Research Report

Reliability of Clinical Pressure-Pain Algometric Measurements Obtained on Consecutive Days

Background and Purpose. Algometers have been used to measure muscle and other soft tissue tenderness. The purpose of this study was to investigate (1) "normal" pressure-pain threshold (PPT) in the biceps brachii muscle, (2) the reliability of repeated measurements of PPT in subjects without pain over 3 consecutive days, (3) the reliability of measurements of PPT between examiners, and (4) the number of measurements required to obtain a best estimate of PPT. **Subjects.** Thirty-five subjects participated in the study. **Methods.** Pain-pressure threshold of the biceps brachii muscle was measured using a Fischer algometer. Three test trials were done on each subject on each of 3 days by each of two examiners. Intraclass correlation coefficients (ICCs) and graphical methods were used to analyze the results. **Results.** The ICCs revealed almost perfect reliability for measurements of PPT within and across 3 days and substantial reliability between examiners. The best estimate of PPT was obtained using the mean of the second and third trials each day. Graphical methods demonstrated that agreement between examiners was greatest at low mean pain thresholds. There was no effect for order of examiner. **Conclusion and Discussion.** The PPT is a reliable measure, and repeated algometry does not change pain threshold in healthy muscle over 3 consecutive days. The PPT can be used to evaluate the development and decline of experimentally induced muscle tenderness. Reliability is enhanced when all measurements are taken by one examiner. [Nussbaum EL, Downes L. Reliability of clinical pressure-pain algometric measurements obtained on consecutive days. *Phys Ther.* 1998;78:160–169

Key Words: *Algometry, Delayed-onset muscle soreness, Instrument evaluation, Physical therapy.*

Ethne L Nussbaum

Laurie Downes

Algometry is a method of quantifying soft tissue tenderness. An algometer registers the force (in kilograms per square centimeter) that is applied to the tissues via a small rubber footplate. The force that is recorded is usually the amount of pressure that causes pain, called the *pressure-pain threshold* (PPT). Normal ranges of PPT have been established for some muscles and for the sites of some bony prominences.[1] The PPT has been used with individuals without pain to assess the hypoalgesic effect of physical therapy modalities. For instance, laser therapy has been applied to normal peripheral sensory nerves, and PPT was compared before and after the intervention.[2]

Algometry in Delayed-Onset Muscle Soreness

Algometers have been used to measure tenderness associated with inflammatory conditions.[1] Accumulation of fluid in intracellular or extracellular spaces as a consequence of injury raises tissue pressure and lowers PPT.[3,4] Delayed-onset muscle soreness (DOMS)[3,5] is a condition that occurs when untrained muscles perform strenuous exercise. This condition develops between 24 and 48 hours after exercise, and it can be recognized by the presence of pain on stretching, loss of force, stiffness, and tenderness in the affected muscles. Blood enzyme analysis[3] and muscle biopsy[6] reveal that there are morphological and biochemical changes in untrained muscles following strenuous eccentric exercise. Disruption of myofibrils as well as supporting connective tissue has been noted.[7] Some authors[7] have noted the presence of cellular infiltrates, including neutrophils, macrophages, and other inflammatory mediators. Tenderness is thought to be due to swelling in the myofibrils and extracellular space.

Delayed-onset muscle soreness can be induced for experimental purposes. The premise for using DOMS in research is that it results in a controlled injury being imposed on the muscles. This control allows the time course of the response to be examined, either under different exercise conditions or under different treatment conditions. We believe that algometers are quick and safe to administer and are preferred over invasive procedures as a daily measure of the effects of strenuous exercise. The absence of any abnormal tenderness in the muscle prior to inducing DOMS is a prerequisite of the model.

Algometry has been used to monitor symptoms of experimental DOMS. Jones et al[3] induced DOMS in the biceps brachii muscle and used PPT and goniometry to measure tenderness and stiffness, respectively, following the exercise. Pressure-pain threshold has also been used to

EL Nussbaum, MEd, BScPT, is Clinical Associate, Physical Therapy, Mount Sinai Hospital, Toronto, Ontario, Canada and Assistant Professor, Department of Physical Therapy, University of Toronto, 256 McCaul St, Toronto, Ontario, Canada M5T 1W5. Address all correspondence to Ms Nussbaum at the second address.

L Downes, MSc, BScPT, is Standards and Practice Coordinator, College of Physiotherapists of Ontario, Toronto, Ontario, Canada. She was Clinical Associate, St Michael's Hospital, Toronto, Ontario, Canada, at the time of the study.

This study was approved by the Ethics Committee of the Department of Physical Therapy, Faculty of Medicine, University of Toronto.

This article was adapted from a presentation at the Teamwork in Action Conference '95, Ontario Society of Occupational Therapists and Ontario Physiotherapy Association; March 24, 1995; Toronto, Ontario, Canada.

This article was submitted September 15, 1996, and was accepted September 3, 1997.

assess differences in development of DOMS. Newham et al,[5] for example, induced DOMS in the quadriceps femoris muscle and used PPT to compare the distribution of tenderness and the degree of tenderness induced by two different exercise protocols. Pressure-pain threshold also has been used to assess the effect of treatment on DOMS. Hasson and colleagues,[4,8] for example, investigated the effect of ultrasound and dexamethasone iontophoresis on DOMS in the quadriceps femoris muscle. They used PPT and a pain measure to evaluate DOMS over a 48-hour period after exercise. Intervention with ultrasound reduced the symptoms of DOMS, as evidenced by less pain being reported and higher PPT values in a treatment group than in a control group. Dexamethasone iontophoresis was found to be ineffective for the treatment of DOMS using the same outcome measures. Jones et al,[3] Hasson et al,[4] and Newham et al[5] reported changes in PPT measurements obtained after exercise as compared with PPT measurements obtained before exercise.

