Interrater Reliability of Algometry in Measuring Pressure Pain Thresholds in Healthy Humans, Using Multiple Raters

Linda S. Chesterton, PhD,* Julius Sim, PhD,* Christine C. Wright, BSc,† and Nadine E. Foster, DPhil* 

Objective: To evaluate interrater reliability using 5 newly trained observers in the assessment of pressure pain threshold (PPT) using a fixed-angle algometer.

Methods: The study design comprised 2 phases. Phase 1: 5 undergraduate physical therapists were trained in algometry at a predefined angle, at a rate of 5 Newtons (N)/s, to the first dorsal interosseous muscle. Each observer then underwent a competency test of the application speed. The aim was to achieve repeated applications at 5 N/s without visual feedback from the algometer. Phase 2: the 5 observers measured PPT of 13 healthy volunteers, at the first dorsal interosseous muscle. The sequence of observer measurements for each participant was randomized.

Mean PPT values for each observer were analyzed using repeated measures analysis of variance, intraclass correlation coefficient (ICC2,1), and standard error of measurement, with 95% confidence intervals (CIs).

Results: No significant differences between observers’ mean values were found (P = 0.094), suggesting no bias. The ICC was 0.91 (95% CI 0.82, 0.97). The standard error of measurement value was 6.27 N/cm² (95% CI 5.35, 7.59). Differences in PPT measurements of more than 17.39 N/cm² (1.77 kg/cm²) are likely to exceed the magnitude of measurement error, and could be used to indicate true change. This margin of error is, however, somewhat larger than a previously proposed minimum clinically important difference in PPT of 14.71 N/cm² (1.5 kg/cm²).

Discussion: This study provides new evidence that trained observers can apply an algometer at a consistent rate and provide highly reliable measures of PPT in healthy humans, when PPT is calculated as the mean of 3 trials.

Key Words: reliability, multiple raters, algometry, pressure pain threshold, quantitative sensory testing

more appropriate way to analyze interrater reliability of interval or ratio data is to use intraclass correlation coefficients (ICCs).\textsuperscript{31–33} The calculation of 95% confidence intervals (CIs), to identify the precision of the estimate, has also been recommended,\textsuperscript{34,35} accompanied by the standard error of measurement (SEM), which is a measure of the (im)precision of the measurements themselves.\textsuperscript{35} To the best of our knowledge, only 2 previous studies have reported both ICCs and estimates of error within the statistical analysis,\textsuperscript{10,17} and from these it is not clear that observers were trained to control the application speed at a fixed level. Finally, fixed algometers, increasingly used in laboratory and some clinical experimental studies to standardize application angles, have not been investigated. The purpose of this study was, therefore, to determine the interrater reliability of PPT measurements in healthy volunteers using 5 inexperienced but specifically trained observers, at a single anatomic site.

**MATERIALS AND METHODS**

**Design**

This study was executed in 2 phases. The first phase comprised basic training and assessment of 5 observers in algometry and PPT measurement techniques. The primary purpose of this was to ensure a controlled rate of algometry application. The second phase was an interrater reliability study of the algometry technique using the same 5 trained observers. Before the study, University Ethics Committee approval was obtained.

**Instrumentation**

PPT was induced using a pressure algometer (Salter Abbey Weighing Machines Ltd, England) with a flat circular metal probe dressed in several layers of lint and measuring 1 cm in diameter. Force was displayed digitally in increments of 0.1 N/cm\textsuperscript{2}. The algometer was mounted vertically on a purpose-built stand to enable force to be applied perpendicular to the measurement site (Fig. 1); there was no automated control or restriction of application rate, which was controlled entirely by the observer.

**Phase 1-training and Testing of Observers**

Five female volunteer observers were recruited from a group of final year undergraduate physical therapy students (mean age 23, range: 20 to 34). We selected this population to enable the results to be generalized to observers in subsequent experimental studies investigating physical therapy interventions and to reflect the basic level of skill of a population of clinical physical therapists. Before the study, we gave each observer a full written explanation of the purpose of the experiment and the training protocol, as recommended by Goulet et al.\textsuperscript{28} After a further verbal briefing, all observers provided written consent to take part. Observers were trained by the same trainer during a single day, in the pressure algometer technique described below, so that repeated applications at the standardized rate of 5 N/s could be achieved.\textsuperscript{16,36} Practice continued until observers were able to repeat applications at the required rate without reference to the digital display of the algometer. The training and protocol incorporated suggestions made by Goulet et al\textsuperscript{28} to increase interrater reliability. These included applying the algometer at 90 degrees to a flat skin surface, standardizing and marking the area where pressure was to be applied, covering the application tip of the algometer to avoid sharp pain from the edge of the plunger,\textsuperscript{37} and applying pressure evenly and at a constant rate.

