Reliability of thermal quantitative sensory testing: A systematic review

Niamh A. Moloney, MSc;1* Toby M. Hall, PhD;2 Catherine M. Doody, PhD1

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

Abstract—The use of quantitative sensory testing (QST) has become more widespread, with increasing focus on describing somatosensory profiles and pain mechanisms. However, the reliability of thermal QST has yet to be established. We systematically searched the literature using key medical databases. Independent reviewers evaluated reliability data using the Quality Appraisal for Reliability Studies checklist. Of the 21 studies we included in this review, we deemed 5 to have high methodological quality. Narrative analysis revealed that estimates of reliability varied considerably, but overall, the reliability of cold and warm detection thresholds ranged from poor to excellent, while heat and cold pain thresholds ranged from fair to excellent. The methodological quality of research investigating the reliability of thermal QST warrants improvement, particularly in terms of appropriate blinding. The results from this review showed considerable variability in the reliability of each thermal QST parameter.

Key words: cold detection threshold, cold pain threshold, detection thresholds, heat pain threshold, neurophysiology, pain thresholds, Quality Appraisal for Reliability Studies, quantitative sensory testing, reliability, warm detection threshold.

INTRODUCTION

In recent years, understanding pain mechanisms among patient populations has become a key focus of many clinical and research groups. In conjunction with this, quantitative sensory testing (QST) has seen increasing use in areas such as musculoskeletal and neuropathic pain for profiling somatosensory phenotypes [1–5] and as an outcome measure in intervention studies [6–7]. Profiling patients using QST involves analyzing multiple parameters of sensory testing to determine whether patients demonstrate dominant features of sensory deficit or sensory hyperexcitability [5,8]. It is thought that this will further the understanding of pain mechanisms and the development or application of more appropriate interventions [8].

QST is a psychophysical means of assessing the function of small and large diameter nerve fibers and their respective pathways [9]. A number of different modalities can be assessed using QST, including vibration, pressure pain thresholds, and thermal thresholds. Thermal thresholds include cold detection threshold (CDT), warm detection threshold (WDT), cold pain threshold (CPT), and heat pain threshold (HPT) [9]. As a psychophysical test, QST is not objective, and consistency in QST data relies heavily on environmental factors, such as ambient temperature and noise; methodological factors, such as test protocol, test application, and test instructions; and the

Abbreviations: CDT = cold detection threshold, CPT = cold pain threshold, CV = coefficient of variation, DFNS = German Research Network on Neuropathic Pain, HPT = heat pain threshold, ICC = intraclass correlation coefficient, MLE = method of levels, MLI = method of limits, QAREL = Quality Appraisal for Reliability Studies, QST = quantitative sensory testing, WDT = warm detection threshold.

*Address all correspondence to Niamh A. Moloney, MSc; School of Public Health, Physiotherapy, and Population Science, University College Dublin, Belfield, Dublin 4, Ireland; +353-86-8106336; fax: +353-1-296-0647.
Email: n_moloney@yahoo.com
http://dx.doi.org/10.1682/JRRD.2011.03.0044
The two primary methods employed in the assessment of thermal QST are the method of limits (MLI) and the method of levels (MLE). The MLI is a reaction-time inclusive method, whereby the applied stimulus increases gradually at a preset rate from the baseline temperature. Participants are then asked to depress a switch when they (1) perceive a change in temperature for detection thresholds or (2) perceive the sensation as painful for pain thresholds. The MLE, sometimes referred to as the forced-choice method, is a reaction-time exclusive method. A set temperature is applied, and the participant is requested to give a “yes” or “no” response on whether or not he or she perceived the sensation. If the participant answers yes, then the temperature is reduced; if he or she answers no, the temperature is increased. This procedure is repeated until the threshold is identified. The staircase method is a variation of the MLI [11].

For any measure to be clinically useful or sufficiently robust for research purposes, it must be reliable [12]. Reliability refers to the consistency of a measurement across time, patients, or observers and the degree to which measurements are free from error [12]. Adequate reliability of a measurement is imperative for clinical decision-making [13]. Reliability of QST also has important consequences for accurate patient profiling. However, a previous literature review of reliability in QST found notable variability in methodology, statistical analyses, and results among the reviewed studies [10]. The use of QST and the body of work in relation to reliability of thermal QST has grown substantially since Chong and Cros’ 2004 review [10]. Therefore, the aim of this article is to systematically review the literature (from January 1990 to May 2010) to determine the level of reliability in thermal QST.

METHODS

Search and Selection

We developed an electronic search strategy through author consensus with a medical librarian and performed it within the following databases: PubMed, Embase, CINAHL (Cumulative Index to Nursing and Allied Health Literature), Web of Science, Science Direct, and Cochrane Library Reviews (covering the period from January 1990 to May 2010). Where possible, we used key words to identify relevant MeSH (medical subject headings) that we then exploded. To gain a list of potentially relevant papers, we combined the QST key words using “or.” We repeated this strategy for the reliability key words. To identify papers on reliability in thermal QST, we combined the two groups of key words using “and.” Table 1 displays the search strategies used for PubMed and Embase. We adapted the search for the other databases using combinations of the search terms outlined in Table 1. We subsequently hand-searched reference lists from retrieved articles for supplementary

<table>
<thead>
<tr>
<th>Phase</th>
<th>MeSH Terms</th>
<th>Emtree Terms</th>
<th>Additional General Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Specific Search Terms for Thermal QST.</td>
<td>Electrophysiology, Neurophysiology, Sensory threshold, Pain threshold, Pain receptors, Pain assessment, Hypesthesia, Reduced/impaired sensation, Thermal hypesthesia</td>
<td>Electroneurology, Sensory system electrophysiology, Perceptive threshold, Pain threshold, Thermal stimulation</td>
<td>QST, Thermal QST, Thermal pain thresholds, Thermal detection thresholds, Psychophysical testing, Sensory testing, Thermal detection, Thermal pain, Pain detection</td>
</tr>
<tr>
<td>3. Combination of Phases 1 and 2.</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

MeSH = medical subject heading, QST = quantitative sensory testing.
studies. Articles were eligible for the review if they fulfilled the criteria outlined in Figure 1.