In none of the studies discussed was a non–DOMS control group used to examine whether PPT changed as a result of measurement procedures alone. Pressure-pain threshold testing involves forcible probing of the muscle surface. In individuals without pain, the PPT in muscle may be as high as 11 kg/cm^2. We have observed that this amount of pressure may cause bruising. We wondered, therefore, whether algometry at high pressures over the same site daily might lead to progressive lowering of PPT. We have not found studies addressing the reliability of algometric measurements over consecutive days. Thus, the reliability of PPT as an outcome measure of DOMS has not been established.

Reliability Issues in Algometry

Fischer[1] studied the reliability of algometric measurements in 10 muscles of 50 subjects without pain on a single occasion. On the basis that there was no difference between single measurements of corresponding muscles on opposite sides of the body, Fischer concluded that PPT was reproducible and proposed a range of normal values. He noted that PPT varied between individual muscles. The quadriceps femoris and biceps brachii muscles are the muscles that are examined most often in DOMS research. Fischer studied the quadriceps femoris muscle, but we have not found any investigation of normal PPT in the biceps brachii muscle. Abnormal tenderness is an exclusion criterion for studies involving DOMS.

Some authors[9–12] tested the reliability of repeated measurements of PPT. They demonstrated that, although measurements were not precise, differences between trials did not exist. Reliability was confirmed by different authors for several patterns of repetition of PPT, including 10 to 50 consecutive measurements,[9,11] trials 45 minutes apart,[9] trials 1 hour apart,[10] and trials 1 week apart.[11] Marking test sites was thought to be one method of improving the reliability of PPT measurements.[10] The reaction time of the examiner and variation in the rate of pressure increase were other factors that affected reliability.[9]

Nonelectronic algometers, such as the Fischer algometer,* depend on the operator to control the rate of pressure increase. Fischer[1] recommended a rate of 1 kg/cm^2/s. Jensen et al[9] emphasized the importance of increasing pressure at a standardized rate, based on their finding that higher PPT scores were recorded at higher application rates. Some authors[9–13] used electronic algometers to reduce variation in the rate of pressure increase; the electronic tool provides examiners with visual cues to improve their timing. Another advantage of an electronic algometer is that the reaction time of the examiner is eliminated; on reaching the pain threshold, the subject activates a button to release pressure. Jensen et al[9] thought that measurements of PPT were most reliable when the measurement site was flat, broad, and bony as opposed to a soft tissue site where the footplate might slide off the target.

Kosek et al[12] used an electronic algometer and studied three trials of PPT. In contrast to other authors, they found a decrease in PPT between trials done 10 seconds apart and an increase in PPT between trials done 20 to 30 minutes apart. The mean PPT of the three trials, however, was not different from the mean PPT of three trials after a 1-week interval. Other investigators[11,13] also used the mean of multiple trials as a criterion score to reduce variation across occasions.

Ohrbach and Gale[13] carried out a study to determine the number of measurements that gave the best estimate of PPT. They used an electronic algometer and measured facial muscles five times each, at 4- to 5-minute intervals. They found that PPT increased and decreased unsystematically from trial to trial but that there was a correlation between pairs of trials (Pearson r=.81–.91). Combining trials showed that the mean of trials 1 and 2 provided a more reliable estimate of PPT than either trial alone, and the authors reported that more than three trials was not justified by their data.

Merskey and Spear[14] investigated PPT using a nonelectronic algometer. They measured PPT twice on two separate occasions. There were no differences across the four trials. Their results, however, appear to support the idea that an electronic algometer provides more reliable

* Pain Diagnostics and Thermography Inc, 233 E Shore Rd, Suite 106, Great Neck, NY 11023.

Table 1.
Characteristics of Subjects

Subjects	Age (y)			Height (m)			Weight (kg)		
	$\bar{X}$	SD	Range	$\bar{X}$	SD	Range	$\bar{X}$	SD	Range
Female (n=30)	29.2	7.39	22–57	1.6	0.07	1.5–1.8	59	7.30	51–86
Male (n=5)	36.4	11.86	23–58	1.8	0.07	1.7–1.9	89	15.77	75–118

measurements, because they reported a lower between-trial correlation (Pearson $r=.65$) than that reported by Orbach and Gale,[13] who used an electronic instrument.

In the same work, Merskey and Spear[14] examined the interrater reliability of PPT measurements. They reported that there was no difference between examiners, although there was a tendency for one examiner to score higher than the other examiner. The correlation between examiners was reported as Pearson $r=.59$. In spite of the low Pearson correlation coefficient, the authors stated that the degree of reliability in their study supported the use of PPT for investigation of the efficacy of analgesia.

Delaney and McKee[15] also examined the interrater and intrarater reliability of PPT measurements. They used a Fischer algometer. The results of their preliminary work showed lower correlation between examiners (Pearson $r<.28$) than that reported by Merskey and Spear.[14] They attributed the finding to a difference in rate of pressure increase, and they addressed the problem prior to another study by training examiners to apply pressure while being timed. Pain-pressure threshold was then measured by two examiners alternately, at 5-minute intervals, for a total of four trials per point. Standardizing the timing of force application appears to have been an effective strategy because high interrater and intrarater reliability were reported in their final study (intraclass correlation coefficients [ICCs]=.80–.92).

