Clinical Pressure Pain Threshold Testing in Neck Pain: Comparing Protocols, Responsiveness, and Association With Psychological Variables

David M. Walton, Lenerdene Levesque, Martin Payne, Julie Schick

Background. Quantitative sensory testing, including pressure pain threshold (PPT), is seeing increased use in clinical practice. In order to facilitate clinical utility, knowledge of the properties of the tool and interpretation of results are required.

Objectives. This observational study used a clinical sample of people with mechanical neck pain to determine: (1) the influence of number of testing repetitions on measurement properties, (2) reliability and minimum clinically important difference, and (3) associations between PPT and key psychological constructs.

Design. This study was observational with both cross-sectional and prospective elements.

Methods. Experienced clinicians measured PPT in patients with mechanical neck pain following a standardized protocol. Subcohorts also provided repeated measures and completed scales of key psychological constructs.

Results. The total sample was 206 participants, but not all participants provided data for all analyses. Interrater and 1-week test-retest reliability were excellent (intraclass correlation coefficients [2,1]=.75–.95). Potentially important differences in reliability and PPT scores were found when using only 1 or 2 repeated measures compared with all 3. The PPT over a distal location (tibialis anterior muscle) was not adequately responsive in this sample, but the local site (upper trapezius muscle) was responsive and may be useful as part of a protocol to evaluate clinical change. Sensitivity values (range=0.08–0.50) and specificity values (range=0.82–0.97) for a range of change scores are presented. Depression, catastrophizing, and kinesiophobia were able to explain small but statistically significant variance in local PPT (3.9%–5.9%), but only catastrophizing and kinesiophobia explained significant variance in the distal PPT (3.6% and 2.9%, respectively).

Limitations. Limitations of the study include multiple raters, unknown recruitment rates, and unknown measurement properties at sites other than those tested here.

Conclusions. The results suggest that PPT is adequately reliable and that 3 measurements should be taken to maximize measurement properties. The variance explained by the psychological variables was small but significant for 3 constructs related to catastrophizing, depression, and fear of movement. Clinical implications for application and interpretation of PPT are discussed.

D.M. Walton, PT, PhD, Manipulative Therapy Program, School of Physical Therapy, Western University, 1201 Western Rd, Room EC1443, London, Ontario, Canada N6G 1A1. Address all correspondence to Dr Walton at: dwalton5@uwo.ca.

L. Levesque, PT, MClSc, Manipulative Therapy Program, School of Physical Therapy, Western University.

M. Payne, PT, MClSc, Manipulative Therapy Program, School of Physical Therapy, Western University.

J. Schick, PT, MClSc, Manipulative Therapy Program, School of Physical Therapy, Western University.

Post a Rapid Response to this article at:
ptjournal.apta.org

New knowledge in the field of mechanical neck pain and disability has necessitated an evolution of assessment techniques for the condition. This evolution has been driven by consistent evidence that the clinical presentation of neck pain, including the intensity of symptoms and the magnitude of movement restriction, are only weakly if at all associated with findings of structural pathology using conventional diagnostic imaging techniques (radiography, MRI).[1] Furthermore, recent findings indicate that some types of neck pain, most notably traumatic (ie, whiplash-related) neck pain, exhibit abnormal local and anatomically distinct mechanical and thermal pain thresholds[2,3] that may have prognostic utility, especially when sensitivity is identified distally,[4,5] and that challenge conventional understandings of the etiology of neck pain. Lastly, psychology has provided consistent evidence that the experience of neck-related pain and disability is influenced by negative emotional valence, either through diagnosable emotional dysfunction (eg, anxiety, depression, posttraumatic stress) or through more cognitively oriented constructs such as fear and catastrophizing.[6,7] It has become clear that assessment of the person with neck pain must be conducted through a diverse biopsychosocial lens.

A recent survey[8] suggests that although rehabilitation professionals appear to be aware of the biopsychosocial influences on neck pain and disability, relatively few are actively capturing multidomain variables in routine practice. In order to facilitate adoption of multisystem assessment, tools need to be clinically accessible, valid, and reliable and provide rich data with minimum burden. One such tool that has received increasing attention due to cost, safety, and ease of use is assessment of pressure pain threshold (PPT) through algom-

etry.[9] The PPT represents a hybrid test, falling somewhere between self-reported paper-and-pencil type tools and objective diagnostic techniques. It provides a quantitative value on a linear scale but is influenced by operator performance, patient cognitions, and operator-patient interaction. Despite these influences, standardized protocols exist that can provide potentially useful information.[10,11] Previous research indicates that even novice raters can reliably collect measurements of PPT in people with neck pain[12] and that PPT may have prognostic utility, especially in the case of widespread hyperalgesia.[4,13]

Previous PPT research intended for clinical application has used the mean of 3 measurements taken at each site of interest, with measurements usually separated by 1 minute.[3,14] If 3 measurements are taken bilaterally at each of 2 sites (12 total measurements), the entire protocol takes 12 minutes or longer, arguably too long for routine clinical use, which begs the question of whether the properties of PPT differ if fewer measurements are taken at each site. If so, clinical feasibility could be improved. Furthermore, properties such as test-retest reliability in a clinical sample and the minimum clinically important difference (MCID) provide important guidance for interpreting change scores but are currently unavailable for use in clinical samples with neck pain. Finally, a state of general uncertainty appears to exist regarding the mechanisms that explain reduced mechanical pain threshold, that is, the extent to which reduced PPT can be explained by abnormal biological processes or by a more general negative (psychological) orientation toward pain. The purposes of this study were: (1) to compare PPT values when using different clinical measurement protocols at the local (neck) and distal (lower leg) sites,

(2) to determine test-retest reliability and responsiveness of PPT in a subset clinical sample with mechanical neck pain, and (3) to determine the extent to which key psychological constructs such as fear, catastrophizing, depression, and anxiety are associated with PPT values in subsets of people seeking rehabilitation for mechanical neck pain.