When training was complete, we tested the observers’ competence in using the algometer. This test required each observer to make 5 consecutive algometry applications at the prescribed rate of 5 Newtons (N)/s, 15 seconds apart. Each application lasted 10 seconds and was made at the first dorsal interosseous muscle of a volunteer participant, who played no other part in the study than to act as a model for this procedure. Observers were masked to the application rate by covering the digital display of the algometer, such that only the study coordinator could view and record the results. The test was considered successful if all 5 applications were carried out at a rate of 5 N/s (± 1.0 N/s) over a 10-second period,

---

**FIGURE 1.** Fixed-angle pressure algometer.
because PPT at this anatomic point is usually reported well within this timescale. All observers passed the test at the first attempt and progressed to take part in the reliability study.

**Phase 2-reliability Study**

**Participants**

For this study, the sample size was calculated so as to detect a difference in the ICC between $\rho = 0.9$ and $\rho = 0.7$ at a 2-tailed 5% significance level and with 80% power, assuming that 5 observers rated each participant. We chose $\rho = 0.7$ for the null value on the basis of its being the midpoint of the interval proposed by Shrout to constitute “moderate” reliability (0.61 to 0.80). From this calculation, the minimum number of participants required was 13 (Table 1).

Volunteers ($n = 13$; mean age 22) were recruited from the student population; they were all white and all but one was female (Table 1). Although age and sex are known to influence PPT between individuals, we found no literature to suggest that these population characteristics influence the reliability of an individual’s repeated verbal report of PPT. Before taking part, volunteers were screened for current and previous pain conditions, medication, peripheral neuropathy, neurologic pathologies, or injury to either hand. No participants were excluded. The experimental procedure was explained and written informed consent was obtained from each participant.

**PPT Measurement Procedure**

The reliability study was conducted in one laboratory at a mean temperature of 21.5°C (SD 0.73). Participants were seated in a comfortable upright position. A single study coordinator marked the location of the PPT measurement site, at the midpoint of the muscle belly of the first dorsal interosseous muscle of the dominant hand. The hand was then placed beneath the algometer such that the algometer probe would be lowered perpendicular to the marked spot. The algometer setup was the same as used previously in training the observers. Two identical calibrated algometers were available.

We explained and demonstrated the method of algometer application to each participant. Each participant then underwent 2 practice PPT measures using the nondominant hand and they were coached in differentiating their report of tactile and painful stimuli. PPT was taken as the amount of pressure required to elicit a sensation of pain distinct from pressure or discomfort. We asked participants to say “stop” as soon as a discernible sensation of pain was felt; at this point the algometer pressure was immediately released and the plunger was retracted by the observer. The algometer probe was lowered at a constant rate of approximately 5 N/s until PPT was reached, as indicated by a participant’s verbal report. Each observer took 3 PPT readings, from each participant, 15 seconds apart. This replicates recommended clinical practice. Participants were given 10 minutes rest between measurements taken by each observer, which has previously been shown to be sufficient. Neither the participants nor the observers were able to see the algometer display and all readings were recorded by the study coordinator. Thus, observers and participants were unaware of their own PPT recordings and those of other observers or participants. To avoid order effects, each observer tested participants in a computer-generated random sequence. All readings were taken on the same day.

**Data Analysis**

The mean of each observer’s 3 PPT measurements per participant was calculated, and used as the data for the main analysis. For purposes of comparison, each observer’s first measurement was also analyzed.

Graphical analysis of each observer’s ratings and repeated measures analysis of variance (ANOVA) were used to identify systematic bias (against an assumed null hypothesis of no difference between the 5 observers). The assumptions of this ANOVA model were checked before the analysis. Although the residuals showed some positive skew, this was considered to be within acceptable limits, given that the skew was largely unidirectional.

The ICC (model 2,1) was computed for interrater reliability, because the analysis of interest was to determine whether the observers could be interchangeable. This ICC is based on a 2-way random effects model with rater variance included in the ICC denominator, so that differences in level across raters count as error. To express the precision of the reliability coefficient, and identify the limits likely to contain the true reliability in the population, 95% CIs were computed for the ICC values.

