RELIABILITY OF THERMAL QUANTITATIVE SENSORY TESTING OF THE HAND IN A COHORT OF YOUNG, HEALTHY ADULTS

NIAMH A. MOLONEY, MSc,1 TOBY M. HALL, PhD2 TOMAS C. O’SULLIVAN, BSc,1 and CATHERINE M. DOODY, PhD1

1 School of Public Health, Physiotherapy, and Population Science, University College Dublin, Belfield, Dublin 4, Ireland
2 School of Physiotherapy, Curtin Innovation Health Research Institute, University of Technology, Perth, Western Australia, Australia

ABSTRACT: Introduction: The reliability of thermal quantitative sensory testing (QST) has yet to be fully established. In this study we investigated intra- and interrater reliability of thermal QST in a blinded manner. Methods: Two investigators recorded thermal detection and pain thresholds on the hand of 22 volunteers, twice on two occasions. Results were analyzed using descriptive statistics, intraclass correlation coefficients (ICCs), and coefficients of variation (CVs). Results: Mean intra-individual differences were small for all measures except cold pain thresholds. ICC values for intra- and interrater reliability were: cold detection, 0.27–0.55; warm detection, 0.33–0.68; and heat pain, 0.39–0.86. Cold pain yielded high ICC values (0.87–0.94), but also high CV (84.9–90.2%). Conclusions: In young, healthy adults, thermal detection and heat pain thresholds of the hand demonstrated good reliability for group comparisons and individual analyses. Cold pain threshold measures may be suitable for group comparisons, but a large variance in the data limits individual analyses.

Muscle Nerve 44: 547–552, 2011

The use of thermal quantitative sensory testing (QST) in assessing painful musculoskeletal disorders has increased substantially in recent years with many different syndromes being studied, including whiplash and low back pain.1–5 Thermal QST is a psychophysical means of assessing the function of small-diameter nerve fibers and their respective pathways and has been used in the development of somatosensory profiles or phenotypes within patient populations.6–8 The identification of somatosensory profiles may be useful in understanding underlying pain mechanisms.7 Furthermore, thermal QST has been used as an outcome measure in intervention studies.9,10 Thermal QST comprises cold detection (CDT), warm detection (WDT), cold pain (CPT), and heat pain (HPT) thresholds.11 As a psychophysical measure, QST is not objective and consistency in QST data relies heavily on environmental factors such as ambient temperature and noise, methodological factors, as well as the cooperation and attention of the individual being tested.12 For any measure to be clinically useful or robust enough for research purposes, it must be both reliable and valid. Reliability refers to the consistency of a measurement across time, patients or observers, and the extent to which it is error-free.13 Three categories of reliability have traditionally been assessed for QST: intrarater reliability (the reliability within a single rater); interrater reliability (the reliability between two or more raters); and test–retest reliability (the reproducibility across repeated measures irrespective of the raters).14,15 Reliability relating to thermal QST has yet to be established as shown in a previous literature review whereby reliability results varied from poor to excellent.12 This affects the utility of such tests in both research studies and clinical practice.

On reviewing the literature relating to thermal QST, it is apparent that there is marked variability in the methodological quality of published studies, with many studies providing insufficient information regarding: (i) the test investigators (in terms of training of investigators in use of QST)14; (ii) blinding procedures;14,17–20,22,23; and (iii) stability or precision of a measurement in support of reliability estimates.16,18,20,22,25,26 The importance of these features has been highlighted by Lucas et al.27 in their development of QAREL, a quality appraisal tool for reliability studies where, for example, 5 of its 11 items pertain to blinding.

Given the growing use of QST in profiling patients with musculoskeletal disorders and the more recent use of QST as an outcome measure in intervention studies, it is important to establish the robustness of these measures through well-designed studies. It is also important to assess whether thermal QST measures are suitable for group comparisons and/or clinical use on individual patients. In addition, the reliability of thermal pain thresholds has been less well explored than detection thresholds and warrants further analysis. Therefore, the aim of this study was to assess intra- and interrater reliability of thermal detection and pain thresholds incorporating the fundamental aspects of reliability study design as identified by Lucas et al.27 These aspects consist of adequate training of investigators, blinding, and standardization of test protocols including environmental factors.