In summary, algometric measurements have been shown to have good interrater and intrarater reliability when the measurements were performed once or repeatedly (2–50 repetitions) on a single day, at weekly intervals (1–5 weeks), and at longer intervals (8–12 weeks).[8–14] The reliability of measurements taken over consecutive days has not been studied. Investigators have suggested that electronic algometers provide more reliable measurements than do nonelectronic algometers.[9–12] The latter type of instrument, however, is more convenient to use and is more commonly available.

Our study was designed to (1) examine the range of "normal" PPT in the biceps brachii muscle, (2) reexamine the intertrial and interrater reliability of PPT measurements using a nonelectronic algometer on asymptomatic muscle over 3 consecutive days, and (3) establish the number of measurements needed for the best estimate of PPT.

Method and Materials

Two examiners participated in the study. They were physical therapists with many years of clinical experience but no prior experience using an algometer. One week prior to the study, the examiners practiced using an algometer while being timed. The standard was to increase pressure linearly to 5 kg/cm^2 over 5 seconds according to the method recommended by Fischer.[1] Ten practice trials were performed by each examiner. A Fischer algometer was used for the practice and test trials. The instrument has a 1-cm^2 rubber footplate and a scale marked from 2 to 20 kg/cm^2, in increments of 0.2 kg/cm^2. A new instrument was acquired for the purpose of the study. No calibration was performed.

Thirty-five subjects without complaints of pain volunteered and gave informed consent to participate in the study. There was an imbalance of female subjects in the sample (Tab. 1). The PPT of the biceps brachii muscle in the nondominant arm of each subject was measured on 3 consecutive days by each examiner.

Subjects were seated with their test arm positioned on a padded support in 90 degrees of horizontal abduction, with full elbow extension and forearm supination. The upper arm was measured, and the skin overlying the biceps muscle belly was marked with indelible ink, at a point one fourth of the distance from the elbow crease to the lateral border of the acromion. This mark established the site for all testing. Each subject's non-test arm was similarly measured and marked for the purpose of a practice session to familiarize subjects with the sensation of PPT.

Standardized instruction was given prior to each trial on all occasions. Subjects were instructed to "report as soon as the sensation of pressure changes to pain by saying 'pain,' and I will stop." The footplate of the algometer was held perpendicular to the muscle belly with the gauge turned away from the subject and the examiner. Pressure was increased at a rate of approximately 1 kg/cm^2/s until the subject reported "pain." The examiner then released the pressure and lifted the algometer off

the muscle to read the gauge and record the measurement. The needle on the gauge was returned to baseline before each trial using the pressure-release button on the algometer. Subjects were kept uninformed of their scores throughout the study to prevent subject bias from influencing the results.

The first examiner did three practice trials on each subject's non-test arm. The practice trials were followed by three trials of measuring PPT on the subject's test arm, with 10-second intervals between trials. After a 20-minute interval, the procedure was repeated by the second examiner, using the marked sites on the non-test arm followed by the marked sites on the test arm. On days 2 and 3, procedures were repeated on the test arm only, using the same sequence and timing. Thus, each subject's test arm was measured three times on each of 3 days (9 trials) by each examiner for a total of 18 trials. For 20 subjects, the order of daily testing was examiner A followed by examiner B. The order of examiners was reversed for 15 subjects. During the study, examiners did not have access to each other's scores or to their own scores of previous days. No analysis was done until data collection was complete.

Data Analysis

Intraclass correlational analyses (Shrout and Fleiss formula, ICC[2,1], a two-way random-effects layout[16]) were used to estimate interrater, trial-to-trial, and day-to-day reliability. Interrater reliability was estimated for each of the nine trials and for scores derived from the mean score of various combinations of trials.

Trial-to-trial reliability was estimated by correlating the trial 1 and trial 2 scores and the trial 2 and trial 3 scores each day, as well as by computing the correlation among all three trials on each day. Day-to-day reliability was estimated by computing the correlation between single trials of like number in the sequence of daily trials. Correlations were also calculated for day-to-day scores derived from the mean score of a combination of like-numbered trials in the sequence of daily trials.

A plot of the data showing the relationship in scores between trials and between examiners suggested that the data varied considerably from the line of equality. We considered it necessary, therefore, to further analyze the data using graphical techniques, as recommended by Bland and Altman[17] and as described.

Graphical analysis: interrater reliability of single trials. A subject's scores in a single trial recorded by each of the two examiners were paired for comparison. The mean of each pair was plotted against their difference. The overall (n=35) mean difference ($\bar{d}$) and standard deviation of the differences were calculated for each of the nine trials.