Method
This study was a secondary analysis of 3 existing databases. Participants were recruited between May 2009 and December 2012 through 1 of 12 different outpatient orthopedic physical therapy clinics across Canada. Inclusion criteria were broad, including pain in the region of the neck, as evaluated through use of a body diagram, that was not the result of major systemic disease. In one cohort, only people with acute traumatic etiology (eg, whiplash) were targeted, whereas in the other 2 cohorts, no restrictions were placed on cause or duration of the symptoms. Other inclusion criteria were age 18 to 65 years and ability to read and understand English at a conversational level. People with serious comorbidities, including cancer, heart, liver, or kidney disease; blood clotting disorders; or neuromuscular disorders were excluded, as were those who sustained a cervical fracture or dislocation as a result of a trauma.

In all cases, upon providing informed consent, participants completed a general background questionnaire that captured age, sex, cause and duration of symptoms, and compensation or litigation status. Validated self-report questionnaires were used to measure pain intensity (numeric rating scale [NRS])[15] and neck-related disability (Neck Disability Index [NDI]).[16] Subsets of the overall sample also completed the following questionnaires: Pain Catastrophizing Scale (PCS, n=114),[17]

Tampa Scale for Kinesiophobia–11 (TSK, n=156),[18] Hospital Anxiety and Depression Scale (HADS, n=105),[19] and Fear of Pain Questionnaire–Short Form (FPQ, n=50).[20]

The PCS is a 13-item scale intended to measure exaggerated negative orientation toward pain. It is the most widely used pain catastrophizing scale currently available and has shown adequate psychometric properties for use in people with a variety of pain conditions, including neck pain.[6,21]

The TSK is an 11-item self-report scale that is intended to measure fear of movement or of injury or reinjury. Each item is scored on a 4-point scale from "strongly disagree" to "strongly agree." It has shown good psychometric function among a sample of patients with neck pain of various causes and durations.[22]

The HADS is a 14-item self-report scale that consists of 2 subscales: depression ($HADS_{Dep}$) and anxiety ($HADS_{Anx}$). It has proven adequately valid for clinical and research use[19,23] and is a commonly used scale in studies of musculoskeletal pain, including neck pain.[24]

The FPQ is a 20-item self-report scale that measures general fear of minor pain (eg, cutting your tongue, getting an injection) and major pain (eg, being hit by a heavy object, having a tooth drilled). Only the fear of minor pain subscale was used for this study, as it was deemed the most relevant to our purposes. The scale has been used in studies of pain-related fear, including pain of musculoskeletal origin.[25]

PPT Evaluation

Pressure pain threshold was assessed bilaterally at the angle of the upper fibers of the trapezius muscle (UFT, local) and belly of the tibialis anterior muscle (TA, distal) sites using a digital algometer (Wagner Instruments FDX-25, Greenwich, Connecticut), following the protocol described previously.[12] Briefly, a digital pressure algometer was applied over the standardized test sites with pressure increasing at a rate of approximately 5 N/s. Participants were instructed to tell the examiner the precise moment the sensation changed from pressure to slightly unpleasant pain. The examiner then repeated the test on the opposite side, and 3 tests of each site were conducted with a 1-minute rest between tests. Both the examiner and the participants were blinded to the current pressure while the testing was being conducted by ensuring the digital readout screen was facing away from both. Participants were seated for testing the UFT site and were positioned supine with the feet flat on the bed for testing the TA site. All raters underwent a standardized training program prior to initiating data collection that included written and video demonstrations and required a demonstration of adherence to the 5-N/s rate of application. In total, 15 different raters were involved in collecting data across the 3 databases. At the time, each rater was enrolled in an advanced manual and manipulative therapy training program at Western University (London, Ontario, Canada), enrollment in which is restricted to those with at least 3 years of orthopedic physical therapist practice.

One subcohort was followed over the course of 4 consecutive weeks, each providing 5 data points (1 at baseline, 1 weekly for the 4 subsequent weeks). At each of the subsequent visits, participants completed a 15-point Global Perceived Rating of Change (GPRC) asking them to indicate whether their neck condition had improved, worsened, or stayed consistent over the previous week. In the case of all but stable conditions, they also indicated the extent to which their condition had changed (1="almost the same" through 7="a very great deal"). The GPRC was specifically chosen as on omnibus indicator of overall change in neck health status, recognizing that this choice would necessarily involve consideration of symptoms and physical and emotional capacity. Participants received physical therapy treatment at the discretion of the therapist between rating sessions.

A second subcohort was evaluated by 2 independent raters. Both were experienced orthopedic physical therapists who had been trained in performing PPT measurement. The order of raters was randomized through a coin flip, and the interval between rating sessions was standardized to 5 minutes to allow time for any minor redness to subside and to prevent problems with temporal summation of pressure sensitivity. In all cases, the second rater was blinded to the findings of the first rater.

Data Analysis

In the interest of clarity, the analyses conducted are described in terms of the 3 purposes of the study.

Purpose 1: comparison of different measurement protocols. Of interest here was balance between strong clinimetric properties (inter-rater and test-retest reliability, equality of scores) and patient/clinician burden. We hypothesized that the mean of all 3 tests at both sites (mean of 6 total values) would provide the strongest measurement properties. Therefore, the properties calculated from all 3 measurements were compared against the following measurement protocols, each conducted bilaterally: result of the first test only, mean of the first 2 tests only, and mean of the last 2 tests only (disregarding the first tests as "calibration"). Differences in the mean values obtained across each of the 4 test

conditions (first test, first 2 tests, last 2 tests, all 3 tests) were evaluated using a one-way repeated-measures analysis of variance with Tukey post hoc test to control for type I error.