METHODS

Participants. Twenty-two healthy subjects (10 males; mean age 30.7 years, SD 7.3 years, range 23–50 years) volunteered for this study. All participants...
were volunteers recruited from the staff and student population of University College Dublin through an e-mail/poster campaign. Inclusion criteria consisted of good health status with no history of neck or upper quadrant pain. A medical history was taken, and participants were excluded if they had any of the following: diabetes; thyroid or endocrine dysfunction; significant spinal pain; generalized neurological/rheumatological disorders; psychiatric disorders; previous chemotherapy; or if they were pregnant. Participants on regular pain or psychotropic medication were also excluded. The study was approved by the Human Research Ethics Committee, Life Sciences, University College Dublin, in accordance with the Helsinki Declaration of 1975. All participants provided written informed consent prior to participation.

Investigators. Data were collected by two investigators from the University College Dublin. Both investigators were physiotherapists, one with 11 years of clinical experience (N.M.) and the other newly qualified (T.O.S.). The primary investigator (N.M.) had undergone training in QST (10 hours formal training with 8 months experience using QST in both healthy and patient populations) and had trained the second investigator (T.O.S.) in its use (20 hours training on healthy subjects). Prior to commencement of the study, independent trials were conducted on persons not included in the main study to ensure consistency between investigators in relation to probe placement, instructions given, and recording of results.

Study Design. Testing was conducted in a quiet, isolated room, free from outside distractions, with an ambient temperature between 21° and 25°C. Participants were tested twice (once by each investigator) on two occasions 2 weeks apart, that is, four tests in total. The second test on each occasion was performed immediately after the first. This was considered reasonable as the testing duration was short at approximately 10 minutes, and previous studies on a normal population showed no effect for repeated measurements of CDT, WDT, or CPT and, although differences were found for HPT, they were still considered reliable. Tests were performed unilaterally, but the side to be tested and the order of first investigator were randomized for each participant using a random number generator. The same side was then tested in each participant on all four test occasions, and the order of investigator was maintained for both days. All tests were performed using a NeuroSensory Analyzer (TSA II; MEDOC, Israel).

A Peltier thermode (16 × 16 mm) was attached directly over the area of the hand innervated by C7 (dorsum of the second metacarpal). This site was chosen as it is one of the test sites to be used in a parallel investigation, and has been used in previous studies. The baseline temperature was preset to 30°C and increased or decreased at a rate of 1°C per second. CDT, WDT, CPT, and HPT were measured using the method of limits, in that order. The interstimulus interval was at least 4–6 seconds for detection thresholds and at least 10 seconds for pain thresholds. To prevent thermal injury, cut-off levels were set at 50°C and 0°C. Standardized instructions were read out to the participants prior to testing. In the case of CDT and WDT, participants were asked to indicate when they noticed the temperature of the thermode changing from neutral to warm/cool, and in the case of CPT and HPT participants were asked to indicate when they noticed the heat or cold becoming painful. Participants were asked to press a patient control switch at each point. They were advised that pain threshold was not a measure of pain tolerance. All measures were taken in triplicate with the mean values used for analysis. Participants were positioned such that they were unable to see the computer screen and output generated and were blinded to their previous results. Investigators were blinded both to their own and the other investigators’ results throughout the testing sessions.