Table 2.
Repeated Measurements of Pressure-Pain Threshold (PPT) for 35 Subjects

	PPT (kg/cm²)			
	Examiner A		Examiner B	
Trial	$\bar{X}$	SD	$\bar{X}$	SD
Day 1				
1	3.41	0.98	3.27	1.37
2	3.50	1.09	3.23	1.36
3	3.54	1.23	3.26	1.46
Day 2				
1	3.41	1.32	3.05	1.16
2	3.39	1.38	2.98	1.24
3	3.39	1.44	3.12	1.48
Day 3				
1	3.39	1.28	3.01	1.15
2	3.41	1.31	3.00	1.37
3	3.44	1.34	3.09	1.52

Using the SAS Univariate procedure,[†] differences were found to be normally distributed. Most of the differences (95%) could, therefore, be expected to lie between the mean difference and approximately two standard deviations ($\bar{d} \pm 1.96SD$), which was interpreted as the "limits of agreement."[17]

Graphical analysis: interrater reliability of repeated measurements. Each examiner calculated each subject's mean score for two consecutive trials within the same day. Means were computed for trials 1 and 2 and trials 2 and 3 daily. Each subject's mean scores, derived from like-numbered trials by each examiner, were paired for comparison. The mean of the paired scores was plotted against their difference. Overall mean difference (n=35) and limits of agreement were calculated as described previously.

A correction, according to the method of Bland and Altman,[17] was applied to calculate the standard deviation of the differences between paired scores, which were derived from the means of multiple measurements. The correction was to compensate for removal of some of the measurement error. The graphical technique was not used for analysis of the mean of more than two repeated measurements because of the risk of underestimation of the standard deviation of the differences.

Graphical analysis: trial-to-trial and day-to-day reliability of single trials. The scores of each examiner were analyzed separately. A subject's score in one trial was

[†] SAS Institute Inc, SAS Campus Dr, Cary, NC 27513.

Table 3.
Intraclass Correlation Coefficients (ICC[2,1]) Between Two Examiners for Measurements of Pressure-Pain Threshold for Single Trials or for Scores Derived From the Mean of Multiple Trials for 35 Subjects

	Single Trials	Interrater ICC	Multiple Trials		Interrater ICC
Day 1	1	.74	1, 2	(mean)	.81
	2	.84	1, 2, 3	(mean)	.85
	3	.89	2, 3	(mean)	.88
Day 2	1	.75	1, 2	(mean)	.82
	2	.84	1, 2, 3	(mean)	.84
	3	.84	2, 3	(mean)	.86
Day 3	1	.78	1, 2	(mean)	.82
	2	.82	1, 2, 3	(mean)	.84
	3	.86	2, 3	(mean)	.85

paired for comparison with the score in the subsequent trial on the same day to assess trial-to-trial reliability and with the score in the like-numbered trial in the sequence of trials on the subsequent day to assess day-to-day reliability. Means and differences were plotted, and the overall mean difference (n=35) and limits of agreement were calculated as described previously.

Effect of order of examiners. A one-way analysis of variance (ANOVA) on the mean difference between examiners in trial 1 was used to assess whether the order of examiners affected the results.

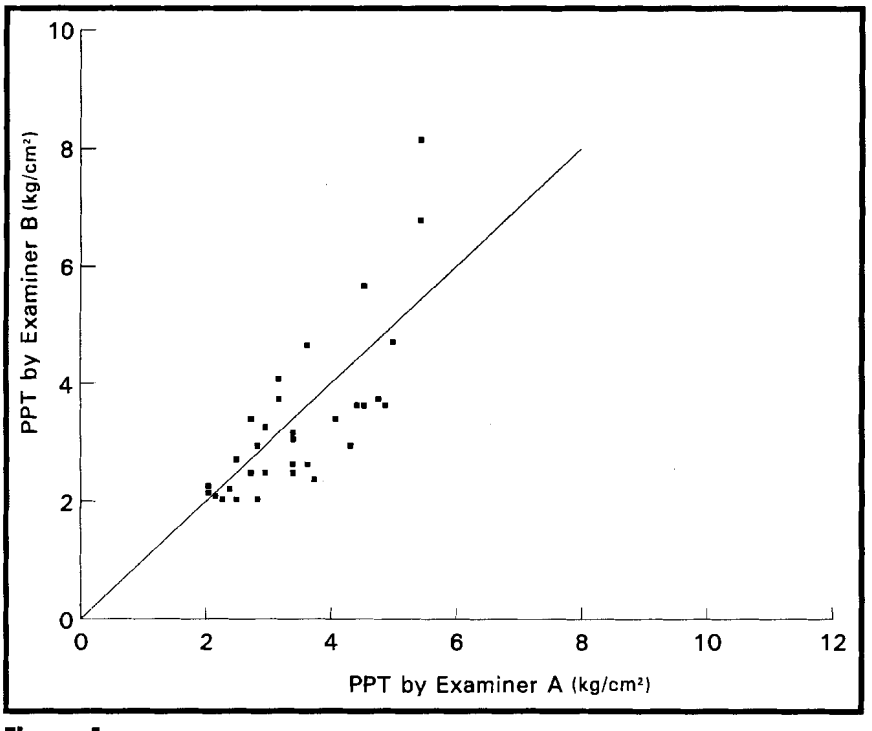

Figure 1.
Pressure-pain threshold (PPT) in trial 1 measured by two examiners, with line of equality (ICC[2,1]=.74).

Results

A total of 630 PPT scores were collected. The mean PPT and standard deviation are shown for each trial and each examiner in Table 2. Some subjects had bruising at the measurement site by the third day of the study.

Normative PPT Values for the Biceps Brachii Muscle

The mean PPT in the biceps brachii muscle was 4.63 kg/cm^2 (range=2.04–10.32) for the male subjects (90 scores) and 3.05 kg/cm^2 (range=1.81–6.80) for the female subjects (540 scores).