Interrater agreement was evaluated by calculating the random-effects model intraclass correlation coefficient (ICC [2,1]) with 95% confidence limits for each of the 4 conditions. The data for evaluation of test-retest reliability were drawn from the longitudinal cohort (n=50), including only those participants who reported a weekly interval in which the GPRC changed by no more than 1 point (either no change at all or almost the same), suggesting a subjectively stable condition over 1 week. An ICC (2,1) value greater than .80 was considered adequate for clinical use.[26] The minimum detectable change at the 90% confidence level (MDC_{90}) was calculated from the ICC [2,1] and standard deviations of the test-retest difference for each protocol. A 95% confidence interval (95% CI) was subsequently calculated around the MDC_{90} to facilitate comparison, using the technique described by Stratford and Goldsmith.[27]

Purpose 2: responsiveness of PPT.
A variety of methods have been described for estimating responsiveness of a measure.[28] For our purposes, an anchor-based method was used[28,29] in which the first weekly interval in the longitudinal subcohort was used to evaluate the ability of PPT at the UFT and TA sites to discriminate between participants who reported change on the GPRC of at least 3 points (at least somewhat better) from those who had not changed by that amount. The decision of a cutoff score of 3 rather than 2 or 4 made the most clinical sense. Change from time 1 (inception) to time 2 (1 week later) was the independent variable, and receiver operating characteristic (ROC) curves for

change at each of the 2 sites were constructed, where meaningful change was the dependent variable. Area under the curve (AUC) was estimated as an omnibus indicator of responsiveness, with an AUC of 0.50 indicating no discriminative ability beyond chance. Sensitivity, specificity, and positive and negative predictive values were calculated for a range of cutoff scores to facilitate clinical interpretation.

Purpose 3: concurrent associations between PPT and psychological constructs. An intercorrelation matrix using the Pearson r statistic was constructed using PPT at each site and NDI, NRS, FPQ–minor pain, TSK, $HADS_{Dep}$, $HADS_{Anx}$, and PCS score, assuming no major violations of normality. Where 2 or fewer responses were missing on a questionnaire, the responses were imputed with the mean of all other items on that tool. Where more than 2 responses were missing, the questionnaire was removed from the analysis. Independent group comparisons also were conducted to identify mean PPT differences between levels of sex (male/female), cause (traumatic/nontraumatic), and duration of symptoms (≤6 months/>6 months) using the Student t test after assumptions of normality and equality of variance were satisfied. In order to determine the percent variance in PPT value that could be explained by each of the psychological constructs, a series of hierarchical stepwise linear regression models was created. The sample size was too small to evaluate the influence of all variables in a single model; therefore, each model was constructed to determine the contribution of each psychological variable individually. Control variables were chosen for known group differences (male/female, traumatic/atraumatic), known association with PPT (NRS rating), or known heterogeneity within the sample (duration of

symptoms). After control variables were entered, percent variance explained (r^2) and significance of the change in F value when the psychological variable was added were used to determine the relative value of each variable. All analyses were conducted using IBM SPSS version 20 software (IBM Corp, Armonk, New York).

Sample Size Justification
Sample size is difficult to estimate with observational studies; however, the planned analyses provided enough structure upon which to calculate appropriate sample sizes for some calculations. Walter and colleagues[30] provided a robust mathematical approach to estimating the required number of participants for reliability studies. The hypothesis was that the test-retest reliability would be lower than the interrater reliability, estimated at ICC [2,1] of .8, with a null hypothesis level set at .4. Using these parameters, a sample of 27 participants provided 80% statistical power with 95% confidence ($P<.05$).

For the planned multiple linear regression, regression models with 5 variables each (sex: male/female, cause: traumatic/nontraumatic, duration: ≤6 months/>6 months, pain intensity, and one psychological variable) were created. Conservatively, we estimated that 20% of the variance in PPT would be explainable by the model with the lowest explanatory power. Using G*Power (version 3.1.2, University of Kiel, Kiel, Germany)[31] software, a minimum sample of 58 participants would be required to identify an r^2 value that deviated significantly from zero, with 95% confidence ($P<.05$) and 80% power ($\beta=0.20$).

Role of the Funding Source

Results

The entire database included 206 individuals with neck pain (76% female) of various causes and durations sampled from 6 outpatient physical therapy centers. The modal cause was motor vehicle accident (44%). Of the full sample, 73 participants were tested by 2 independent raters at the same visit, and their data were used in the evaluation of interrater reliability, whereas data for 35 participants who showed at least one stable weekly interval in their condition were used to evaluate test-retest reliability. Table 1 provides the characteristics of the entire sample and the 3 subsamples. Missing data were identified in fewer than 2% of all scale responses, requiring removal of 3 forms' data (2 PCS, 1 TSK) from the entire database.

Comparison Across Measurement Protocols

Mean values of the 4 different measurement combinations (first only, first 2 only, last 2 only, all 3 tests) showed potentially important differences (Figure). At both the UFT and TA sites, using only the results of the first test resulted in significantly higher values than any of the other combinations. Conversely, using the mean of the final 2 tests resulted in significantly lower values than any other combination. At the TA and UFT sites, all combinations differed statistically from each other at the $P<.01$ level. Although statistically significant, none of the mean differences fell beyond the standard error of measurement (SEM).

Table 2 shows the differences in interrater and test-retest reliability estimates across the 4 measurement combinations. Interrater reliability (ICC [2,1]) ranged from .85 (first test

Table 1.
Characteristics of the Overall Sample and Subcohorts Used in This Study

Characteristic	Total Sample (N=206)	Interrater Reliability (n=73)	MCID[a] (n=50)
Sex (% female)	74.5	64.4	76.5
Age (y), $\overline{X}$ (range)	44.2 (18–65)	41.7 (19–65)	39.7 (18–64)
Symptoms >6 mo (%)	55.5	41.1	29.4
Cause (%)			
Traumatic	54.9	41.1	55.0
Nontraumatic	45.1	58.9	45.0
Pain intensity (/10), $\overline{X}$ (range)	5.0 (0–10)	3.8 (0–9)	5.5 (1–9)
Neck Disability Index (/50), $\overline{X}$ (range)	15.3 (2–34)	11.3 (2–26)	15.3 (3–33)

[a] The data for test-retest reliability was a subgroup of the database used for the minimum clinically important difference (MCID). This was the same subcohort that included the Fear of Pain Questionnaire for analysis.

only, last 2 tests only, TA site) to .89 (first 2 tests only, all 3 tests, UFT site). One-week test-retest reliability ranged from .75 (first 2 tests only, UFT site) to .95 (all 3 tests, TA site) with MDC$_{90}$ varying accordingly (80.2–139.0 kPa, UFT site; 31.7–65.8 kPa, TA site). The mean time to complete all 3 tests was 13.5 minutes.