Reviewer 1 (Ms. Moloney) initially screened titles of articles and article abstracts and only included those that mentioned reliability and/or reproducibility of thermal QST in the review. Two reviewers (Ms. Moloney and Dr. Doody) independently reviewed titles and abstracts to ensure that articles met the inclusion criteria. Where uncertainty arose regarding the eligibility of an article from its abstract, we retrieved the full-text version of the article and evaluated it against the inclusion criteria. We discussed disagreements and achieved consensus for all articles to be included. We then retrieved full-text versions of the studies to be included in the review for quality assessment and data extraction. The reviewers consisted of two physiotherapists (one PhD student and one college lecturer), each with at least 12 years of postgraduate experience.

Quality Assessment

We used a recently devised data extraction form, the Quality Appraisal for Reliability Studies (QAREL) [14], to extract and record data. We then completed a QAREL checklist to facilitate a quality appraisal of the studies using the guidelines suggested by Lucas et al. [14] (Figure 2 and Table 2). Using the standard of Van Trijffel et al. [15], we considered studies to be of high quality if they received a yes score on at least 50 percent of relevant checklist items. As there are 11 items on the checklist, we required a minimum of six yes answers for the study to achieve high quality status. Alternatively, if we considered some checklist items inapplicable for that type of study, we required a yes score on at least 50 percent of the remaining relevant items. The two primary reviewers conducted quality assessment independently using the QAREL checklist. They discussed disagreements, and in all cases, reached a consensus. We obtained statistical advice from a biostatistician regarding the appropriateness of the last item on the QAREL checklist, i.e., statistical measures and their interpretation.

Data Extraction

The two reviewers independently extracted data from the original studies using the QAREL data extraction form, which includes publication details, type of study, subject and observer characteristics, inclusion and exclusion criteria, blinding, randomization, considerations of stability of measure, interpretation of data, and statistical analysis methods [14]. In addition, we extracted and assessed specific data pertaining to the methodological issues, e.g., control of environmental factors.

QST output, measured in degrees Celsius, is continuous in nature; thus, intraclass correlation coefficients (ICCs) and coefficients of variation (CVs) are the most

---

**Inclusion criteria—**
- Experimental studies that assessed thermal QST for intra-rater, inter-rater, and test-retest reliability.
- Experimental studies that compared thermal QST with other methods of assessment but with reliability of thermal QST as stated aim and/or objective.
- Experimental studies involving at least two separate test sessions.
- Studies investigating both nondisabled and/or patient populations.
- Articles published in English between 1990 and 2010.
- Studies with adult participants (>18 yr).

**Exclusion criteria—**
- Studies that did not have reliability as stated primary or secondary aim or objective.
- Studies not involving at least two separate test sessions.
- Studies that described methods and statistical analysis insufficiently to allow adequate analysis.
- Studies that involved manipulation of test scenario, e.g., simulation of results.
- Letters, editorials, or comments.

**Figure 1.**
Article selection criteria. QST = quantitative sensory testing.
commonly used estimates of reliability. In the absence of an accepted standard for the qualitative interpretation of ICC values, we used the interpretation of ICC values by Shrout and Fleiss [16], whereby <0.4 is considered poor agreement, 0.40 to 0.59 is fair, 0.60 to 0.75 is good, and >0.75 is excellent. Note that reliability estimates such as the ICC can be difficult to interpret in the context of an individual score, and as such, an estimate of precision (e.g., standard error of measurement) is important for judging about the degree that measurements vary for an individual [14,17–18].

RESULTS

Search Strategy Yield

The initial search yielded 2,214 references, of which reviewer 1 removed 2,124 irrelevant and duplicate articles (Figure 3). Both reviewers reviewed the titles and abstracts of the remaining 90 articles. Of these, we dismissed 63 articles based on the inclusion and exclusion criteria. We retrieved a further 8 articles after hand-searching the remaining 27 articles. Both reviewers reviewed a total of 35 articles in full. Following the review of the complete articles, we excluded a further 14 articles because they did not meet the inclusion criteria. Consequently, we included 21 total articles in the review.

Quality Assessment and Data Extraction

Of the 21 studies included in the review, we deemed that only 5 studies illustrated high quality using the criteria outlined [19–23]. On closer examination of the QAREL checklist results (Table 2), it is clear that the majority of studies investigated asymptomatic cohorts and few studies described the examiners, which limits the external validity of these studies. Furthermore, recruitment strategies for both study subjects and examiners were poorly outlined. With respect to internal validity, details concerning the blinding of examiners and randomization of the test procedures were the main weaknesses, with the majority of scores interpreted as “unclear.” Of the five studies that we deemed high quality, one study presented statistics that we did not deem comprehensive [19], i.e., ICC values only, without measures of precision or sufficient raw data [14].