Statistical Analysis. All statistical analyses were performed using SPSS, version 15 (SPSS, Inc., Chicago, Illinois). As CPTs and HPTs were not normally distributed, data were log-transformed before further analysis. Cold and heat pain thresholds were normally distributed and, therefore, the absolute values were used for analysis. Mean values and standard deviations (SDs) were calculated for all continuous variables. In addition, mean intrapersonal differences were calculated for each pair of data sets used for analysis. Intra- and interrater reliability were analyzed using intraclass correlation coefficients (ICCs) with 95% confidence intervals. For intrarater reliability, the ICC model 3,2 was used, accounting for measures taken over two test sessions. For interrater reliability, the ICC model 2,2 was used. ICC values were interpreted according to guidelines established by Shrout and Fleiss, where values >0.75 indicate excellent reliability, 0.6–0.75 is good reliability, 0.4–0.59 is fair reliability, and <0.4 is poor reliability. The coefficient of variation (CV) was used to assess response stability and was calculated as: SD / mean × 100. CVs should be as low as possible with values <10% considered good. However, there is dispute regarding the value of using arbitrary cut-off points, with recommendations that data should be interpreted in the context of the actual results.
How reliability is determined statistically is an area of debate, and no standard method is currently widely accepted. Given the limitations that exist with each estimate of reliability, it has become more common to see compound measures being presented in research, a practice recommended by Lucas et al. They proposed that appropriate statistical analysis involves the presentation of both an appropriate estimate of reliability and an appropriate measure of precision. As such, in our analysis, we used a compound analysis of the mean–intraindividual differences and ICC and CV values to determine reliability.

RESULTS
All subjects completed the study. Table 1 lists means, SDs, and ranges for all measures. Tables 2 and 3 show mean intraindividual differences for intra- and interrater reliability. Mean intraindividual differences were small for CDT and WDT and, although they were larger for HPT, values were still small when compared with mean values. Mean intraindividual differences for CPT had the highest SDs, which reflects the wide variation in the data.

The ICCs, 95% confidence intervals, and CVs for intra- and interrater reliability are presented in Tables 2 and 3. CDT and WDT demonstrated poor to good ICC values for inter- and intrarater reliability, but CV values were very low for CDT and moderately low for WDT. CPT demonstrated the highest ICC values for intra- and interrater data. However, CV values were extremely high, as were the range of values and SDs. ICC values for HPT ranged from poor to excellent; conversely, CVs were consistently low. Compound measures of reliability are presented in Tables 2 and 3 for each parameter.

CPT could not be established within the temperature interval used for three participants during all four test sessions; that is, participants did not determine cold as painful before the cut-off of 0°C was reached. This was found in a further 3 participants during at least one of their four test sessions. HPT could not be established for the same reason in one participant but only during the second test session. CPT consistently >15°C was found in 7 participants.

DISCUSSION
The aim of this study was to assess intra- and interrater reliability of thermal QST of the hand incorporating fundamental aspects of reliability study design as suggested by Lucas et al. CDT and WDT data displayed very small differences in mean values and mean intraindividual differences between examinations or between raters and thus can be considered to have good intra- and interrater reliability. CDT was found to have better CV values than WDT, but both could be considered good. Despite low CVs for CDT and WDT, ICC values ranged between poor and good. This apparent contradiction may reflect a limitation in using the ICC statistic as a reliability measure in some circumstances. ICC is a measure of relative

<table>
<thead>
<tr>
<th>Table 1. Means and standard deviations of thermal detection log-transformed data and thermal pain absolute thresholds.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDT</td>
</tr>
<tr>
<td>I1 S1</td>
</tr>
<tr>
<td>I2 S1</td>
</tr>
<tr>
<td>I1 S2</td>
</tr>
<tr>
<td>I2 S2</td>
</tr>
<tr>
<td>Data expressed as mean (SD), range. Thresholds presented as absolute values in degrees Celsius. CDT, cold detection threshold; WDT, warm detection threshold; CPT, cold pain threshold; HPT, heat pain threshold; I1, investigator 1; I2, investigator 2; S1, session 1; S2, session 2.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2. Intrarater reliability: compound measure for thermal detection and pain thresholds.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID (SD)</td>
</tr>
<tr>
<td>CDT session 1</td>
</tr>
<tr>
<td>CDT session 2</td>
</tr>
<tr>
<td>WDT session 1</td>
</tr>
<tr>
<td>WDT session 2</td>
</tr>
<tr>
<td>CPT session 1</td>
</tr>
<tr>
<td>CPT session 2</td>
</tr>
<tr>
<td>HPT session 1</td>
</tr>
<tr>
<td>HPT session 2</td>
</tr>
</tbody>
</table>