Taken together, the "all 3 tests" protocol showed the strongest measurement properties and, therefore, was used for all subsequent analyses.

Responsiveness

Using a GPRC change of at least 3 points, 13 of 50 participants changed from week 1 to week 2 in the longitudinal cohort. The ROC curves for both UFT and TA percent change suggested that PPT at the TA site was unable to discriminate between participants who had changed and those who had not, with an AUC of 0.65 (95% CI=0.46–0.84). The PPT at the UFT site showed a significant ability to discriminate change status, with an AUC of 0.76 (95% CI=0.57–0.89). Table 3 provides the sensitivity, specificity, and predictive values for a range of change scores between 50.0 and 221.5 kPa.

Associations Between PPT and Psychosocial Variables

Table 4 provides the cross-sectional associations between conditional PPT at the 2 sites and the independent variables of sex, duration, cause, PCS, FPQ, HADS, and TSK. The PPT values at the UFT site were significantly higher in male participants, in those whose symptoms had been present for greater than 6 months, and in those with nontraumatic injury mechanisms. The same pattern was seen for PPT at the TA site, with the exception of cause. Of the psychological variables, only PCS showed a significant cross-sectional correlation with PPT at the UFT and TA sites, being small and inverse (−.24 and −.21, respectively).

Table 5 presents the unique variance in PPT at either site that could be explained by each of the psychological variables after controlling for sex, duration, cause, and pain intensity. The 4 descriptive variables together explained between 8.8% and 17.3% of the variance in UFT PPT, depending on the subcohort. The PCS, HADS$_{Dep}$, and TSK scores each led to a significant increase in explanatory power over the base model at the UFT. The strongest vari-
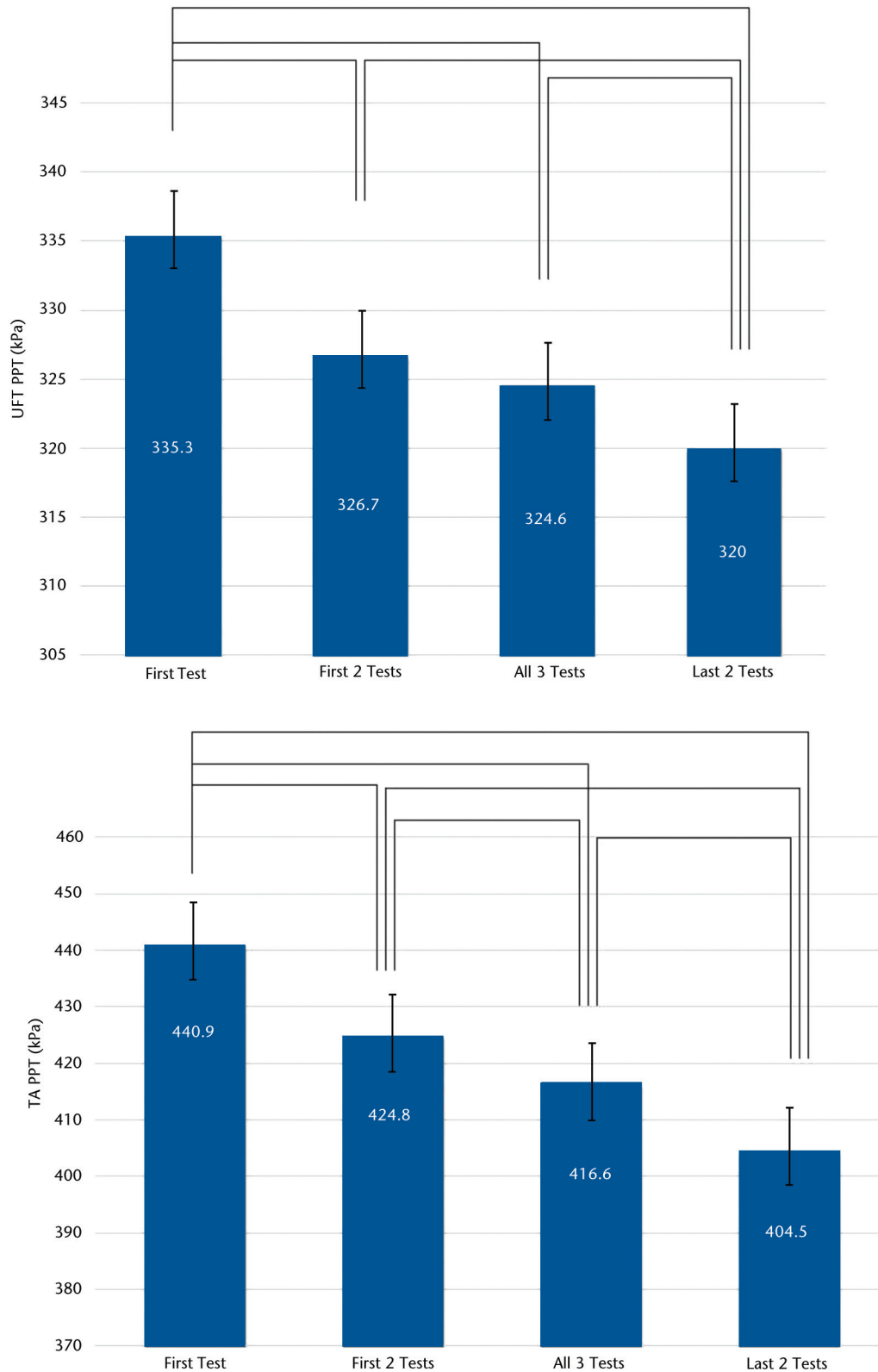
Figure.
Mean and standard error of pressure pain threshold (PPT) values across the 4 testing protocols at the (top) upper fibers of the trapezius muscle (UFT, local sites) and (bottom) belly of the tibialis anterior muscle (TA, distal sites). Lines denote significant differences at the $P<.01$ level using repeated-measures analysis of variance.

able was PCS scores, which explained an additional 5.9% of variance in UFT PPT values. At the TA site, only PCS and TSK scores led to a significant increase in explanatory power over the base model, explaining an additional 3.6% and 2.0%, respectively.