Interrater Reliability

Examiner A recorded higher scores than those recorded by examiner B in about 70% of the paired measurements. As mean PPT increased, however, examiner A tended to score increasingly lower than examiner B did. At a mean PPT of approximately 7.0 kg/cm^2, examiner A recorded a score that was 2.5 kg/cm^2 lower than the score recorded by examiner B. On balance, however, the mean difference between examiners (0.14 kg/cm^2 in trial 1) was small.

Table 3 shows interrater ICCs (2,1) for single trials and for mean scores derived from various combinations of trials; all correlations were significant at $P<.0001$. Each day reliability was lowest for the first of the single trials and highest for the third of the single trials (ICC=.74–.89).

Reliability improved when the mean score of the three daily trials was used rather than the score of the first or second trial of the day. The highest reliability, however, was seen when the score of the first trial of each day was omitted and the mean of the second and third trials of the day (ICC=.85–.88), or the score of the third trial alone (ICC=.84–.89), was used as the criterion score.

Figure 1 shows the relationship between examiners of the scores recorded in trial 1. The line of equality is shown on which all points would lie if the two examiners recorded identical scores. The variation from the line of equality shown in Figure 1 is typical of the results of all the single trials and prompted the additional analyses using Bland and Altman's methods.[17]

Figures 2 and 3 show the agreement between examiners using Bland and Altman's methods.[17] Each subject is

represented by a point (n=35) that shows the difference in scores between examiners against their mean score. Points on the zero line show perfect agreement. The overall mean difference between examiners is shown by a broken line. The limits of agreement are also shown ($\bar{d} \pm 1.96SD$).

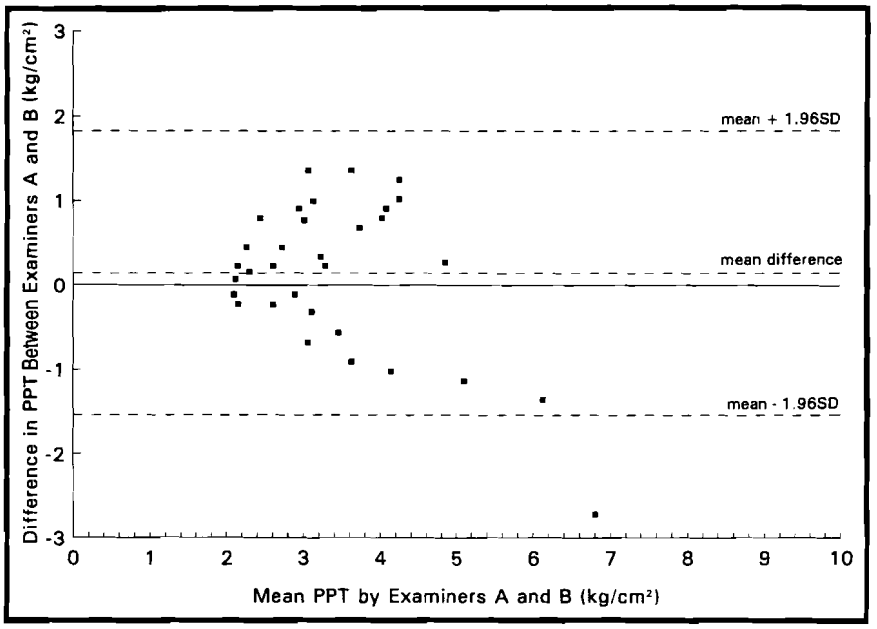

Figure 2.
Agreement between examiners for measurements of pressure-pain threshold (PPT) in trial 1. $\bar{d}$=0.14 kg/cm^2 (SD=0.86); limits of agreement were from −1.54 to 1.83 kg/cm^2.

Figure 2 illustrates the added benefit of using Bland and Altman's methods[17] for the data obtained in trial 1. The mean difference between examiners in this trial was 0.14 kg/cm^2 (SD=0.86). From the limits of agreement, it can be projected that if one examiner measures PPT once, a second examiner would score the same subjects within 1.55 kg/cm^2 below and 1.83 kg/cm^2 above the first examiner's measurement 95% of the time. Agreement between examiners was better in some later trials than in trial 1. The limits of agreement for day 2 of trial 3, for example, were from −1.22 to +1.78 kg/cm^2.

There was little change in agreement between examiners when the measure was derived from the mean score of the first two trials rather than the scores of the first trial each day. For example, when the measurement was based on subjects' mean score for day 1 of trials 1 and 2, the limits of agreement between examiners were from −1.29 to +1.71 kg/cm^2.

On all 3 days, however, agreement between examiners was best when the measurement was derived from the subjects' mean score of trials 2 and 3. Figure 3 shows the results for day 1; the limits of agreement lie between −0.97 and +1.47 kg/cm^2.

Trial-to-Trial Reliability

Tables 4 and 5 show the trial-to-trial and day-to-day reliability (ICC[2,1]) of measurements of PPT. With the exception of day 1, trial-to-trial reliability was higher between trials 2 and 3 than between trials 1 and 2. Day-to-day reliability for a single measurement of PPT was highest in trial 3, and day-to-day reliability for a measurement derived from the mean of multiple trials was highest for the mean of trials 2 and 3.

Figure 4 shows the trial-to-trial agreement of the scores recorded by examiner B on day 1 of trials 2 and 3, using Bland and Altman's method of graphical analysis.[17] The results were similar on days 2 and 3. A point is plotted for each subject (n=35) showing the difference in scores between trials 2 and 3 against their mean. Perfect agreement (zero line) and overall mean difference (−0.03 kg/cm^2, SD=0.42) between the two trials are shown. The limits of agreement lie between −0.86 and +0.79 kg/cm^2. Thus, the PPT in the third trial was between 0.86 kg/cm^2 below and 0.79 kg/cm^2 above the PPT in the second trial 95% of the time.

