the UFT site was substantial (0.81) and was excellent for the TA site (0.90). **TABLE 2** also presents SEM and MDC. As was seen in the healthy participants, the changes detectable at the UFT site were smaller than those at the TA site.

FIGURE 3 presents the Bland-Altman plots for the UFT and TA sites. Visual inspection of the plots for intrarater reliability suggested overall small measurement error, with proportional increases

at the higher ends of the scale.

DISCUSSION

Overall, this study supports the reliability of testing PPT in asymptomatic and symptomatic individuals at the upper trapezius and tibialis anterior using a handheld digital algometer. The innovative work of Sterling and colleagues[16,17] has provided motivation for clinical investigation of quantitative sensory testing in individuals with acute WAD, and our study moves some of their work further into the clinical arena. This group has demonstrated both reliability[17] and predictive validity[16] of PPT testing over both the cervical spine and tibialis anterior sites under controlled experimental conditions. The results of this study and 2 related manuscripts[20,21] suggest that similar measurement properties and clinical phenomena are observable when testing is performed in circumstances more reflective of routine clinical practice.

The results of our study suggest that PPT testing can be undertaken reliably and is acceptable regardless of whether the individual tested is currently experiencing symptomatic neck pain or not. Both novice and experienced therapists were able to obtain reliable results after approximately 1 hour of training. Within-session repeatability of a single PPT measure taken by the same rater was excellent in both injured and noninjured participants. Between-rater reliability was substantial but lower, as is common in many clinical measurement situations due to increased variability in techniques between practitioners. Test-retest reliability was also substantial in the uninjured sample. Absolute error, reflected in the SEM and MDC, suggests that the clinical measurement properties for PPT evaluated in this study were appropriate. Clinicians will be able to use these values in their communications with funders or other members of the healthcare team, indicating their level of confidence in the extent and meaning of change scores in their patients. However, clinicians should note that the MDC values at the UFT site, which ranged between approximately 44.5 and 111.2 kPa (0.45-1.13 kgf), indicate that the tool may not be accurate enough to detect a decrease in PPT if the baseline value starts below the MDC, as it did in some of the participants in both the healthy and clinical samples. This particular concern is tempered by the fact that PPT would be expected to increase rather than decrease following intervention; but the cautionary note remains prudent, as the tool may not accurately detect worsening in a patient who starts with a very low PPT.

Our results compare favorably to those of other investigators. Persson and colleagues[12] found intrarater ICCs in the range of 0.70 to 0.90 for testing the upper trapezius muscles of 27 healthy females. Chesterton and colleagues[3] trained undergraduate physiotherapists to perform PPT testing over the first dorsal interos-