Discussion

Using a sample of 206 people with neck pain of various causes and durations who appear to be representative of previous cohorts of people seeking rehabilitation for neck pain, potentially important measurement properties were found when comparing across protocols. Use of a single or only 2 repetitions led to worse absolute and relative reliability estimates that are potentially nontrivial compared with use of the last 2 (disregarding the first) or all 3 tests. The one caveat here is in the case where the first 2 measures varied by less than the SEM as calculated in a previous article,[12] which was 20.5 kPa at the UFT site and 42.3 kPa at the TA site. Under those conditions (not shown), absolute and relative reliability estimates were nearly identical to the "all 3 tests" protocol (for UFT and TA, respectively: interrater ICC=.89 and .87, test-retest ICC=.83 and .95, and MDC$_{90}$=157.60 and 100.32 kPa) with the potential benefit of time savings. Of additional note is that the SEM for the "all 3 tests" protocol in this study was nearly identical at the TA site compared with our previous control group (42.3 kPa previous, 43.9 kPA current) but was considerably higher in this sample with neck pain (20.5 kPa previous, 66.9 kPa current). The simple interpretation is that people with neck pain represent a heterogeneous group that are more labile in their pain sensitivity about the neck.

Relying on only one measure resulted in significantly higher group mean values and worse reliability estimates, especially across a 1-week interval, which may miss important changes in the clinical status of a patient. The 1-week test-retest reliability estimates in the clinical sample are novel findings that suggest PPT at both the UFT and TA sites is adequately stable over that period when using the all 3 tests, only the last 2 tests, or a conditional measurement protocol. The interrater reliability findings are in keeping with the estimates found in a previous study using an independent cohort, reported as an ICC (2,1) value of .89 (versus .81) for the UFT site and .87 (versus .90) for the TA site.[12]

Clinicians also need to know a tool is responsive to change when it occurs. The GPRC was used to determine the ability of PPT at both sites to identify change. Using this approach, PPT at the TA site was not responsive to global change in our sample of people with neck pain, with an area under the ROC curve not statistically greater than parity (0.5). However, PPT at the UFT site did show a significant ability to detect global change (AUC=0.76). Using change scores within a clinically reasonable range (between approximately 50 and 220 kPa), PPT appears to be more valuable for ruling change in than for ruling it out. This finding was demonstrated by the considerably higher specificity and negative predictive values within this range. For example, only 8% of participants who did not report improvement changed by at least 83.85 kPa (specificity=0.92), providing confidence that change of at least that amount is unlikely to be a false-positive result. However, 50% of participants who did report improvement changed by less than 83.85 kPa (sensitivity=0.50), leading to a large proportion of false-negative results. The simple interpretation is that PPT may be one useful aid in identifying change, but a decision of whether change has occurred

Table 2.
Interrater and 1-Week Test-Retest Reliability Estimates for Each of the 4 Conditions Tested in This Study[a]

	ICC (2,1) (95% CI)							
	All 3 Tests		First Test		First 2 Tests		Last 2 Tests	
Measure	UFT	TA	UFT	TA	UFT	TA	UFT	TA
Interrater reliability (n=73)	.89 (.83–.93)	.88 (.81–.92)	.87 (.80–.92)	.85 (.77–.90)	.89 (.83–.93)	.88 (.82–.93)	.88 (.82–.92)	.85 (.77–.90)
Test-retest reliability (n=35)	.83 (.69–.91)	.95 (.89–.97)	.78 (.61–.88)	.90 (.80–.95)	.75 (.56–.86)	.94 (.88–.97)	.85 (.72–.92)	.93 (.86–.96)
MDC$_{90}$ (kPa) (n=35)	155.73 (127.87, 207.12)	103.84 (83.99, 136.05)	190.96 (154.47, 250.21)	147.04 (118.95, 192.67)	196.54 (158.98, 257.51)	112.71 (91.78, 148.66)	146.49 (118.50, 191.94)	121.34 (98.16, 158.99)

[a] Values presented are point estimate (intraclass correlation coefficient [ICC]) and 95% confidence interval (95% CI). Note that only those participants who reported no change in their condition over the course of 1 week, as indicated by the Global Perceived Rating of Change score, were used for the analysis of test-retest reliability. UFT=upper fibers of the trapezius muscle (local sites), TA=belly of the tibialis anterior muscle (distal sites). MDC$_{90}$=minimum detectable change at the 90% confidence level.

Table 3.
Sensitivity (Sn), Specificity (Sp), Positive (PPV) and Negative (NPV) Predictive Values Normalized to a 50% Pretest Likelihood of Improvement, and Positive (PLR) and Negative (NLR) Likelihood Ratios for Different Change Thresholds at the Upper Fibers of the Trapezius Muscle (UFT) Site Using the "All 3 Tests" Protocol[a]

Change (kPa)	Sn	Sp	PPV	NPV	PLR	NLR
50.04	0.50	0.82	0.46	0.84	2.78	0.61
62.50	0.50	0.85	0.50	0.85	3.33	0.59
76.51	0.50	0.87	0.55	0.85	3.85	0.57
81.62	0.50	0.90	0.60	0.85	5.00	0.56
83.85	0.50	0.92	0.67	0.86	6.25	0.54
85.18	0.42	0.92	0.63	0.84	5.25	0.63
94.52	0.33	0.92	0.57	0.82	4.13	0.73
107.42	0.25	0.92	0.50	0.80	3.13	0.82
124.77	0.17	0.92	0.40	0.78	2.13	0.90
143.46	0.08	0.92	0.25	0.77	1.00	1.00
166.36	0.08	0.95	0.33	0.77	1.60	0.97
221.52	0.08	0.97	0.50	0.78	2.67	0.95

[a] The comparator was a Global Perceived Rating of Change change of 3 points ("somewhat better") or greater.