The same methods of analysis for the scores recorded by examiner A in trials 2 and 3 showed an overall mean difference between trials of −0.04 kg/cm^2 (SD=0.35). Limits of agreement were between −0.72 and +0.64 kg/cm^2.

Order of Examiners

The ANOVA revealed that the order in which the examiners measured PPT had no effect on the differences between their scores (P=.33).

Discussion

The purpose of our study was to establish the normal range of PPT values in the biceps brachii muscle because this muscle is frequently used in studies of experimentally induced DOMS and to examine interrater, trial-to-trial, and day-to-day reliability of algometric measurements in healthy muscle. If PPT proved to be a stable measure in the absence of pathology, then it could be used as an outcome measure in studies of experimentally induced DOMS.

We observed PPT in a group of individuals without pain whose ages were within a fairly restricted range. The group included a disproportionate number of female

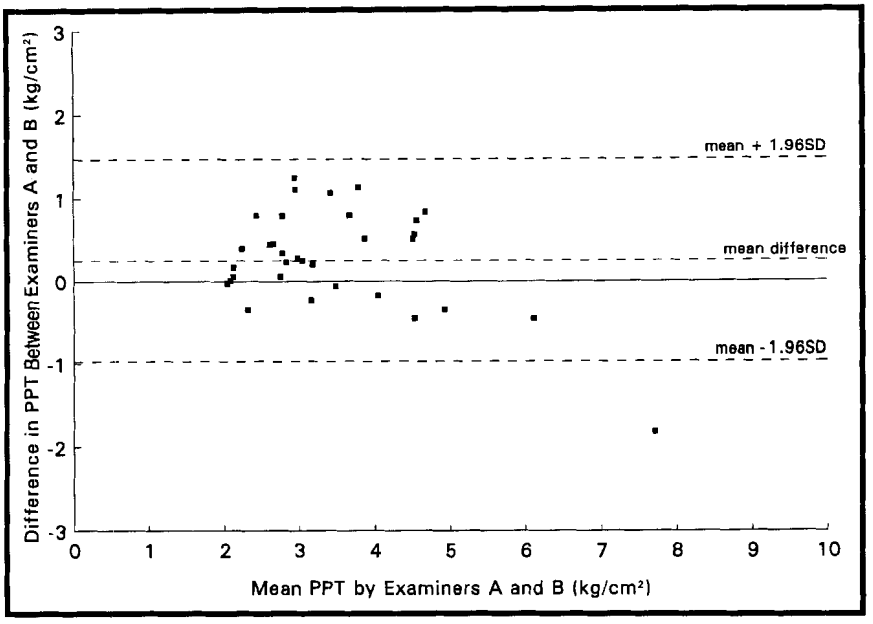

Figure 3.
Repeated-measurement agreement between examiners based on the mean score derived from measurements of pressure-pain threshold (PPT) in trials 2 and 3. $\bar{d}$=0.25 kg/cm^2 (SD=0.62); limits of agreement were from −0.97 to 1.47 kg/cm^2.

Using a measurement derived from the mean score of trials 2 and 3 daily, PPT appears to yield reliable measurements of muscle tenderness over a 20-minute period and over 3 consecutive days, according to ICC analyses. Moreover, the ICCs suggest that two examiners could be used interchangeably to measure PPT.

Delaney and McKee,[15] using the Fischer algometer on muscle, also reported lower reliability for the first of two trials. They considered their examiners to be experienced in algometric techniques, and they timed their examiners' rate of pressure application in an attempt to improve reliability. The reported reliability (ICC=.80–.92) was similar to ours, and our examiners were not timed during the testing. Ohrbach and Gale[13] similarly concluded that measurements obtained in their first trial did not agree well with the mean of five trials of PPT. On the basis of the 95% confidence interval around the mean of five trials, they recommended the use of data recorded during trial 2 or the mean of trials 1 and 2 to estimate PPT. Our results did not support the use of trial 2 alone or the mean score of trials 1 and 2.

subjects. These factors might limit the applicability of our results. The height and weight of the subjects were fairly representative of the adult population. All subjects tolerated 18 measurements over 3 days, although some subjects showed bruising at the measurement site.

Fischer[1] suggested that PPT is reproducible between individual subjects. He calculated PPT from the mean of two measurements taken on contralateral sides of the body and examined the distribution of values in a study of nine healthy muscles. Based on the standard deviation from the average logarithmic values of the PPT findings in his study, Fischer proposed that for diagnostic purposes and for quantifying pain, 84.1% of mean PPT should be considered a cutoff value for "normal." Fischer did not study PPT in the biceps brachii muscle. In our study, mean PPT for the biceps brachii muscle in the female subjects was 3.05 kg/cm^2. Using 84.1% of mean PPT as a cutoff, as recommended by Fischer, the lowest normal value for the biceps brachii muscle in female subjects would be 2.44 kg/cm^2. This value should be taken into consideration when screening subjects for admission to studies of DOMS involving the biceps brachii muscle.