The SEM provides a quantitative indicator of the measurement error in the units of the original measurements (which in this instance are the mean of each observer’s 3 consecutive measurements per participant). The SEM is the SD of the measurement errors, and is calculated as $s \times \sqrt{1 - r}$, where $s$ is the SD of the observed ratings and $r$ is the reliability coefficient (here, the ICC). A small value for the SEM indicates low measurement error, and thus high reliability. We can be 95% confident that an interval of 1.96 × SEM either side

<table>
<thead>
<tr>
<th>TABLE 1. Summary of Reliability Study Protocol Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. participants (n)</td>
</tr>
<tr>
<td>Male: female</td>
</tr>
<tr>
<td>Mean (range) age</td>
</tr>
<tr>
<td>No. observers</td>
</tr>
<tr>
<td>Male: female</td>
</tr>
<tr>
<td>Mean (range) age</td>
</tr>
<tr>
<td>No. measures taken by each observer from a single participant</td>
</tr>
<tr>
<td>Interval between measurements for a single observer</td>
</tr>
<tr>
<td>Interval between each observer per participant</td>
</tr>
<tr>
<td>Total no. measures per participant</td>
</tr>
<tr>
<td>Total no. measures per observer</td>
</tr>
</tbody>
</table>
of an individual’s observed score contains that individual’s true score (assuming that measurements are not systematically biased). The SEM can also be used to identify the minimum difference in the measure that must be exceeded to demonstrate true change, as opposed to random measurement fluctuation. Thus, for example, we can be 95% confident that a difference of greater than $1.96 \times \sqrt{2}\text{SEM}^2$ between 2 measurements represents true change.35 We also calculated CIs for the SEM, via the formula for the CI of a variance.36 Spearman rho ($r_S$) was used to assess the correlation between observers’ mean ratings and the order in which they were taken.

Data were analyzed using SPSS version 14. We took $P \leq 0.05$ (2-tailed) as the cut-off for statistical significance.

RESULTS

The data for the reliability analysis (ie, mean values of the 3 measurements taken from each participant by each of the 5 observers) are given in Table 2. Figure 2 provides a plot of these values. This plot suggests that agreement between observers is generally high, being slightly lower for high PPT values (eg, participants 1, 6, and 10).

The ICC was very high (ICC = 0.91; 95% CI 0.82, 0.97); this value is significantly larger than the null value of 0.7 ($P < 0.0005$), and the narrow CI indicates the high precision of the estimate. The SEM was calculated as $6.27 \, \text{N/cm}^2$ (95% CI 5.35, 7.59). On the basis of a SEM of this magnitude, we can be 95% confident that (1) an interval of $12.30 \, \text{N/cm}^2$ (1.25 kg/cm$^2$) either side of the observed PPT measurement contains the true value, and (2) a difference between 2 observers’ PPT measurements of more than $17.39 \, \text{N/cm}^2$ (1.77 kg/cm$^2$) represents a true difference.

The ICC of 0.91 for the mean of the 3 measurements for each observer was somewhat higher than that obtained when only the observers’ first measurement was analyzed; for the latter analysis, ICC = 0.90 (95% CI 0.79, 0.96), and $P = 0.001$ for the statistical test against the null value of 0.7. The SEM calculated on the observers’ first measurement was $6.66 \, \text{N/cm}^2$ (95% CI 5.68, 8.06), such that we can be 95% confident that a difference between 2 observers’ first PPT measurements of more than $18.47 \, \text{N/cm}^2$ (1.88 kg/cm$^2$) represents a true difference.

The bottom row in Table 2 shows the means for each observer across all 13 participants. Repeated measures ANOVA showed no statistically significant difference across these values ($F_{4,48} = 1.000, \, P = 0.417$), which suggests that there was no bias in ratings across observers. Additionally, there was a nonsignificant correlation between the means of the 15 ratings and the sequence in which they were performed ($r_S = 0.343, \, P = 0.211$), which suggests that there was no systematic change in the participants’ PPT values over time.

DISCUSSION

The value of most outcome measures and assessment techniques is in part dependent upon high intrarater and interrater reliability. The purpose of this study was to establish the interrater reliability of PPT measurement made using a pressure algometer, and with multiple observers trained in applications at a consistent rate.