Narrative Analysis

For the purpose of this review, we present a narrative analysis of the results. Meta-analysis was not possible because of the variation in study quality and statistical methods used across studies. We reviewed the various aspects of the studies under three sections: (1) type of study, sample, and raters; (2) methodological issues; (3) statistical analysis; and (4) results. Table 3 presents details of methods, statistical analyses, and results.
<table>
<thead>
<tr>
<th>Study</th>
<th>1</th>
<th>2</th>
<th>3*</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>High Quality†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agostinho et al. (2009) [1]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear/NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Becser et al. (1998) [2]</td>
<td>No</td>
<td>Unclear</td>
<td>Unclear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Yes</td>
<td>No (fixed)</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Bird et al. (2006) [3]</td>
<td>Yes</td>
<td>No</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Bravenboer et al. (1992) [4]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear/NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Claus et al. (1990) [5]</td>
<td>No</td>
<td>Unclear</td>
<td>Unclear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No†</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Claus et al. (1993) [6]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear/NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>De Neeling et al. (1994) [7]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear/NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Dyck et al. (1991) [8]</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>No†</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Felix and Widerström-Noga (2009) [9]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear/NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Gelber et al. (1995) [10]</td>
<td>No</td>
<td>Unclear</td>
<td>Unclear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Heldestad et al. (2010) [11]</td>
<td>No</td>
<td>Unclear</td>
<td>Unclear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Kemler et al. (2000) [12]</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>No (fixed)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Krassioukov et al. (1999) [13]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear</td>
<td>NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Moravcová et al. (2005) [14]</td>
<td>Yes</td>
<td>Unclear</td>
<td>NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>No†</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Peltier et al. (2009) [15]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>No (fixed)</td>
<td>Unclear</td>
<td>Yes</td>
<td>No†</td>
<td>No</td>
</tr>
<tr>
<td>Pigg et al. (2010) [16]</td>
<td>No</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Valensi et al. (1993) [17]</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>No†</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Wasner and Brock (2008) [18]</td>
<td>No</td>
<td>Unclear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>No†</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Yarnitsky and Sprecher (1994) [19]</td>
<td>No</td>
<td>Unclear</td>
<td>NA</td>
<td>Unclear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yarnitsky et al. (1995) [20]</td>
<td>No</td>
<td>Unclear</td>
<td>NA</td>
<td>Unclear</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Zwart and Sand (2002) [21]</td>
<td>Yes</td>
<td>Yes</td>
<td>NA</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table 2. (cont).
Results from Quality Appraisal for Reliability Studies (QAREL) checklist.

Type of Study, Sample, and Raters

Eight studies investigated intrarater reliability [11,21–22,24–28]. Of those, three also assessed interrater reliability [22,24,26]. Six studies investigated test-retest reliability [23,29–33], but the remaining seven studies did not specify which type of reliability they intended to assess [25,28,34–38]. Of the 21 studies we reviewed, 11 provided details of the raters. In most studies, details of the raters is limited to “the authors of the papers” [23–24,27,31] or brief information such as “a single technician or one of a number of observers” [11,28,32–33]. Of the three remaining studies, the raters were (1) individuals trained and certified by the Central Reading and Coordinating Center, Department of Neurology, University of Pennsylvania (Philadelphia, Pennsylvania) [34]; (2) neurologists [19]; and (3) authors trained by the German Research Network on Neuropathic Pain (DFNS) [22].

Of the 21 articles we reviewed, 14 investigated non-disabled populations, 7 investigated people with diabetes with and without associated neuropathy [19,28,30,33–36], 2 investigated people with spinal cord injuries [29,37] (one with neuropathic pain [29]), and 3 investigated people with diffuse pain syndromes and/or musculoskeletal disorders [21,23,38].
Environmental factors and instructions. Environmental factors reported in the studies we reviewed included standardizing room temperature, controlling noise and distractions, and recording skin temperature. Eight of the studies described how they controlled environmental factors [11,21,24,26,29,31–33], while two studies stated that they used the protocol described by the DFNS [29–30]. The remainder did not provide either sufficient or any details about environmental factors. Regarding instructions, 10 studies either described their instructions or stated that they used standardized instructions [20–22,24,26–27,30–32,38]; the remaining 11 did not specify.

Blinding. Blinding may incorporate blinding of testers to their previous results, to the results of other testers, and to clinical information that may influence their testing. Surprisingly, 18 of the 21 studies did not specify any details relating to blinding within their design. In the remaining studies, aspects of blinding were conducted in two studies [19,22] while the remaining study by Pigg et al. [22] was the only article we deemed to have reported appropriate blinding.

Randomization. The testing order was randomized in four studies and fixed in a further four studies that described the order. The remaining studies did not describe whether they controlled the order of testing. The sequence of examiner was randomized in one study and fixed in another. It was not relevant in two studies and not known whether relevant or not in the remainder of the studies, because the type of reliability study being performed was not clearly stated. The side to be tested was randomly assigned in five studies and not discussed in the remainder of the articles.

Statistical Analysis
We found large variation in the statistical methods used to analyze the data in these studies. Lucas et al. have recommended that appropriate statistical analysis should utilize a measure of reliability (e.g., ICC) as well as a measure of precision or stability (e.g., 95% confidence intervals or standard error of measurement) [14]. If studies achieved these criteria, a yes score was given on QAREL checklist item 11. We categorized six studies as using appropriate statistical analysis but with insufficient detail provided to truly determine reliability [19,25,27–28,30–31]; for example, measures of precision were not provided in conjunction with the reliability, or ICC or insufficient actual raw data were provided [19,30–31]. The ICC was the most common estimate of reliability.
### Table 3.
Testing methods, statistical analysis, and results found in electronic search results.