Table 4.
Differences and Associations for Pressure Pain Threshold (PPT) at the Upper Fibers of the Trapezius Muscle (UFT) and Tibialis Anterior Muscle (TA) Sites and the Descriptive and Psychological Variables Measured in This Study[a]

Categorical Variables	UFT PPT $\overline{X}$ (SD)	TA PPT $\overline{X}$ (SD)
Sex		
Male (n=49)	422.2 (199.3)	583.6 (280.1)
Female (n=157)	295.8 (158.1)[b]	370.6 (177.0)[b]
Duration		
≤6 mo (n=94)	295.1 (164.7)	381.7 (209.7)
>6 mo (n=112)	348.5 (181.6)[c]	444.9 (232.5)[c]
Cause		
Traumatic (n=113)	300.8 (164.9)	399.2 (222.1)
Nontraumatic (n=93)	343.5 (184.7)[c]	437.6 (227.0)
Continuous Variables	**r (P)**	**r (P)**
NDI (n=206)	−.25 (<.01)[b]	−.21 (<.01)[b]
NRS (n=206)	−.09 (.21)	−.08 (.27)
PCS (n=114)	−.24 (.01)[c]	−.21 (.03)[c]
FPQ–minor pain (n=50)	.04 (.78)	.22 (.12)
HADS$_{Dep}$ (n=105)	.17 (.09)	.01 (.98)
HADS$_{Anx}$ (n=105)	.06 (.55)	−.01 (.89)
TSK (n=156)	−.13 (.12)	−.06 (.45)

[a] Differences in the categorical variables were analyzed using an independent-samples t test. Associations with the psychological variables were analyzed using the Pearson r correlation coefficient. NDI=Neck Disability Index, NRS=numeric rating scale, PCS=Pain Catastrophizing Scale, FPQ=Fear of Pain Questionnaire–short form, HADS$_{Dep}$=Hospital Anxiety and Depression Scale depression subscale, HADS$_{Anx}$=Hospital Anxiety and Depression Scale anxiety subscale, TSK=Tampa Scale for Kinesiophobia.
[b] Difference or association is significant at the P<.01 level.
[c] Difference or association is significant at the P<.05 level.

should be based on more than just PPT. Interpretation of change scores also should occur within the context of the MDC$_{90}$ of approximately 157 kPa. Stratford and Riddle[32] recently explored the paradox that arises when normal statistical variation of a measure exceeds clinically relevant change scores, and readers are directed to that article for greater exploration of this paradox. For comparison purposes, Fuentes and colleagues[33] estimated a clinically relevant change in PPT of the lumbar paraspinal muscles in healthy volunteers to be 1.16 kg/cm^2, which equates to roughly 113.9 kPa, a value that would be reasonable in our sample of people with neck pain. Our results also must be considered in light of the indicator of change used (ie, global perceived change). Global perceived change requires the respondent to consider all aspects of neck-related health, not just pain or function. It is possible that different responsiveness estimates may have been found using a different change anchor.

There are 2 broad schools of thought regarding the mechanisms of reduced pain threshold. One school of thought is that mechanical hyperalgesia represents some biological change in nociceptive processing, such as increased permeability of peripheral mechanosensitive nociceptive afferents[34,35] or increased propagation (reduced inhibition) of central nociceptive pathways.[36-38] The other potential mechanism is psychological—people who are generally more "pain averse" or in an emotionally negative state might be more apt to halt a pain threshold test early than would those who are more stoic in the face of pain.

Although it is likely that both biological and cognitive mechanisms interact to influence the point at which a patient opts to label a sensation as painful, the results of this analysis

provide an estimate of the role that psychological factors alone play in this pathway. Accordingly, it would appear that such factors play a rather small and arguably trivial role. The factor with the strongest contribution to the explanatory models, after controlling for sex, cause, duration, and pain intensity, was pain-related catastrophizing. Although the increase in explanatory power was statistically significant at the UFT site, the absolute value of 5.9% of unique variance in PPT is not large in practical terms. At the TA (anatomically distinct) site, the results were even more equivocal; the PCS and TSK were the only variables that provided a significant increase in explanatory power, and both were of small magnitude (<4%). Although we are unable to comment on the biological mechanisms of PPT, these results suggest that the psychological factors of pain catastrophizing, fear of pain, fear of movement or of injury or reinjury, general depression, and general anxiety do not appear to play a large role in influencing PPT measurement.

These findings are in keeping with those of Rivest and colleagues,[39] who found no significant relationship between PCS and local PPT in a sample of 37 people reporting neck pain in the acute stage following a motor vehicle accident. The magnitude of the correlation between PCS and PPT at the neck in that report ($r=-.22$) was nearly identical to that found in the current study ($r=-.24$). Such findings also are consistent with quantitative sensory testing in other conditions. George and Hirsh[40] evaluated the association among pain-related fear, catastrophizing, and performance on a cold pressor task in 59 patients with shoulder pain. Modest associations were reported between fear of pain and cold pressor performance of the magnitude of $r=.28$ to $.38$, whereas no significant association was found

with pain catastrophizing. Similarly, George and colleagues[41] evaluated the associations between fear-avoidance beliefs and catastrophizing with thermal pain sensitivity in a sample of 33 patients with chronic low back pain. Sex was the only unique individual factor to explain variance in thermal pain tolerance, explaining 38% of that measure. Of note was that an alternative measure of nociceptive system function, temporal summation, was more strongly associated with the psychological variables. A model including sex and catastrophizing was able to explain 63% of overall summation pain variance. It is possible, therefore, that the association between psychological variables and quantitative sensory testing differs by the mode of stimuli application. The current study also demonstrated differences in PPT by sex (female participants lower than male participants), duration (acute and subacute lower than chronic), and mechanism of injury (traumatic lower than nontraumatic). These findings are consistent with previous research[3,42-44] but do little to help explain the mechanism or identify potential treatment targets.

The question of mechanism remains unanswered, and it is likely that multiple factors converge to influence a patient's decision to label a stimulus as painful. Biology, psychology, and the environmental context all likely play a role, perhaps with one mediating the effect of another. Integrated models that include cognitive, biological, and contextual variables in the same analysis provide an interesting direction for further investigation in this field. In addition, the relative value of PPT for clinical decision making compared with other clinical techniques to assess components of nociceptive function, such as temporal summation (ie, "wind-up" pain) or effectiveness of the descending nociceptive inhibitory pathways through conditioned pain modulation, has yet to be empirically explored.