The results showed that the training of observers in the algometry application technique was successful, with all 5 observers passing the competency test at the first attempt and within a single day. This addresses the observer error introduced by variable rates of application and is, therefore, an important feature of the study. Most previous studies have not reported whether, or how, such testing of observers’ performance has taken place. The interrater ICC for the mean measurements was excellent at 0.91, with a narrow CI. These results are consistent with those of Antonaci et al,17 Delaney and McKee,20 and Nussbaum and Downes10; notwithstanding the use of different study protocols, these authors also showed high interrater reliability (with ICCs between 0.75 and 0.88), in

| TABLE 2. PPT Measurements in N/cm$^2$ for Each Participant by Observer (Data in the Cells are the Mean of 3 Measurements Per Participant by Each Observer) |
|-----------------|---|---|---|---|---|---|
| Participant     | 1  | 2  | 3  | 4  | 5  | Mean (SD) |
| 1               | 89.4 | 91.6 | 77.9 | 78.4 | 100.7 | 87.6 (9.6) |
| 2               | 32.8 | 33.3 | 33.0 | 33.1 | 34.4 | 33.3 (0.7) |
| 3               | 14.7 | 12.3 | 15.8 | 25.8 | 21.4 | 16.5 (5.3) |
| 4               | 31.1 | 25.7 | 28.3 | 28.4 | 26.9 | 28.1 (2.0) |
| 5               | 32.7 | 27.8 | 39.7 | 34.3 | 31.5 | 33.2 (4.4) |
| 6               | 67.1 | 47.3 | 71.1 | 63.3 | 57.6 | 61.3 (9.3) |
| 7               | 32.7 | 14.1 | 25.8 | 27.9 | 22.5 | 24.6 (6.9) |
| 8               | 26.7 | 23.5 | 40.0 | 36.9 | 32.4 | 31.9 (6.9) |
| 9               | 36.2 | 54.1 | 37.8 | 26.0 | 34.1 | 37.7 (10.3) |
| 10              | 58.7 | 55.8 | 73.1 | 59.8 | 66.8 | 62.9 (7.0) |
| 11              | 36.3 | 37.5 | 40.0 | 38.6 | 26.4 | 35.8 (5.4) |
| 12              | 13.5 | 15.2 | 20.8 | 18.3 | 20.3 | 17.6 (3.2) |
| 13              | 18.6 | 14.6 | 13.1 | 24.2 | 12.8 | 16.6 (4.8) |
| Mean (SD)       | 37.7 (21.8) | 34.8 (22.7) | 39.7 (21.5) | 38.1 (17.9) | 37.0 (24.6) | — |

© 2007 Lippincott Williams & Wilkins

763
healthy humans. The current study, however, used a larger number of observers and reports the statistical conclusions with a stated power.

Not unexpectedly, reliability was somewhat greater for the observers’ mean values than for their first ratings, as has been reported previously. It is advisable, therefore, that mean rather than single measurements should be used in multiple-observer studies of PPT.

From the SEM of 6.27 N/cm² (0.64 kg/cm²) obtained in this study, it is highly likely that measurement error will be no greater than 12.30 N/cm² (1.25 kg/cm²), and that a difference between 2 observers’ measurements of PPT of more than 17.39 N/cm² (1.77 kg/cm²) will indicate a true difference, when using multiple trained raters. This value for the interobserver SEM is higher than the figure of 5.74 N/cm² (0.59 kg/cm²) calculable for the data reported by Antonaci et al,17 and that of 5.06 N/cm² (0.52 kg/cm²) calculable for the data reported by Nussbaum and Downes10 despite their reporting lower ICCs. The SDs of the PPT measurements were smaller in these studies than in our study, which is likely to have constrained the values they obtained for the ICC, but would not have affected the SEM.

Fischer17 suggested meaningful clinical differences in PPT measures would range from 1.5 to 2 kg/cm² (14.71 to 19.61 N/cm²). This suggestion was based on the comparison of PPTs in healthy tissue with that of pathologic (tender) tissue on the corresponding and contralateral anatomic site in 50 volunteers (tender spots correlated as “hot spots” when viewed thermographically). The magnitude of measurement error of 17.39 N/cm² (1.77 kg/cm²) in our study, using mean measurements, is therefore larger than the lower boundary of Fischer’s estimate of a minimum clinically important difference. Therefore, despite the high ICC, we cannot be highly confident that a clinically important change will not be masked by measurement error, when using multiple observers and mean measurements.