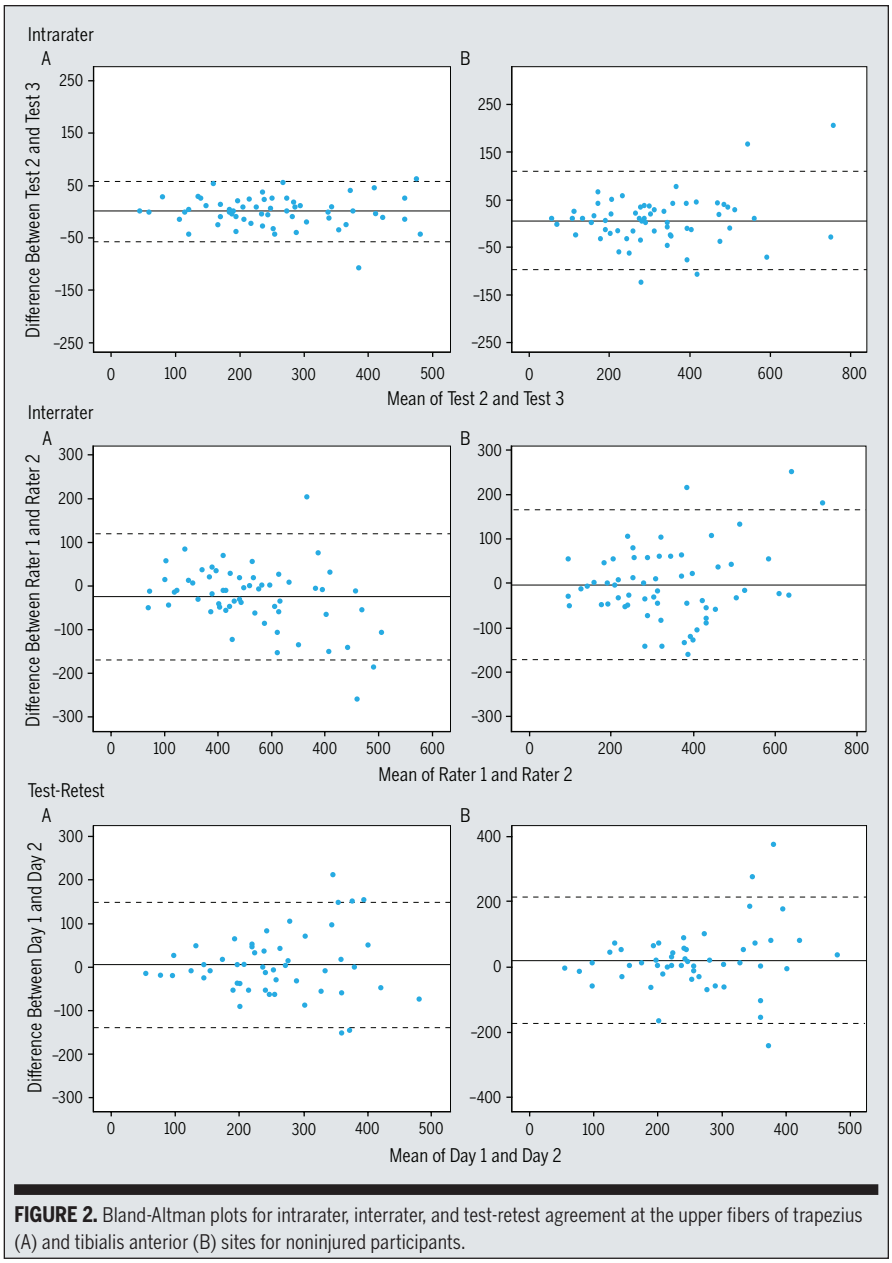

FIGURE 2. Bland-Altman plots for intrarater, interrater, and test-retest agreement at the upper fibers of trapezius (A) and tibialis anterior (B) sites for noninjured participants.

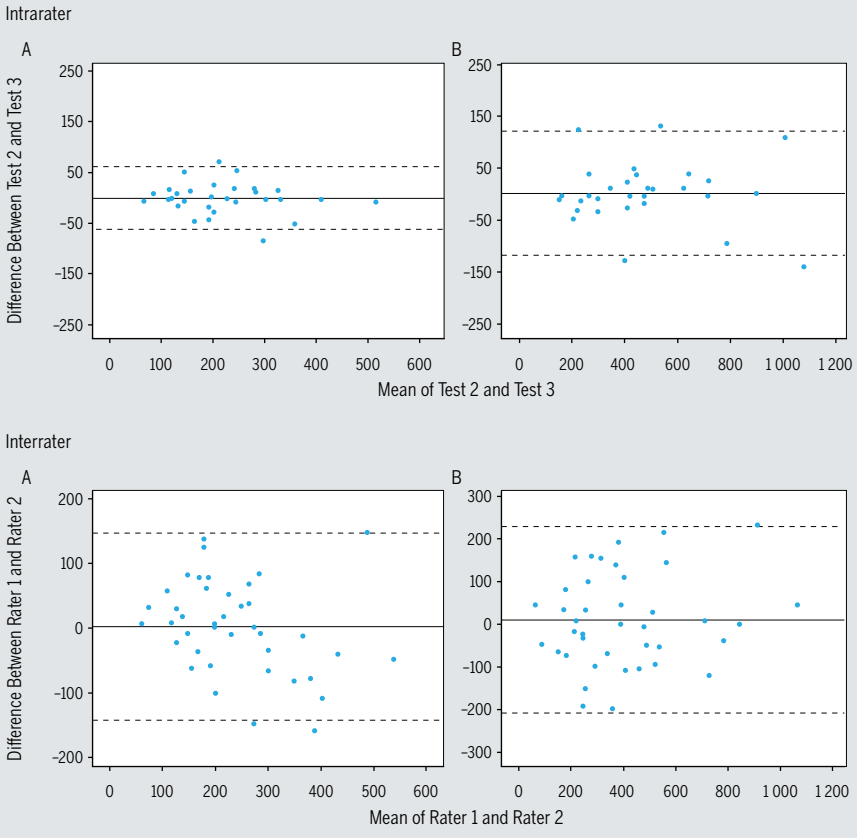

FIGURE 3. Bland-Altman plots for intrarater and interrater reliability at the upper fibers of trapezius (A) and tibialis anterior (B) sites for symptomatic participants.

seous muscle. Their interrater $ICC_{2,1}$ of 0.91 and our ICC for the tibialis anterior are comparable, and both are higher than that found for the UFT. This suggests that the neck muscles may be a more labile site with respect to pain threshold than the muscles of the extremities.