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Reliability/Interval</th>
<th>Subjects/Tests</th>
<th>Modality/Equipment</th>
<th>Environmental Factors/Instructions</th>
<th>Measure of Repeatability</th>
<th>Measure of Precision/Stability</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agostinho et al. (2009)</td>
<td>Test-retest¹</td>
<td>36 nondisabled; 36 non-neuropathic pain/Not specified</td>
<td>CPT, HPT/TSA-II Neuro Sensory Analyzer²</td>
<td>No details on environmental factors/Standardized protocol developed by DFNS that includes verbal instructions</td>
<td>ANOVA</td>
<td>Analysis of absolute variables/Bland-Altman analysis</td>
<td>No significant systematic difference between days for WDT and HPT; systematic difference between days for CDT and CPT. Bland-Altman analysis: No significant difference from baseline for CDT and CPT. r-Values— All: CDT = 0.54, WDT = 0.44, CPT = 0.61, HPT = 0.52. Nondisabled: CDT = 0.43, WDT = 0.49, CPT = 0.62, HPT = 0.51. Patients: CDT = 0.62, WDT = 0.41, CPT = 0.60, HPT = 0.55.</td>
<td>No significant differences between r-values for patients or nondisabled controls. Systematic difference between days for CDT and CPT but differences in absolute values is small. Conclusions— CDT = good r but systematic difference between days indicates fair reliability. WDT = fair reliability. CPT = good r but systematic difference between days indicates fair reliability. HPT = Fair reliability.</td>
</tr>
<tr>
<td>Becser et al. (1998) [2]</td>
<td>Intrarater and inter-rater/Within 7 day limit</td>
<td>20 nondisabled/ Bector and Zwart (no other details)</td>
<td>CDT, WDT/ Somedic thermodest equipment³</td>
<td>Quiet room 22°C–23°C/ Brief details on instructions provided</td>
<td>CR; Bland-Altman analysis, presented as °C, ICC (one-way random effects model)</td>
<td>95% reference limits (upper): CDT = 0.63, WDT = 0.66. 95% reference limits (lower): CDT = 0.35–0.59, WDT = 0.45–0.69. CR average (°C): CDT = 0.53, WDT = 0.60.</td>
<td>Intrarater reliability— ICC average: CDT = 0.63, WDT = 0.66. CR average (°C): CDT = 0.53, WDT = 0.60.</td>
<td>Intrarater reliability— CDT and WDT ICC values: Good. CR: moderate. Interrater reliability— ICC values: Good. Small but significant difference for CDT. Conclusions— Intrarater: CDT and WDT = fair reliability. Interrater: CDT = fair reliability. WDT = good reliability but limited information.</td>
</tr>
<tr>
<td>Claus et al. (1993) [3]</td>
<td>Test-retest/ 2 consecutive days within 1 week</td>
<td>30 nondisabled; 12 diabetes/ Not specified</td>
<td>CDT, WDT/ Modified Marstock thermode⁴</td>
<td>Not specified</td>
<td>CR; Bland-Altman analysis, presented as °C, ICC (one-way random effects model)</td>
<td>95% reference limits (upper): CDT = 0.77, WDT = 0.77. 95% reference limits (lower): CDT = 0.32, WDT = 0.32. CR average (°C): CDT = 1.17, WDT = 1.45.</td>
<td>Intrarater reliability— ICC average: CDT = 0.65, WDT = 0.72. CR average (°C): CDT = 1.04, WDT = 1.65.</td>
<td>Intrarater reliability— CDT and WDT ICC values: Good. CR: moderate. Interrater reliability— ICC values: Good. Small but significant difference for CDT. Conclusions— Intrarater: CDT and WDT = fair reliability. Interrater: CDT = fair reliability. WDT = good reliability but limited information.</td>
</tr>
<tr>
<td>Felix and Widerström (2009)</td>
<td>Test-retest/ 1 week</td>
<td>10 SCI and neuropathic pain; 10 non-disabled/Not specified</td>
<td>CPT, HPT/TSA-II Neuro Sensory Analyzer²</td>
<td>Quiet room with temperature controlled; skin temperature recorded/No details on instructions</td>
<td>ICC (one-way random effects model)</td>
<td>ICC (95% CI)— SCI: CDT = 0.00 (0.83–0.94), WDT = 0.95 (0.91–0.95), CPT = 0.50 (0.28–0.67), HPT = 0.50 (0.28–0.66). Nondisabled: CDT = 0.68 (0.54–0.78), WDT = 0.70 (0.57–0.80), CPT = 0.49 (0.31–0.64), HPT = 0.68 (0.55–0.79).</td>
<td>CDT demonstrates good correlation coefficient but high variability and day to day differences. WDT demonstrates excellent correlation coefficient, moderate difference, and small day to day difference. Conclusions— CDT = poor reliability. WDT = good reliability.</td>
<td></td>
</tr>
</tbody>
</table>

1. Test-retest: repeated testing at the same time interval.
2. r-Values: significance of test-retest.
3. CR: comprehensive reliability.
4. ICC: interclass correlation coefficient.

Note: The table provides a summary of testing methods, statistical analysis, and results found in electronic search results. The studies are referenced as follows: 1. Noga (2009); 2. Widerström-Felix and Claus et al. (1993); 3. Becser et al. (1998); 4. Felix and Noga (2009).
### Table 3. (cont).

Testing methods, statistical analysis, and results found in electronic search results.