CDT, cold detection threshold; WDT, warm detection threshold; CPT, cold pain threshold; HPT, heat pain threshold; MID, mean intraindividual difference (i.e., mean difference between participants' individual scores); SD, standard deviation of intraindividual differences; ICC, intraclass correlation coefficients; CI, confidence interval; CV, coefficient of variation (%). Intra- and interrater reliability for all continuous variables: ICC <0.4 considered poor agreement; 0.4–0.59 fair agreement; 0.6–0.75 good agreement; and >0.75 excellent agreement. CV% should be interpreted in the context of the overall data. Lower CV percentages indicate less variance between the data. CV values <10% have been suggested as good.29
Table 3. Interrater reliability: compound measure for thermal detection and pain thresholds.

<table>
<thead>
<tr>
<th></th>
<th>MID (SD)</th>
<th>ICC (95% CI)</th>
<th>CV%</th>
<th>Overall reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CDT session 1</strong></td>
<td>0.14 (2.1)</td>
<td>0.27 (-0.16–0.61)</td>
<td>5.50</td>
<td>Good</td>
</tr>
<tr>
<td><strong>CDT session 2</strong></td>
<td>0.25 (1.4)</td>
<td>0.55 (0.17–0.78)</td>
<td>4.80</td>
<td>Good</td>
</tr>
<tr>
<td><strong>WDT session 1</strong></td>
<td>0.3 (3.6)</td>
<td>0.38 (-0.35–0.68)</td>
<td>8.70</td>
<td>Good</td>
</tr>
<tr>
<td><strong>WDT session 2</strong></td>
<td>0.4 (2.8)</td>
<td>0.69 (0.38–0.86)</td>
<td>7.80</td>
<td>Good</td>
</tr>
<tr>
<td><strong>CPT session 1</strong></td>
<td>2.2 (4.6)</td>
<td>0.88 (0.73–0.93)</td>
<td>84.9</td>
<td>Fair</td>
</tr>
<tr>
<td><strong>CPT session 2</strong></td>
<td>0.82 (3.6)</td>
<td>0.94 (0.85–0.97)</td>
<td>90.2</td>
<td>Fair</td>
</tr>
<tr>
<td><strong>HPT session 1</strong></td>
<td>0.48 (2.6)</td>
<td>0.52 (0.14–0.87)</td>
<td>5.30</td>
<td>Good</td>
</tr>
<tr>
<td><strong>HPT session 2</strong></td>
<td>0.05 (1.3)</td>
<td>0.86 (0.69–0.94)</td>
<td>5.70</td>
<td>Good</td>
</tr>
</tbody>
</table>

CDT, cold detection threshold; WDT, warm detection threshold; CPT, cold pain threshold; MID, mean intraindividual difference (i.e., mean difference between participants’ individual scores); SD, standard deviation of intraindividual differences; ICC, intraclass correlation coefficients; CI, confidence interval; CV, coefficient of variation (%). Intra- and interrater reliability for all continuous variables: ICC < 0.4 considered poor agreement; 0.4–0.59 fair agreement; 0.6–0.75 good agreement; and >0.75 excellent agreement.36 CV% should be interpreted in the context of the overall data. Lower CV percentages indicate less variance between the data. CV values < 10% have been suggested as acceptable.29

consistency, and this is reflected in the fact that the magnitude of the ICC depends on between-subject variability.31 For example, when the actual data set has a narrow range of values, that is, when subjects differ little from one another, ICC values can appear very low and vice versa.31,32 This appears to be the case for CDT and WDT.

In relation to CPT, the ICC values in our study for both intra- and interrater reliability were excellent, but, given the large CVs and SDs found for CPT, ICC values may overestimate this measure’s actual reliability for reasons previously discussed.31,32 Thus, CPT could be considered a less robust measure in repeated measures comparisons or for clinical use. HPT data demonstrated poor to fair ICC values for intrarater reliability, but fair to excellent ICC values for interrater reliability. CV values for HPT were relatively low across all tests. Despite wide-ranging data, as compared with detection thresholds, the SD and mean intraindividual differences were quite small for HPT. The combination of results could be interpreted as indicating good intra- and interrater reliability.