There are limitations that must be considered when interpreting these results. Perhaps the most important limitation is that the data for these analyses were drawn from 3 independent cohort studies, collected by 15 different raters across the studies. Although all raters were experienced clinical physical therapists

Table 5.
Percentage of Variance in Pressure Pain Threshold (PPT) at the Upper Fibers of the Trapezius Muscle (UFT) and Tibialis Anterior Muscle (TA) Sites Explained by Each of the Psychological Variables After Controlling for the Effects of Sex, Chronicity, Cause of Symptoms, and Pain Intensity Using Forward Hierarchical Multiple Linear Regression[a]

Measure	Unique Variance Explained (P)	
	UFT PPT	TA PPT
PCS (n=108)	5.9 (.01)	3.6 (.03)
HADS$_{Dep}$ (n=105)	4.4 (.03)	0.0 (.87)
TSK (n=156)	3.9 (.01)	2.9 (.02)
HADS$_{Anx}$ (n=105)	1.1 (.27)	0.2 (.57)
FPQ (n=50)	0.4 (.64)	1.9 (.28)

[a] The value in parentheses is the P value associated with the change in F value when each variable was added to the 3 base descriptive variables equation. PCS=Pain Catastrophizing Scale, FPQ=Fear of Pain Questionnaire–short form, HADS$_{Dep}$=Hospital Anxiety and Depression Scale depression subscale, HADS$_{Anx}$=Hospital Anxiety and Depression Scale anxiety subscale, TSK=Tampa Scale for Kinesiophobia. Note that with the exception of the 2 HADS subscales, each psychological variable was captured in a slightly different cohort.

and were trained in proper application of the protocol, it is not beyond reason to believe a systematic bias may exist in the actual application between raters. Despite rigorous attempts at standardization, PPT is still an inherently subjective measure and as such should be subject to all biases associated with subjective measurement of pain. These biases include the participant-rater relationship, environment, and level of distraction, among others. Although there is little doubt that these biases influenced the results of some participants, the degree of concern is tempered somewhat by a belief that such bias would have attenuated our findings and have hence led to overly conservative estimates of reliability. In a more controlled environment, the reliability estimates and associations reported here would likely be stronger rather than weaker, but these results provide a snapshot of the properties of this tool in a real-world environment. Related to this limitation, the data available on number of patients screened versus number eligible were inadequate to calculate recruitment ratio, allowing the potential of selection bias to have influenced the results. Another limitation is inherent to all observational study designs: only those variables that were collected can be found to be predictive of the dependent variable in linear regression. In other words, these results can only shed light on the associations between the scales used and the PPT values.

The results are only as strong as the scales used, and there may well be other psychological variables that can explain greater variance. Similarly, it is possible that a suprathreshold or pain tolerance test (maximum tolerable pain) may have led to associations of a different magnitude with the psychological constructs than the pain threshold tests used here. This suggestion has found empirical support using thermal stimuli of the trunk and extremities[45] and may also hold for neck pain. This is an avenue worth considering for those interested in using suprathreshold testing. Given the cross-sectional nature of the regression design, the study cannot provide evidence for cause-and-effect beyond strength of the association and biologic plausibility. Longitudinal designs or randomized trials are better suited for this purpose. Finally, it is possible that the angle of the UFT is not the most sensitive area in all people with neck pain, which may have reduced assay sensitivity, but for research purposes, a standardized testing location facilitates interpretation.

This study was conducted to address important pragmatic gaps in knowledge about PPT measurement. The results suggest that clinicians should expect to conduct 3 trials of PPT at each site, although 2 trials may be adequate when the first 2 trials vary by less than the SEM. This protocol provides adequate interrater and test-retest reliability for use in clinical practice and research. Local (UFT) PPT appears to be a useful tool for measuring change over time, but distal (TA) PPT is not useful for this purpose. Providing sensitivity and specificity for different magnitudes of change facilitates interpretation of change scores. Pain catastrophizing, general depression, and fear of movement or of injury or reinjury may explain some of the variance in PPT at the local (UFT) site but less so at the distal (TA) site, and none are strong enough to suggest that changing cognition will lead to appreciable change in PPT. At a correlation of magnitude $r = -.25$, it does not appear that self-reported pain intensity on an NRS and PPT are redundant but that both may provide clinicians with a better understanding of different aspects of a patient's pain experience. Clinicians are encouraged to consider both NRS and PPT when conducting a full assessment of patients with neck pain.

References

1 Nordin M, Carragee EJ, Hogg-Johnson S, et al; Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. Assessment of neck pain and its associated disorders: results of the Bone and Joint Decade 2000–2010 Task Force on Neck Pain and Its Associated Disorders. *Spine (Phila Pa 1976)*. 2008;33(4 suppl):S101–S122.

2 Sterling M. Differential development of sensory hypersensitivity and a measure of spinal cord hyperexcitability following whiplash injury. *Pain*. 2010;150:501–506.

3 Walton DM, MacDermid JC, Nielson W, et al. 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.

4 Walton DM, MacDermid J, Nielson W, et al. Pressure pain threshold testing demonstrates predictive ability in people with acute whiplash. *J Orthop Sports Phys Ther*. 2011;41:658–665.

5 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.

6 Sullivan MJ, Stanish W, Sullivan ME, Tripp D. Differential predictors of pain and disability in patients with whiplash injuries. *Pain Res Manag*. 2002;7:68–74.

7 Carroll LJ, Cassidy JD, Cote P. Frequency, timing, and course of depressive symptomatology after whiplash. *Spine (Phila Pa 1976)*. 2006;31:E551–E556.

8 Lamb SE, Williams MA, Withers E, et al. A national survey of clinical practice for the management of whiplash-associated disorders in UK emergency departments. *Emerg Med J.* 2009;26:644–647.

9 Stone AM, Vicenzino B, Lim EC, Sterling M. Measures of central hyperexcitability in chronic whiplash associated disorder: a systematic review and meta-analysis. *Man Ther.* 2013;18:111–117.

10 Petzke F, Gracely RH, Park KM, et al. What do tender points measure? Influence of distress on 4 measures of tenderness. *J Rheumatol.* 2003;30:567–574.