Intraclass correlation coefficients appear frequently in the literature as an index of reliability. The examiners in our study were not experienced in algometric techniques. Thus, they represent examiners as broadly defined, not a particular set of examiners, and we believe that the random-effects model (ICC[2,1]) applies.[16]

Some authors[9–12] have argued that the reliability of measurements of PPT was improved by the use of

Table 4.
Trial-to-Trial Intraclass Correlation Coefficients (ICC[2,1]) for Pressure-Pain Threshold,[a] Rated by Examiner B, for 35 Subjects

	Trials	ICC	Trials	ICC	Trials	ICC
Day 1	1 and 2	.98	2 and 3	.96	All trials (1–3)	.96
Day 2	1 and 2	.94	2 and 3	.95	All trials (1–3)	.93
Day 3	1 and 2	.94	2 and 3	.98	All trials (1–3)	.95

[a] Correlations are between single trials on the same day.

Table 5.
Intraclass Correlation Coefficients (ICC[2,1]) for Day-to-Day Pressure-Pain Threshold,[a] Rated by Examiner B, for 35 Subjects

Trial	ICC
Trial 1 × 3 days	.88
Trial 2 × 3 days	.88
Trial 3 × 3 days	.89
Trials 2 and 3 (mean) × 3 days	.90

[a] Correlations are between single trials of like number in the sequence of trials on different days or scores derived from the mean of multiple trials on one day correlated with the mean of multiple trials of like number in the sequence of trials on subsequent days.

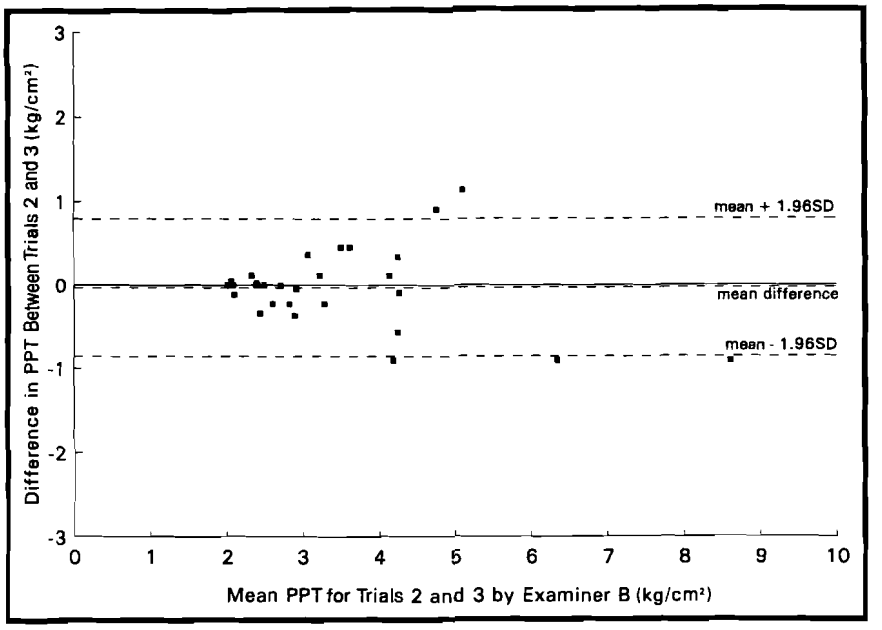

Figure 4.
Trial-to-trial agreement for measurements of pressure-pain threshold (PPT) obtained by examiner B in trials 2 and 3. $\bar{d}=-0.03$ kg/cm^2 (SD=0.42); limits of agreement were from −0.86 to 0.79 kg/cm^2.

electronically controlled instruments. In our study, we used a nonelectronic instrument. In spite of the difficulty of maintaining the recommended rate of pressure increase of 1 kg/cm^2/s using our type of algometer, there was still a positive correlation between trials in our study, and our reliability was comparable to that obtained with electronic algometers.

The use of correlation coefficients to assess the repeatability of measurements is misleading, according to Bland and Altman.[17] These authors noted that correlation coefficients measure the relationship between two measurements, not the agreement between them. Because the examiners in our study measured the same subjects, and measured them repeatedly, we would expect the scores between examiners and between trials to be strongly related. In accordance with the recommendations of Bland and Altman,[17] we plotted the measurements of one examiner against those of the other examiner to assess visually whether the data varied from the line of equality. Bland and Altman[17] noted that two sets of measurements that agree perfectly lie on the line of equality; measurements that are highly correlated lie along any straight line. A plot of our data (Fig. 1) showed considerable variation around the line of equality, especially at higher values of PPT. We believed, therefore, that additional methods of analysis were indicated.

When we used graphical methods to assess agreement, our findings supported the opinion of Bland and Altman[17] that high correlations can exist with concurrent lack of good agreement between measurements (Figs. 1–4). Graphical analysis of the data provided information about trials and raters that was not obvious from the ICCs.

From the distribution of the measurements around zero, it was apparent that although the examiners agreed well on average (ie, small mean difference), there were quite large differences between them for individual subjects (Figs. 2–4). Furthermore, examination of the relationship between differences and means showed that the differences were affected by the size of the measurement, in that differences between examiners were larger and in an opposite direction, at high mean values of PPT than at low mean values. The change of direction of differences between the examiners as mean PPT increased was unexpected and warrants some explanation. We speculated that at low mean PPT, examiner A used a faster rate of pressure application than examiner B used. Jensen et al[9] noted that higher rates tended to produce a higher PPT. We think that the reason for this finding is that examiner reaction time is slowed by high rates of pressure increase, leading to overestimation of PPT. Why then did examiner A not overestimate at high levels of PPT? A self-limiting factor might have been that examiner A was not able to maintain the high rate of pressure increase at the highest levels of force encountered in our study. This explanation would account for the change of direction of differences between examiners.