There is significant disparity in the literature regarding reliability of thermal QST, as highlighted in an earlier review by Chong and Cros.12 The variation in statistical tests used and lack of standard reporting of estimates of reliability and precision makes interpretation of study findings difficult. Furthermore, a review of the literature (submitted for publication) highlighted only one study that appropriately described blinding21 and this was for the intraoral area, whereas randomization or control of sequence of testing varied enormously. Many studies have reported poor to fair reliability in relation to thermal detection.15,23,31–36 Conversely, there have been a number of studies citing good to excellent reliability of thermal detection thresholds.17,18,25,37–40 Thermal pain thresholds have been investigated to a lesser extent than detection thresholds and, again, reliability results ranged from poor26 to fair38 to good.25

Some recent reliability studies reported findings similar to ours, with ICC values varying considerably, whereas mean differences tended to be small for detection thresholds but larger for pain thresholds. Agostinio et al.14 investigated the reproducibility of thermal QST in the palmar thenar area. They found that CDT and WDT ICC values for intrarater reliability ranged from 0.43 to 0.62 and 0.41 to 0.49, respectively, whereas CPT and HPT ICC ranged from 0.6 to 0.62 and 0.51 to 0.55, respectively. Further examination of their results reveals small mean differences and ranges for detection thresholds, compared with CPT and HPT, which had wider ranges. Their findings are similar to those in our study. Wasner and Brock22 also investigated thermal QST over the dorsum of the hand and reported good–excellent ICC values for intrarater reliability for CPT and HPT. Again, analysis of the range of data revealed wide ranges for CPT and medium ranges for HPT, although small intraindividual variations were reported. In a more recent study, Pigg et al.21 examined the intra- and interrater reliability of intraoral thermal QST. For interrater reliability, they reported ICC values ranging from 0.21 to 0.61 for CDT, 0.13 to 0.65 for WDT, 0.44 to 0.91 for CPT, and 0.58 to 0.87 for HPT. Intrarater ICC values were slightly better for detection thresholds and slightly worse for pain thresholds in comparison to those from this study. However, Pigg et al.21 also reported low mean intraindividual differences both within and between examiners, reflecting better reliability of thermal QST than ICC values might reflect.

It is of interest to note that many studies of the reliability of thermal QST did not provide detailed information regarding the investigators and their training.14–25,30–38 The current investigators were considered to have sufficient training to ensure competency in QST testing. However, one
investigator had significantly more experience than the other. Given the lack of information in previous studies, the level of experience required to ensure good reliability is currently not known. The influence of extent of examiner experience (above competency) on reliability of QST requires further investigation. This is particularly important as QST is being used more readily outside the neuropsychology domain.

The results from this study indicate that CPT and WDT demonstrate good reliability, and their use, particularly in a research setting, may be warranted. HPT also demonstrated good reliability, although CPT was found to be less reliable and further reliability studies are warranted. The levels of reliability required for clinical utility is debatable but should be higher than that considered for research purposes. Similar studies are warranted in patient populations.

Limitations. The investigators acknowledge that the small sample size in this study may have impacted estimates of reliability. In addition, participants were asymptomatic, which limits generalization of the results to other populations. A potential further limitation is the level of training of the two examiners. Although the examiners were trained to be competent in QST evaluation, one examiner had less experience than the other.

In conclusion, in this study we have investigated the reproducibility of thermal QST over the dorsal of the hand in healthy young adults. Cold detection and warm detection thresholds showed good intra- and interrater reliability despite low ICC values. ICC values for cold pain thresholds were high but the large CVs and wide ranges in the data collected are limitations to the reliability of our results. Heat pain thresholds showed good intra- and interrater reliability.

The authors thank Dr. Catherine Blake and Dr. Caroline O’Kelly for statistical support, and Dr. Sean Connolly for technical QST advice. This study was funded by a PhD scholarship from the Irish Research Council for Science Engineering and Technology. These findings were presented at the Research, Therapy and Rehabilitation Society Conference, May 2010, Dublin, Ireland, and the NOI Conference, April 2010, Nottingham, UK.

REFERENCES