The inclusion of values from healthy noninjured participants has demonstrated that both relative and absolute reliability are not adversely affected by the presence or absence of neck pain. Researchers should find these normal values useful when designing the lab-based pain stimulation studies that are routinely conducted using healthy uninjured participants. The fact that physiotherapy students were able to achieve reliable results with a handheld digital algometer suggests that previous clinical experience is not a prerequisite for adequate measurement properties. It is notable that the mean PPT values for the UFT site (251.8 kPa in healthy participants, 238.9 kPa in those with neck pain) were not noticeably different; in fact, the mean values at the TA might have been even higher in the clinical sample than in the healthy sample (334.1 kPa in the healthy sample, 401.7 kPa in those with neck pain). No group-to-group comparison was planned or conducted, and the somewhat counterintuitive results might simply be indicative of nonmatched groups or an influence of the measurement environment. The design of the study was such that direct comparison would be inappropriate and any further interpretation of these findings would be pure speculation.

We believe that important questions remain to be answered. For example, while intrarater and interrater reliability were good to excellent in people with neck pain, and similar to those of healthy participants, separate-days test-retest reliability in people with neck pain is still unknown. It is tempting to assume that, based on the similarities in findings for the other 2 types of reliability estimates, values would be similar to the test-retest estimates found for the healthy participants. However, this cannot be assumed and should be formally evaluated in a follow-up study. Visual inspection of the Bland-Altman plots (**FIGURES 2** and **3**) reveals that mean differences in PPT values tend to increase at higher levels of the scale, suggesting that SEM and MDC may be greater for those who have higher, as compared to lower, PPT values. The present study's sample size was not sufficiently large to further explore these differences across levels of the scale. This would be an important issue to clarify in future research, especially if PPT were to be recommended for use as an evaluative measure over time.

An important clarification is required when interpreting the results of this study. The protocol for intrarater reliability evaluation, comparing single values taken consecutively, was distinct from that of interrater and test-retest reliability, both of which used the mean of 6 values (3 values on each of 2 sides). Given that measurement precision tends to increase with repeated measurement,[8] we would expect the reliability estimates for intrarater reliability to be even higher had the mean of 3 values over 2 separate tests during the same session been used. Given the already very high estimates for intrarater reliability in both groups using a single value, it would appear that the added burden of 3 repeated measures to evaluate change during a single session would be unnecessary. However, when the desire is to evaluate change across sessions, as is more likely the case in clinical practice, the mean of at least 3 values is currently recommended.

The lack of rigorous scientific control in the present study was intentional, as the study's purpose was to evaluate the performance of digital algometry under real-world conditions, in which multiple

raters collect data on multiple patients in a variety of settings. Presumably, the properties found here would be as good, if not better, under tightly controlled conditions. While intentional, however, the lack of control does influence interpretation of the results. The number of raters introduces greater random error. Given the number of raters, the sample size was not large enough to allow investigation of potential systematic biases by rater. It is unknown whether raters with less training would achieve the same results.

CONCLUSION

We have been able to adequately reproduce the reliability of PPT measures in previous studies using a more clinically accessible device. Values for MDC in both injured and noninjured groups have been presented, which may facilitate the interpretation of clinical findings. Limitations in interpretation of the results have been outlined. The accessible digital algometer used in this study appears to be adequately reliable for use as a both a clinical and research tool. ◉

KEY POINTS

FINDINGS: PPT can be tested at the upper trapezius and belly of the tibialis anterior muscles, using an accessible digital algometer, with adequate intrarater, interrater, and test-retest reliability for research and clinical use.

IMPLICATION: Clinicians may use the information contained in this study to determine whether the change they see between testing sessions or between raters is reflective of true change rather than random error.

CAUTION: Multiple raters and multiple testing sites introduce the risk of bias due to rater. The expected impact of this is underestimation of the true reliability of the device.