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Reliability/Interval</th>
<th>Subjects/Equipment</th>
<th>Method of Limits</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heldestad et al. (2010) [5]</td>
<td>Test-retest/Between 1–8 days</td>
<td>Nondisabled—Not specified</td>
<td>CDT, WDT, CPT, HPT, Somedic thermostest equipment*</td>
<td>Inter- and intra-subject reproducibility; Analyses of inter- and intra-individual differences</td>
<td>No difference between repeated testing on days 1, 2, and 7. Repeatability between measurements within participants (as first test), mean CR (°C)—First test: CDT = 1.00, WDT = 1.06, CPT = 6.50, HPT = 5.99. After thermal pain assessment: CDT = 2.92, WDT = 2.08. Intraindividual variation for first test (CV %)—Absolute values (mean): CDT = 3.10, WDT = 1.80, CPT = 0.63, HPT = 1.60. A (%): CDT = 65.80, WDT = 35.30, CPT = 0.29, HPT = 5.80. Within days repeatability (CV%) absolute values—CPT = 0.89–6.07 (first test), WDT = 0.33–4.10 (first test), HPT = 0.63–8.10.</td>
</tr>
<tr>
<td>Kra-siukov et al. (1999) [6]</td>
<td>Test-retest/SCI: 3 weeks; Nondisabled controls: 1 week</td>
<td>21 SCI; 14 nondisabled/Not specified</td>
<td>CDT, WDT, CPT/TSA-II Neuro-Sensory Analyzer*</td>
<td>ICC SD; CV</td>
<td>Nondisabled—ICC: CDT = 0.75–0.90, WDT = 0.36–0.84, CPT = 0.91–0.95. CV%: CDT = 4.00–10.80, WDT = 1.20–3.80, CPT = 56.30–100.30. SCI—ICC: CDT = 0.45–0.81, WDT = 0.23–0.69, CPT = 0.65–0.89. CV%: CDT = 42.60–75.50, WDT = 7.10–12.00, CPT = 72.7–139.30.</td>
</tr>
<tr>
<td>Pigg et al. (2010) [7]</td>
<td>Intra- and interrater/Twice day 1 by 2 examiners, again 1–2 weeks later by 1 examiner</td>
<td>21 nondisabled/2 authors trained by DFNS</td>
<td>CDT, WDT, CPT, HPT, MSA Thermal Stimulator*</td>
<td>ICC Mean ± SD/MID</td>
<td>Intrarater reliability—ICC: CDT = 0.45–0.77, WDT = 0.23–0.67, CPT = 0.55–0.87, HPT = 0.64–0.80. MID: CDT = 0.40–4.70, WDT = 0.50–2.40, CPT = 2.20–4.20, HPT = 1.20–2.30. Interrater reliability—ICC: CDT = 0.21–0.61, WDT = 0.13–0.65, CPT = 0.44–0.91, HPT = 0.58–0.87. MID: CDT = 1.20–5.70, WDT = 0.90–2.40, CPT = 2.00–4.60, HPT = 1.40–1.70.</td>
</tr>
</tbody>
</table>
### Conclusions

ICC (Bland-Altman) No CR; ICC for between variability.

Reliability coefficient (Rtt)

Confidence

Testing methods, statistical analysis, and results found in electronic search results.

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Interval</th>
<th>Subjects/</th>
<th>Modality/ Equipment</th>
<th>Environmental Factors/ Instructions</th>
<th>Measure of Repeatability</th>
<th>Measure of Precision/ Stability</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wasner and Brock (2008) [8]</td>
<td>Test-retest/ 3 times over 3 weeks (days 0,1, and 21)</td>
<td>20 nondisabled/ Wasner</td>
<td>CPT, HPT/ TSA-II Neuro Sensory Analyzer2</td>
<td>Room held at 22°C–23°C with relative humidity of 50%–60% Used DFNS protocol instructions; no information on noise control</td>
<td>ICC</td>
<td>No</td>
<td>ICC (r-value) — Day 0 vs day 1: CPT = 0.948, HPT = 0.648. Day 0 vs day 21: CPT = 0.781, HPT = 0.887.</td>
<td>Conclusions — CPT and HPT = good to excellent reliability but limited information.</td>
</tr>
<tr>
<td>Zwart and Sand (2002) [9]</td>
<td>Test-retest/ Tested twice 1 to 2 hours between tests</td>
<td>19 lumbosacral radiculopathy/ Zwart</td>
<td>CDT, WDT/ Tsa-II some medical thermostest equipment</td>
<td>Not specified</td>
<td>CR; ICC for between variation, repeated measures of ANOVA</td>
<td>Not applicable</td>
<td>Conclusions — Mixed results with ICC values varying from poor to excellent. Coefficients of repeatability were high throughout. Conclusions — CDT = fair reliability, WDT = fair reliability.</td>
<td></td>
</tr>
</tbody>
</table>

### Mixed Method of Limits and Method of Levels

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Interval</th>
<th>Subjects/</th>
<th>Modality/ Equipment</th>
<th>Environmental Factors/ Instructions</th>
<th>Measure of Repeatability</th>
<th>Measure of Precision/ Stability</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Claus et al. (1990) [10]</td>
<td>Test-retest/3 consecutive days</td>
<td>55 nondisabled/Not specified</td>
<td>CDT, WDT/ Modified Marstock, thermode2 (MLI, MLE)</td>
<td>Not specified</td>
<td>Reliability coefficient (Rtt)</td>
<td>Confidence limits mentioned but values not provided</td>
<td>MLI— Rtt 1/2: WDT = 0.73, CDT = 0.71. Rtt 2/3: WDT = 0.83, CDT = 0.71. MLE— Rtt 1/2: WDT = 0.82, CDT = 0.82. Rtt 2/3: WDT = 0.78, CDT = 0.78.</td>
<td>Conclusions — MLI— CR (%): CDT = 0.40–0.83, WDT = 0.35–0.67. Asymptomatic side — CR (%): CDT = 34–52, WDT = 40–65. ICC: CDT = 0.27–0.86, WDT = 0.43–0.82.</td>
</tr>
<tr>
<td>Kemler et al. (2000) [11]</td>
<td>Intrarater/1 month</td>
<td>53 CRPS/ Not specified</td>
<td>CDT, WDT/ TSA-II Neuro Sensory Analyzer2 (MLE vs MLI)</td>
<td>Temperature-controlled laboratory (22°C–24°C); no visual access to computer; no visual/auditory cues/ Instructions described</td>
<td>CR</td>
<td>Bland-Altman analysis</td>
<td>CR — MLE: CDT unaffacted wrist = 0.8, CDT affected wrist = 0.7, CDT unaffected foot = 4.1, CDT affected foot = 5.8, WDT unaffected wrist = 1.0, WDT affected wrist = 2.0, WDT unaffected foot = 5.8, WDT affected foot = 4.0. MLI: CDT unaffected wrist = 2.3, CDT affected wrist = 3.7, CDT unaffected foot = 5.3, CDT affected foot = 3.4, WDT unaffected wrist = 1.7, WDT affected wrist = 5.0, WDT unaffected foot = 2.9, WDT affected foot = 4.4.</td>
<td>Conclusions — All measures demonstrated poor reliability at foot. MLE CDT and WDT = good reliability at wrist. MLI CDT and WDT = poor reliability at all sites except unaffected wrist.</td>
</tr>
<tr>
<td>Moravcová et al. (2005) [12]</td>
<td>Intrarater/ Twice over 1 week</td>
<td>58 small-fiber neuropathy; 30 nondisabled/Moravcová</td>
<td>CDT, WDT/ Nicolet IV Viking IV electrodiagnostic unit3, TSA-II NeuroSensory Analyzer2 (MLI random and nonrandom), (MLE)</td>
<td>Protocol description for thermal QST very brief; authors state that “conditions were standardized” but detail insufficient/ Standardized instructions used</td>
<td>CR</td>
<td>No</td>
<td>Thescan cold— Nondisabled: MLI nonrandom = 1.06, MLI random = 0.71, MLE = 0.48. Patients: MLI nonrandom = 2.18, MLI random = 1.40, MLE = 1.22. Thescan warm— Nondisabled: MLI nonrandom = 0.76, MLI random = 0.72, MLE = 0.54. Patients: MLI nonrandom = 1.38, MLI random = 1.56, MLE = 1.24.</td>
<td>Reliability better for MLE than MLI. Reliability better for patient group than nondisabled participants. Conclusions — CDT and WDT = good reliability but limited information.</td>
</tr>
</tbody>
</table>
Table 3. (cont).