Graphical methods, in contrast to ICCs, demonstrate measurement error in units that are clinically meaningful (eg, kilograms per square centimeter) so that the consequence of differences between methods can be assessed. For example, Figure 2 shows that the limits of agreement between examiners based on a single rating of PPT were −1.5 to +1.8 kg/cm^2. Figure 3 shows that by using the mean of multiple measurements, the measurement error between examiners is reduced to −1.0 to +1.5 kg/cm^2. In our opinion, however, a difference between examiners of up to 1.5 kg/cm^2 on a measurement that has a "true" value (mean measurement of two examiners) of 3 to 5 kg/cm^2 is large, especially in view of our finding that the measurement error between trials by one examiner was from −0.9 to +0.8 kg/cm^2 (Fig. 4). Our findings from the graphical analysis suggest that one examiner should perform all measurements of PPT.

Graphical analysis of trial-to-trial agreement (Fig. 4) showed that there was no systematic bias of one trial relative to another trial (consistent upward or downward

shift in PPT). Because the results were similar for all trial-to-trial comparisons, we conclude that there was no effect of repeated use of the algometer. There was also no effect of size of measurement on the size or direction of trial-to-trial differences. Trial-to-trial results were similar for both examiners. Thus, we conclude that examiners A and B were consistent in their individual techniques, even though there were differences between their scores.

Our study lends support to previous work that has shown that measurements of PPT are highly reliable in individuals without pain.[9–13] Reliability improves when three trials are performed and data from the last two trials are used to determine PPT. Variation in rate of pressure increase may be the factor most affecting reliability. To minimize this effect, we believe that testing should be performed by one examiner.

The six PPT ratings that we performed daily for 3 days, in two sets of three with a 20-minute interval each day, is typical of measurement in intervention studies involving DOMS. Our testing procedure did not, in itself, effect a change in PPT. We have shown that PPT can be used as an outcome measure in the treatment of persons with DOMS.

Conclusion

Measurements of PPT in healthy muscle obtained with a simple nonelectronic algometer were reliable from trial to trial within the same day and from day to day over 3 consecutive days. Measurements by one examiner were more reliable than measurements between examiners. We have demonstrated that reliability is improved when the first of three trials is excluded for estimating the "true" PPT. The algometer appears to have potential for measuring day-to-day changes in soft tissue tenderness in persons with DOMS.

References

1 Fischer A. Pressure algometry over normal muscles: standard values, validity, and reproducibility of pressure threshold. *Pain.* 1987;30: 115–126.

2 Wylie L, Baxter GD, Walsh DM, Robinson L. The hypoalgesic effects of low-intensity infrared laser therapy upon mechanical pain threshold. *Lasers Surg Med Suppl.* 1995;17:9.

3 Jones DA, Newham DJ, Clarkson PM. Skeletal muscle stiffness and pain following eccentric exercise of the elbow flexors. *Pain.* 1987;30: 233–242.

4 Hasson S, Mundorf R, Barnes W, et al. Effect of pulsed ultrasound versus placebo on muscle soreness perception and muscular performance. *Scand J Rehabil Med.* 1990;22:199–205.

5 Newham DJ, Mills KR, Quigley BM, Edwards RHT. Pain and fatigue after concentric and eccentric muscle contractions. *Clin Sci.* 1983;64: 55–62.

6 Stauber WT, Clarkson PM, Fritz VK, Evans WJ. Extracellular matrix disruption and pain after eccentric muscle action. *J Appl Physiol.* 1990;69:868–874.

7 MacIntyre DL, Reid WD, McKenzie DC. Delayed muscle soreness: the inflammatory response to muscle injury and its clinical implications. *Sports Med.* 1995;20:24–40.

8 Hasson S, Wible C, Reich M, et al. Therapeutic effect of iontophoretically delivered dexamethasone on delayed muscle soreness. *Phys Ther.* 1989;69:389. Abstract.

9 Jensen K, Andersen HO, Olesen J, Lindblom U. Pressure-pain threshold in human temporal region: evaluation of a new pressure algometer. *Pain.* 1986;25:313–323.

10 Vatine J, Shapira SC, Magora F, et al. Electronic pressure algometry of deep pain in healthy volunteers. *Arch Phys Med Rehabil.* 1993;74: 526–530.

11 Brennum J, Kjeldsen M, Jensen K, Jensen T. Measurements of human pressure-pain thresholds on fingers and toes. *Pain.* 1989:38: 211–217.

12 Kosek E, Ekholm J, Nordemar R. A comparison of pressure-pain thresholds in different tissues and body regions. *Scand J Rehabil Med.* 1993;25:117–124.

13 Ohrbach R, Gale EN. Pressure-pain thresholds in normal muscles: reliability, measurement effects, and topographic differences. *Pain.* 1989;37:257–263.

14 Merskey H, Spear FG. The reliability of the pressure algometer. *British Journal of Social and Clinical Psychology.* 1964;3:130–136.

15 Delaney GA, McKee AC. Inter- and intra-rater reliability of the pressure threshold meter in measurement of myofascial trigger point sensitivity. *Am J Phys Med Rehabil.* 1993;72:136–139.

16 Shrout P, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull.* 1979;86:420–428.

17 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* February 8, 1986:307–310.