ACKNOWLEDGEMENTS: *The authors would like to thank Matthew White, PT, Asterios Gougoulias, PT, and Jordan Lamothe, PT for their help in data collection for this study.*

REFERENCES

1. Adekoya N. Reasons for visits to emergency departments for Medicaid and State Children's Health Insurance Program patients: United States, 2004. *N C Med J*. 2004;71:123-130.
2. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135-160.
3. Chesterton LS, Sim J, Wright CC, Foster NE. Interrater reliability of algometry in measuring pressure pain thresholds in healthy humans, using multiple raters. *Clin J Pain*. 2007;23:760-766. http://dx.doi.org/10.1097/AJP.0b013e318154b6ae
4. Donoghue D, Stokes EK. How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *J Rehabil Med*. 2009;41:343-346. http://dx.doi.org/10.2340/16501977-0337
5. Fredheim OM, Kaasa S, Fayers P, Saltnes T, Jordhoy M, Borchgrevink PC. Chronic non-malignant pain patients report as poor health-related quality of life as palliative cancer patients. *Acta Anaesthesiol Scand*. 2008;52:143-148. http://dx.doi.org/10.1111/j.1399-6576.2007.01524.x
6. Gracely RH, Kwilosz DM. The Descriptor Differential Scale: applying psychophysical principles to clinical pain assessment. *Pain*. 1988;35:279-288.
7. Harvill L. Standard error of measurement. *Educ Meas*. 1991;10:33-41.
8. Koppenhaver SL, Parent EC, Teyhen DS, Hebert JJ, Fritz JM. 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. *J Orthop Sports Phys Ther*. 2009;39:604-611. http://dx.doi.org/10.2519/jospt.2009.3088
9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
10. Melzack R. Pain and the neuromatrix in the brain. *J Dent Educ*. 2001;65:1378-1382.
11. Merskey H, Bogaduk N. *Classification of Chronic Pain: Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms*. Seattle, WA: IASP Press; 1994.
12. Persson AL, Brogardh C, Sjolund BH. Tender or not tender: test-retest repeatability of pressure pain thresholds in the trapezius and deltoid muscles of healthy women. *J Rehabil Med*. 2004;36:17-27.
13. Scott D, Jull G, Sterling M. Widespread sensory hypersensitivity is a feature of chronic whiplash-associated disorder but not chronic idiopathic neck pain. *Clin J Pain*. 2005;21:175-181.
14. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
15. Sterling M, Jull G, Kenardy J. Physical and psychological factors maintain long-term predictive capacity post-whiplash injury. *Pain*. 2006;122:102-108. http://dx.doi.org/10.1016/j.pain.2006.01.014
16. Sterling M, Jull G, Vicenzino B, Kenardy J. Sensory hypersensitivity occurs soon after whiplash injury and is associated with poor recovery. *Pain*. 2003;104:509-517.
17. Sterling M, Treleaven J, Edwards S, Jull G. Pressure pain thresholds in chronic whiplash associated disorder: further evidence of altered central pain processing. *J Musculoskel Pain*. 2002;10:69-81. http://dx.doi.org/10.1300/J094v10n03_05
18. Stratford PW, Binkley J, Solomon P, Finch E, Gill C, Moreland J. Defining the minimum level of detectable change for the Roland-Morris questionnaire. *Phys Ther*. 1996;76:359-365; discussion 366-358.
19. Sullivan M. Toward a biopsychomotor conceptualization of pain: implications for research and intervention. *Clin J Pain*. 2008;24:281-290.
20. Walton D, MacDermid J, Teasell R, Nielson W, Nailer T, Maheu P. A descriptive study of pressure pain threshold at 2 standardized sites in people with acute or subacute neck pain. *J Orthop Sports Phys Ther*. 2011;41:651-657. http://dx.doi.org/10.2519/jospt.2011.3667
21. Walton D, MacDermid J, Teasell R, Nielson W, Reese H, Levesque L. Pressure pain threshold testing demonstrates predictive ability in people with acute whiplash. *J Orthop Sports Phys Ther*. 2011;41:658-665. http://dx.doi.org/10.2519/jospt.2011.3668
22. Walton D, Pretty J, Teasell R, MacDermid J. Risk factors for persistent problems following whiplash injury: results of a systematic review and meta-analysis. *J Orthop Sports Phys Ther*. 2009;39:334-350. http://dx.doi.org/10.2519/jospt.2009.2765