Testing methods, statistical analysis, and results found in electronic search results.

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Reliability/Interval</th>
<th>Subjects/Tests</th>
<th>Modality/Equipment</th>
<th>Environmental Factors/Instructions</th>
<th>Measure of Repeatability</th>
<th>Measure of Precision/Stability</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yarnitsky and Sprecher (1994) [13]</td>
<td>Intrarater/2 weeks</td>
<td>106 nondisabled/Not specified (&quot;single technician&quot;)</td>
<td>CDT, WDT/ TSD-II NeuroSensor Analyzer² (MLI, MLE, SC)</td>
<td>Soundproof air-conditioned room with distractions minimized/Standard instructions used</td>
<td>Repeatability r</td>
<td>MISD</td>
<td>r-Value— Thenar: MLE CDT = 1.040, MLE WDT = 0.572, MLI CDT = 1.964, MLI WDT = 1.587, SC CDT = 1.144, SC WDT = 0.720. Foot: MLE CDT = 3.016, MLE WDT = 3.758, MLI CDT = 3.778, MLI WDT = 4.298. Mean intersession difference— Thenar: MLE CDT = –0.086, MLE WDT = –0.006, MLI CDT = 0.419, MLI WDT = 0.249, SC CDT = 0.013, SC WDT = –0.013. Foot: MLE CDT = –0.044, MLE WDT = 0.352, MLI CDT = 0.197, MLI WDT = –0.115.</td>
<td>Intersession bias found for MLI complicates reliability study. Higher r for lower limb correlates with higher threshold values. Conclusions— MLI: CDT and WDT for thenar area = poor reliability, MLE and SC for thenar area: CDT = fair reliability. WDT = good reliability. MLI and MLE for foot area: CDT = fair reliability, WDT = fair reliability.</td>
</tr>
<tr>
<td>Yarnitsky et al. (1995) [14]</td>
<td>Test-retest/2 weeks</td>
<td>72 nondisabled/Not specified (&quot;single technician&quot;)</td>
<td>HPT/TSA-II NeuroSensor Analyzer² (MLI, MLE, SC)</td>
<td>Soundproof air-conditioned room with distractions minimized/Standard instructions used</td>
<td>Repeatability r</td>
<td>MISD</td>
<td>Thenar eminence: r = 5.85. Foot: r = 4.47</td>
<td>Large coefficients of repeatability. MISD data not presented, but authors report “intersession bias” for heat pain at thenar eminence. No specific data provided for MLI, MLE, or SC. Conclusions— HPT = poor reliability but limited information.³</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Reliability/Interval</th>
<th>Subjects/Tests</th>
<th>Modality/Equipment</th>
<th>Environmental Factors/Instructions</th>
<th>Measure of Repeatability</th>
<th>Measure of Precision/Stability</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bird et al. (2006) [15]</td>
<td>Test-retest/3 separate days within 4-week period</td>
<td>1,100 clinically stable diabetes with mild neuropathy/CRCR trained neurologists and technologists</td>
<td>CDT/CASE IV system² (MLE: 4-2-1 stepping algorithm)</td>
<td>Temperature controlled room/No details on instructions or noise</td>
<td>Total variance (SD); ICC</td>
<td>CV</td>
<td>Variance— Total: 20.88. Due to site: 1.15 (6%). Due to patient: 14.41 (69%). Random error: 5.58 (27%). ICC range: 0.68–0.73. CV: 30.22%</td>
<td>Low variance between sites. High intrasubject variation. ICC values = good. CV = 30.22% (moderate variance). Main methodological limitation: Not primarily designed as a reliability study. Conclusions— CDT = fair reliability.</td>
</tr>
<tr>
<td>Bravenboer et al. (1992) [16]</td>
<td>Test-retest/2 weeks</td>
<td>39 diabetes without known neuropathy/Not specified</td>
<td>CDT, WDT/ Triple T Thermal Threshold Tester ²⁶</td>
<td>Not specified</td>
<td>CR Bland-Alman analysis</td>
<td>No</td>
<td>Correlation of reliability— Normal: Warm hand = 0.19, cold hand = 0.17, warm foot = 4.34, cold foot = 0.60. Abnormal: Warm hand = 1.17, cold hand = 1.01, cold foot = 4.69.</td>
<td>Conclusions— CDT and WDT = fair reliability in hand and poor reliability in foot but limited information.³</td>
</tr>
<tr>
<td>De Neeling et al. (1994) [17]</td>
<td>Test-retest/13–24 days</td>
<td>19 nondisabled; 20 with without non-insulin dependent diabetes/Not specified (&quot;one of three observers&quot;)</td>
<td>TDT (combination of CDT and WDT)</td>
<td>Quiet ambience with constant room temperature of 18°C–22°C/No details on instructions</td>
<td>Reliability coefficient 95% CI, SD diff, CV</td>
<td>r (95% CI): 0.54 (0.26–0.73). SD diff (95% CI): 0.49 (0.39–0.61). CV: 0.72.</td>
<td>Fair estimate of reliability with large variance. Conclusions— TDT = poor reliability.</td>
<td></td>
</tr>
</tbody>
</table>
### Table 3. (cont).