Further, findings from the current study show that participants are reliable in making a consistent verbal report of their own PPT with 15 repeated measures over a 1-hour period, which is consistent with findings of Chesterton et al13 and Jensen et al,48 but represents new information because ratings are from multiple observers. There was also evidence of heteroscedasticity (where larger differences between raters were recorded for participants showing higher PPT thresholds). This finding is of interest, and possible explanations are that some examiners were unable to maintain pressure of application at higher levels of PPT, or that participants were unable to discern reliably the threshold point at high levels of pressure and tissue distention. However, the change in angle (or possible movement/wobble) of the probe that may be produced with high pressure manual algometry clearly plays no part. The role of structured training and competency testing may also have played a part in the observed differences with previous studies, but this is difficult to quantify because Nussbaum and Downes10 provided training but did not undertake a competency test and Antonaci et al17 did not document any training. Hence, the consistency of application rate cannot be commented upon. The comparability in terms of developed “skill” in the technique is, therefore, difficult to determine.

Within the data from the present study, there is a small trend toward lower PPT ratings on second assessments compared with the first and third by each observer. Variations in individual readings have been noted by other authors,10,22,41 and mean data are, therefore, likely to give better estimates. Order effects for observers were counteracted by the random order in which observers assessed participants. The findings of this study relate to PPT measurements taken at the first dorsal interosseus muscle of the hand, using a fixed angle (via a stand). They may not necessarily be generalizable to other measurement sites, or to situations in which a purely manual method of measurement is used.

With reference to the statistical analysis used in this study, some authors maintain that the ICC statistic may produce misleading results through underestimating or overestimating reliability.30,49 Heterogeneity in individual levels of PPT was observed between participants, and this is a well documented phenomenon in PPT measurement.4,5,48 In common with the Pearson correlation coefficient, a wide versus a narrow range of PPT measures (ie, heterogeneous population) will lead to relatively higher ICCs (though researchers should not be tempted to select a sample that is more heterogeneous than the population of interest for the sake of larger ICC values). The large interindividual differences demonstrated in the present study may, therefore, have influenced the magnitude of the ICCs obtained. For this reason, the
ICCs should be interpreted alongside the SEMs, as the latter are not affected by the range of the measurements, and thus generalizable across samples with differing degrees of variability.

An approach to analyzing agreement through bias and limits of agreement has been proposed as the method of choice, and was used by Nussbaum and Downes. This method, which relies on graphic display and analysis of paired differences, is impractical in the case of multiple observers; there are, for example, 10 pairwise analyses possible between the observers in this study. We feel the approach to analysis in the current study is, therefore, a useful combination of methods. The familiar skewness of PPT measures can, however, present difficulties for reliability studies.

Our study provides new evidence that multiple, trained observers using a fixed-angle algometer, demonstrate high interrater reliability in the measurement of PPT in healthy humans. The goal of the training was to ensure a consistent and constant speed of application, which can be achieved through basic instruction and practice. This level of competency is easily achieved with this algometer and does not require specialist training; in this respect our findings are generalizable. Also, as our observers were undergraduate students, they can be assumed to represent a basic level of skill in the measurement of PPT; therefore, our findings may well underestimate the reliability of measurements taken by more experienced clinicians. The way in which this algometer is applied should not differ between patients and healthy humans, and it is reasonable to assume that our findings would apply to both healthy and clinical populations in the research situation. It must be acknowledged, however, that the algometer stand used in this study is limited in its application to areas that can be accommodated under the algometer probe, such as the hand or forearm. This apparatus could not be used to measure PPT in regions such as the neck or back, and our hand or forearm. This apparatus could not be used to accommodate under the algometer probe, such as the study is limited in its application to areas that can be edged, however, that the algometer stand used in this populations in the research situation. It must be acknowledged, however, that the algometer stand used in this study is limited in its application to areas that can be accommodated under the algometer probe, such as the hand or forearm. This apparatus could not be used to measure PPT in regions such as the neck or back, and our hand or forearm. This apparatus could not be used to accommodate under the algometer probe, such as the

We conclude that multiple trained examiners and this algometry technique have the potential to achieve highly reliable measurements of PPT in multibserver trials. In this study, changes of greater than 17.39 N/cm² (1.77 kg/cm²) can be confidently regarded as representing true change. Notwithstanding the high value of the ICC, the measurement error estimated in this study could mask a clinically important difference. This highlights the importance of calculating measures of both reliability and measurement error.

REFERENCES