11 Giesecke T, Gracely RH, Grant MA, et al. Evidence of augmented central pain processing in idiopathic chronic low back pain. *Arthritis Rheum.* 2004;50:613–623.

12 Walton DM, Macdermid JC, Nielson W, et al. Reliability, standard error, and minimum detectable change of clinical pressure pain threshold testing in people with and without acute neck pain. *J Orthop Sports Phys Ther.* 2011;41:644–650.

13 Sterling M, Jull G, Vicenzino B, et al. Physical and psychological factors predict outcome following whiplash injury. *Pain.* 2005;114:141–148.

14 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.

15 Jensen MP, Turner JA, Romano JM, Fisher LD. Comparative reliability and validity of chronic pain intensity measures. *Pain.* 1999;83:157–162.

16 Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. *J Manipulative Physiol Ther.* 1991;14:409–415.

17 Sullivan MJ, Bishop SR, Pivik J. The Pain Catastrophizing Scale: development and validation. *Psychol Assess.* 1995;7:524–532.

18 Vlaeyen JW, Kole-Snijders AM, Boeren RG, van Eek H. Fear of movement/(re)injury in chronic low back pain and its relation to behavioral performance. *Pain.* 1995;62:363–372.

19 Zigmond AS, Snaith RP. The Hospital Anxiety and Depression Scale. *Acta Psychiatr Scand.* 1983;67:361–370.

20 Asmundson GJ, Bovell CV, Carleton RN, McWilliams LA. The Fear of Pain Questionnaire–Short Form (FPQ-SF): factorial validity and psychometric properties. *Pain.* 2008;134:51–58.

21 Walton DM, Wideman TH, Sullivan MJ. A Rasch analysis of the pain catastrophizing scale supports its use as an interval-level measure. *Clin J Pain.* 2013;29:499–506.

22 Walton DM, Elliott JM. A higher-order analysis supports use of the 11-item version of the Tampa Scale for Kinesiophobia in people with neck pain. *Phys Ther.* 2013;93:60–68.

23 Snaith RP, Zigmond AS. The Hospital Anxiety and Depression Scale. *Br Med J (Clin Res Ed).* 1986;292:344.

24 Berglund A, Bodin L, Jensen I, et al. The influence of prognostic factors on neck pain intensity, disability, anxiety and depression over a 2-year period in subjects with acute whiplash injury. *Pain.* 2006;125:244–256.

25 George SZ, Dover GC, Fillingim RB. Fear of pain influences outcomes after exercise-induced delayed onset muscle soreness at the shoulder. *Clin J Pain.* 2007;23:76–84.

26 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.

27 Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Phys Ther.* 1997;77:745–750.

28 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol.* 2008;61:102–109.

29 Stratford PW, Binkley J, Solomon P, et al. Defining the minimum level of detectable change for the Roland-Morris Questionnaire. *Phys Ther.* 1996;76:359–365; discussion 366–368.

30 Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med.* 1998;17:101–110.

31 Cunningham JB, McCrum-Gardner E. Power, effect and sample size using GPower: practical issues for researchers and members of research ethics committees. *Evidence Based Midwifery.* 2007;5:132–136.

32 Stratford PW, Riddle DL. When minimal detectable change exceeds a diagnostic test-based threshold change value for an outcome measure: resolving the conflict. *Phys Ther.* 2012;92:1338–1347.

33 Fuentes CJ, Armijo-Olivo S, Magee DJ, Gross DP. A preliminary investigation into the effects of active interferential current therapy and placebo on pressure pain sensitivity: a random crossover placebo controlled study. *Physiotherapy.* 2011;97:291–301.

34 Lolignier S, Amsalem M, Maingret F, et al. Nav1.9 channel contributes to mechanical and heat pain hypersensitivity induced by subacute and chronic inflammation. *PLoS One.* 2011;6:e23083.

35 Xie W, Strong JA, Ye L, et al. Knockdown of sodium channel Nav1.6 blocks mechanical pain and abnormal bursting activity of afferent neurons in inflamed sensory ganglia. *Pain.* 2013;154:1170–1180.

36 Rahman W, Bauer CS, Bannister K, et al. Descending serotonergic facilitation and the antinociceptive effects of pregabalin in a rat model of osteoarthritic pain. *Mol Pain.* 2009;5:45.

37 Rahman W, D'Mello R, Dickenson AH. Peripheral nerve injury-induced changes in spinal alpha(2)-adrenoceptor-mediated modulation of mechanically evoked dorsal horn neuronal responses. *J Pain.* 2008;9:350–359.

38 Wei H, Pertovaara A. Regulation of neuropathic hypersensitivity by alpha(2)-adrenoceptors in the pontine A7 cell group. *Basic Clin Pharmacol Toxicol.* 2013;112:90–95.

39 Rivest K, Cote JN, Dumas JP, et al. Relationships between pain thresholds, catastrophizing and gender in acute whiplash injury. *Man Ther.* 2010;15:154–159.

40 George SZ, Hirsh AT. Psychologic influence on experimental pain sensitivity and clinical pain intensity for patients with shoulder pain. *J Pain.* 2009;10:293–299.

41 George SZ, Wittmer VT, Fillingim RB, Robinson ME. Sex and pain-related psychological variables are associated with thermal pain sensitivity for patients with chronic low back pain. *J Pain.* 2007;8:2–10.

42 Hogeweg JA, Langereis MJ, Bernards AT, et al. Algometry. measuring pain threshold, method and characteristics in healthy subjects. *Scand J Rehabil Med.* 1992;24:99–103.

43 Javanshir K, Ortega-Santiago R, Mohseni-Bandpei MA, et al. Exploration of somatosensory impairments in subjects with mechanical idiopathic neck pain: a preliminary study. *J Manipulative Physiol Ther.* 2010;33:493–499.

44 Magora A, Vatine J, Magora F. Quantification of musculoskeletal pain by pressure algometry. *Pain Clinic.* 1992;5:101–104.

45 Robinson ME, Bialosky JE, Bishop MD, et al. Supra-threshold scaling, temporal summation, and after-sensation: relationships to each other and anxiety/fear. *J Pain Res.* 2010;3:25–32.