Testing methods, statistical analysis, and results found in electronic search results.

<table>
<thead>
<tr>
<th>Study</th>
<th>Type of Reliability/Interval</th>
<th>Subjects/ Testers</th>
<th>Modality/Equipment</th>
<th>Environmental Factors/Instructions</th>
<th>Measure of Repeatability</th>
<th>Measure of Precision/Stability</th>
<th>Results</th>
<th>Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyck et al. (1991) [18]</td>
<td>Intra- and interrater/3-5 days</td>
<td>20 diabetes with and without neuropathy/3 neurologists</td>
<td>CDT, WDT/IV system</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Not specified</td>
<td>ICC</td>
<td>CI on graph but specific measures not provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gelber et al. (1995) [19]</td>
<td>Intra- and interrater/3 test sessions on 3 days (days 1, 2, and 7 for n = 29); 1 test session on 3 days (days 1, 2, and 7 for n = 9)</td>
<td>10 nondisabled for intratester reliability; compared 140 nondisabled at 6 centers/Not specified</td>
<td>CDT/Thermal sensitivity tester</td>
<td>ANOVA linear regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Quiet room free from visual distractions; skin temperature recorded/Standardized instructions used</td>
<td>CV (%)—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peltier et al. (2009) [20]</td>
<td>Test-retest/Twice over 30 days</td>
<td>19 impaired glucose regulation and peripheral neuropathy/Not specified</td>
<td>CDT/CASE IV system (“previoulsy published methodology”)</td>
<td>ICC</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Conditions of the testing were standardized”/Standardized instructions used</td>
<td>ICC—</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valseni et al. (1993) [21]</td>
<td>Intrarater/inter-center/4 weeks</td>
<td>132 diabetes with peripheral neuropathy/1 neurophysiologist in each center</td>
<td>CDT, WDT/Thermal testing system (no additional information)</td>
<td>No information provided/No standards</td>
<td>CV; percentages of total variance</td>
<td>No</td>
<td>Total CV (%)—</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CDT = 0.9, WDT = 0.8. 95% CI: CDT = 0.95–0.99 (approx), WDT = 0.55–0.90 (approx.)</td>
<td>WDT = 64.5, CDT = 116.6. Intercenter variability (%): WDT = 3.9, CDT = 12.5. Intersubject variability (%): WDT = 39.4, CDT = 85.8. Intrasubject variability (%): WDT = 21.2, CDT = 18.3. Large variance for both WDT and CDT. Large inter-subject variability. Small intrasubject or intercenter variability. Conclusions—CDT and WDT = fair reliability but limited information.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Intra- and interrater reliability for all continuous variables. ICC of <0.40 is considered poor; 0.40 to 0.59 fair; 0.60 to 0.75 good; >0.75 excellent agreement [22].

*Method of reliability not stated, therefore assumed to be test-retest.

†Medoc Advance Medical Systems; Ramat Yishai, Israel.

§Somedic AB; Hörby, Sweden.

¶Limited information: Absence of sufficient data or additional measures of precision limits interpretation of estimates of reliability.

**Marstocknervtest; Schriesheim, Germany.

††Nicolet Biomedical; Madison, Wisconsin.

‡‡W. R. Medical Electronics; Stillwater, Minnesota.

§§Medec Ltd; Old Woking, United Kingdom.


Table 3. (cont).

Testing methods, statistical analysis, and results found in electronic search results.

23. MOLONEY et al. Reliability of thermal quantitative sensory testing

Results: Estimates of Reliability

Table 3 presents the actual results and estimates of reliability, which are divided into three sections according to the method used. Given the variability of actual results for all parameters, it is difficult to draw definite conclusions regarding reliability. We found no observable difference in reliability between the MLE and the MLI. Analysis of reliability of the five studies with high methodological quality indicated considerable variability [19–23]. CDT and WDT ranged from poor to excellent (Dyck et al.: excellent [19]; Hedestad et al.: good for CDT and excellent for WDT [20]; Klemmer et al.: good at the wrist site and poor elsewhere for CDT and WDT [21]; Pigg et al.: poor to excellent for CDT, poor to good for WDT [22]; Zwart and Sand: fair for both CDT and WDT, fair to excellent for CPT and HPT [23]; Hedestad et al.: excellent for CPT and good for HPT [20]; Pigg et al.: fair to excellent for both CPT and HPT [22]). When we included the remaining studies, the reliability of CDT ranged from poor to excellent, with the majority of studies indicating fair reliability; WDT ranged from poor to excellent, with the majority of studies indicating fair reliability; CPT was divided between fair and good; and finally, HPT was also divided between fair and good, with one study finding poor reliability.

DISCUSSION

Type of Study, Sample, and Raters

Of the 21 studies we included, only 3 provided adequate information about the raters [19–20,34], while 13 either assessed test-retest reliability or did not state which type of reliability they investigated. This affects the external validity of these studies because the study design does not take into account the possible influence of the rater on the test outcome. Indeed, Becser et al. reported some differences between raters and noted that using more than one rater may introduce bias in a measurement [24]. It has also

used (9 studies) [19,22–24,29–31,34,37]. The coefficient of reliability was the second most frequently used statistical test, with Bland-Altman limits of agreement, CV, and repeatability r being the other measures used.
been previously highlighted that standardized methods of assessment and attention to training are important factors in achieving consistent results in QST [34]. Appropriate training in conducting QST is warranted, and indeed, groups such as the DFNS have developed standardized training days for testers. As such, further information on the raters, their relevant training, and where they conducted the study would allow for better appraisal of the broader applicability of reliability studies.

External validity of the studies we reviewed is also limited by the number of reliability studies that were conducted on nondisabled populations. It appears that reliability in nondisabled participants (n = 14) and people with diabetes (n = 7) have been thoroughly investigated to date. However, this is not the case for musculoskeletal pain, which has been less well investigated despite the growing use of QST in profiling patients with these conditions, e.g., whiplash [1–3], patellofemoral pain syndrome [39], and low back pain [4]. Therefore, further studies on the reliability of QST in populations with musculoskeletal pain are warranted. At least two such studies are underway [40–41].

Methodological Issues

It is clear from this review that the majority of studies favored the use of the MLI (n = 14). This may be partly explained by the greater time involved in using the MLE. Despite the evidence for greater accuracy and reliability with the MLE in two studies [21,27], analysis of all the studies included in this review suggests that the MLI and the MLE demonstrate comparable reliability. In fact, studies using the MLE provided inadequate information regarding analysis and results more frequently (4 out of 7 studies [19,28,30,35]) than those using the MLI alone (2 out of 9 studies [24,31]). Estimates of reliability in studies that provided inadequate information must be interpreted with caution.

Reliability studies of thermal detection thresholds have been explored to a greater extent than thermal pain thresholds, and as such, the reliability of thermal pain thresholds is less well established. Regarding environmental factors and standardization of instructions to subjects, note that these are important components of reliability in QST [10]. Environmental factors may include standardizing room temperature, controlling noise and distractions, and recording skin temperature. With this in mind, studies that controlled environmental factors and issued standardized instructions would be deemed more reliable than those that did not, although this was not clearly reflected in the actual estimates of reliability. Of all methodological factors of importance in a reliability study, blinding appears to be one that has been most poorly described in the studies in this review. Only four of the studies included in this review outlined any blinding procedures, and of those, we only deemed one to demonstrate appropriate blinding according to the QAREL checklist [22]. The importance of blinding in a reliability study is highlighted by Lucas et al. in their development of the QAREL checklist in which 5 of the 11 items in the QAREL checklist pertain to blinding [14].

It is also clear from this review that consideration of the test and examiner sequence varies considerably. Randomization was consistently poorly described throughout the articles reviewed. In this review, only three studies indicate that they randomized their test protocol [20,25,27], while four studies fixed it [21,24,30,37]. The remaining 14 studies did not reference any form of randomization.

We did not find a consistent association between the interval between sessions and the estimate of reliability. Indeed, Wasner and Brock demonstrated better ICC estimates for CPT between days 0 and 1 versus days 0 and 21, but poorer ICC estimates for HPT between days 0 and 1 versus days 0 and 21 [31].

In summary, the main areas of methodology that warrant greater attention in future studies are descriptions of the raters and their training, blinding, and randomization. In addition, standardization of test protocols, environmental factors, and instructions are also important factors to consider. Regarding the study population, future studies of patient populations, particularly those with painful conditions, are warranted so that the subject samples are representative of those who would typically be undergoing QST in clinical or research settings.

Statistical Analysis and Results

The range of statistical measures used in reliability studies limits the ability to perform meta-analysis on the data. Differences in statistical methods probably reflect changing trends within statistics. It has been suggested that ICCs are the most appropriate measures, but only if they are presented in conjunction with a measure of precision, e.g., standard error of measurement [14,17–18]. However, it can be argued that when the data demonstrate large or small variation, the ICC may over- or underestimate reliability, respectively [12]. It has been suggested that the presentation of sufficient data alongside measures
of reliability and precision allows a more accurate analysis of the data by the reader. This has been demonstrated by more recent studies, such as Pigg et al. [22].

As discussed earlier in the “Results” section, no one thermal QST parameter demonstrated consistent estimates of reliability across studies. Therefore, it is difficult to draw definite conclusions about reliability. While at risk of being oversimplistic, the summary of the results suggests that the reliability is fair for CDT and WDT and fair to good for CPT and HPT.

Limitations

We acknowledge a number of limitations to this review. We only included published studies, and as such, results from this review may overestimate reliability. Furthermore, we only included studies published in English. Finally, we only included studies if they were published within the last two decades.

CONCLUSIONS

The methodological quality of research investigating the reliability of thermal QST could adhere to more rigorous guidelines as suggested by the QAREL checklist, particularly in relation to incorporating appropriate blinding procedures into the design. Further studies investigating reliability of QST in populations with pain are warranted. The results of this review found that the reliability of thermal QST varied considerably. CDT and WDT were found to have fair reliability. CPT and HPT demonstrated good reliability in high quality studies and varied from fair to good reliability in the other studies. We found no difference in reliability between studies using the MLI and those using the MLE. The reliability of thermal pain thresholds is less well established than thermal detection thresholds.

ACKNOWLEDGMENTS

Author Contributions:
Study concept and design: N. A. Moloney, T. M. Hall, C. M. Doody.
Acquisition of data: N. A. Moloney.
Analysis and interpretation of data: N. A. Moloney, C. M. Doody.
Drafting of manuscript: N. A. Moloney.
Critical revision of manuscript for important intellectual content: C. M. Doody, T. M. Hall.
Narrative analysis: N. A. Moloney, C. M. Doody, T. M. Hall.
Study supervision: C. M. Doody, T. M. Hall.

Financial Disclosures: The authors have declared that no competing interests exist.
Funding/Support: This material was based on work supported by the Irish Research Council for Science, Engineering and Technology Embark Initiative (grant 1853271).

Additional Contributions: The authors would like to thank Dr. Catherine Blake for her advice regarding the interpretation of statistical methods and results in this review.

REFERENCES


Submitted for publication March 15, 2011. Accepted in revised form July 18, 2011.

This article and any complementary material should be cited as follows: