JRRD Volume 49, Number 2, 2012
Pages 191–208

Reliability of thermal quantitative sensory testing: A systematic review

Niamh A. Moloney, MSc;[1*] Toby M. Hall, PhD;[2] Catherine M. Doody, PhD[1]
[1]School of Public Health, Physiotherapy, and Population Science, University College Dublin, Belfield, Dublin, Ireland;
[2]School 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

cooperation and attention of the individual being tested [10]. 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 1.
Search strategy for PubMed and Embase.

Phase	MeSH Terms	Emtree Terms	Additional General Terms
1. Specific Search Terms for Reliability Studies.	Reliability Reproducibility	Reliability Intratester reliability Intertester reliability Test-retest reliability Reproducibility	Method reliability
2. Specific Search Terms for Thermal QST.	Electrophysiology Neurophysiology Sensory threshold Pain threshold Pain receptors Pain assessment Hypesthesia Reduced/impaired sensation Thermal hypesthesia	Electroneurology Sensory system electrophysiology Perceptive threshold Pain threshold Thermal stimulation	QST Thermal QST Thermal pain thresholds Thermal detection thresholds Psychophysical testing Sensory testing Thermal detection Thermal pain Pain detection
3. Combination of Phases 1 and 2.	—	—	—

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

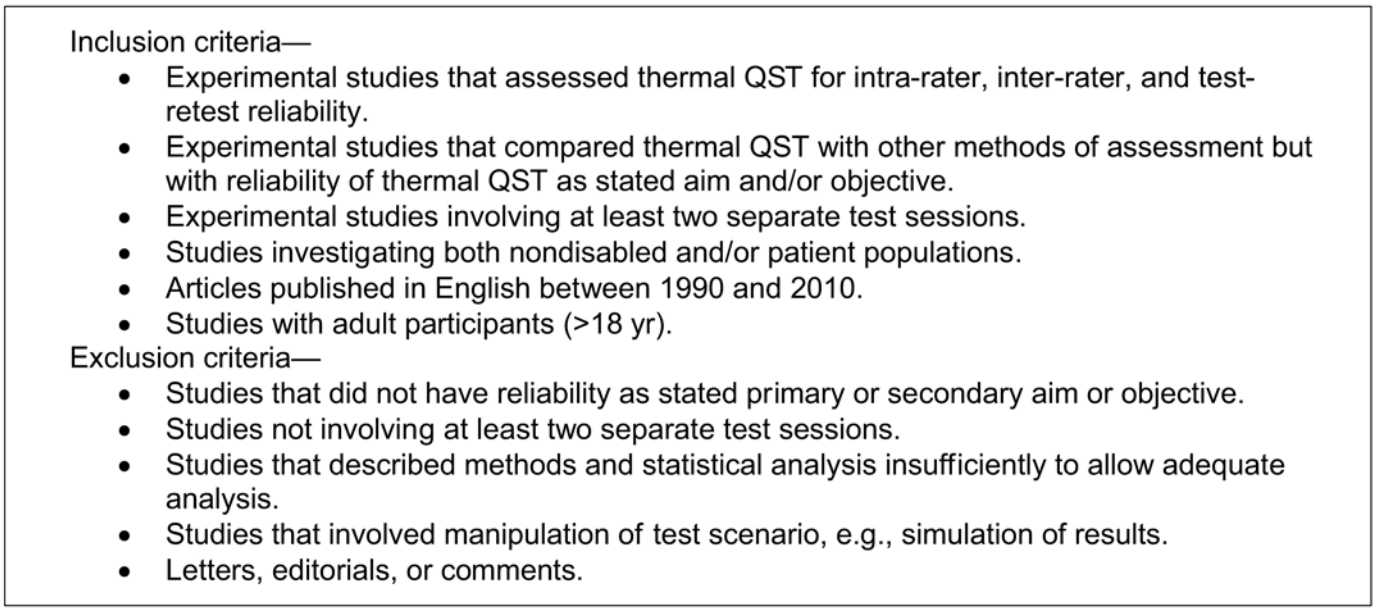

Figure 1.
Article selection criteria. QST = quantitative sensory testing.

1. Was test evaluated in sample of subjects who were representative of those to whom authors intended results to be applied?
2. Was test performed by raters who were representative of those to whom authors intended results to be applied?
3. Were raters blinded to findings or other raters during study?
4. Were raters blinded to their own prior findings of test under evaluation?
5. Were raters blinded to results of accepted reference standard or disease status for target disorder (or variable) being evaluated?
6. Were raters blinded to clinical information that was not intended to be provided as part of testing procedure or study design?
7. Were raters blinded to additional cues that were not part of test?
8. Was order of examination varied?
9. Was stability (or theoretical stability) of variable being measured taken into account when determining suitability of time-interval between repeated measures?
10. Was test applied correctly and interpreted appropriately?
11. Were appropriate statistical measures of agreement used?

Figure 2.
Quality Appraisal for Diagnostic Reliability checklist.

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 2.
Results from Quality Appraisal for Reliability Studies (QAREL) checklist.

Study	QAREL Item											High Quality[†]
	1	2	3[*]	4	5	6	7	8	9	10	11	
Agostinho et al. (2009) [1]	Yes	Unclear	Unclear/ NA	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Yes	No
Becser et al. (1998) [2]	No	Unclear	Unclear	Unclear	NA	NA	NA	No (fixed)	Yes	Unclear	Yes	No
Bird et al. (2006) [3]	Yes	Yes	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Yes	No
Bravenboer et al. (1992) [4]	Yes	Unclear	Unclear/ NA	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Yes	No
Claus et al. (1990) [5]	No	Unclear	Unclear/ NA	Unclear	NA	NA	NA	Yes	Yes	Yes	No[‡]	No
Claus et al. (1993) [6]	Yes	Unclear	Unclear/ NA	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Yes	No
De Neeling et al. (1994) [7]	Yes	Unclear	Unclear/ NA	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Yes	No
Dyck et al. (1991) [8]	Yes	Yes	Yes	Yes	Unclear	Yes	Unclear	Unclear	Yes	Yes	No[‡]	Yes
Felix and Widerström-Noga (2009) [9]	Yes	Unclear	Unclear/ NA	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Yes	Yes	No
Gelber et al. (1995) [10]	No	Unclear	Unclear/ NA	Unclear	NA	NA	NA	Unclear	Yes	Yes	Yes	No
Heldestad et al. (2010) [11]	No	Unclear	Unclear/ NA	Unclear	NA	NA	NA	Yes	Yes	Yes	Yes	Yes
Kemler et al. (2000) [12]	Yes	Yes	NA	Unclear	Unclear	Unclear	Unclear	No (fixed)	Yes	Yes	Yes	Yes
Krassioukov et al. (1999) [13]	Yes	Unclear	Unclear/ NA	Unclear	Unclear	Unclear	Unclear	No (fixed)	Yes	Yes	Yes	No
Moravcová et al. (2005) [14]	Yes	Unclear	NA	Unclear	Unclear	Unclear	Unclear	Yes	Yes	Yes	No[‡]	No
Peltier et al. (2009) [15]	Yes	Unclear	Unclear/ NA	Unclear	Unclear	Unclear	Unclear	No (fixed)	Unclear	Yes	No[‡]	No
Pigg et al. (2010) [16]	No	Unclear	Yes	Yes	NA	NA	NA	Unclear	Yes	Yes	Yes	Yes
Valensi et al. (1993) [17]	Yes	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear	No[‡]	No
Wasner and Brock (2008) [18]	No	Unclear	NA	Unclear	NA	NA	NA	Unclear	Yes	Yes	No[‡]	No
Yarnitsky and Sprecher (1994) [19]	No	Unclear	NA	Unclear	NA	NA	NA	Unclear	Yes	Yes	Yes	No
Yarnitsky et al. (1995) [20]	No	Unclear	NA	Unclear	NA	NA	NA	Unclear	Yes	Yes	Yes	No
Zwart and Sand (2002) [21]	Yes	Yes	NA	Unclear	Unclear	Yes	Unclear	Unclear	Yes	Yes	Yes	Yes

Table 2. (cont).
Results from Quality Appraisal for Reliability Studies (QAREL) checklist.

*In many studies where type of study has not been specified, it was unclear whether it was applicable for testers to be blinded to findings of other testers (noted as "Unclear/NA").

†High quality requirement: 50 percent of checklist items achieving "yes" score. With 11-item checklist, minimum of six "yes" scores was required, or in cases where checklist items were not relevant, score of 50 percent of *relevant* items was required.

‡No "‡" symbol indicates that measure of statistical analysis was appropriate, but study lacked measure of precision or sufficient raw data.

1. Agostinho CM, Scherens A, Richter H, Schaub C, Rolke R, Treede RD, Maier C. Habituation and short-term repeatability of thermal testing in healthy human subjects and patients with chronic non-neuropathic pain. Eur J Pain. 2009;13(8):779–85.
2. Becser N, Sand T, Zwart JA. Reliability of cephalic thermal thresholds in healthy subjects. Cephalalgia. 1998;18(8):574–82.
3. Bird SJ, Brown MJ, Spino C, Watling S, Foyt HL. Value of repeated measures of nerve conduction and quantitative sensory testing in a diabetic neuropathy trial. Muscle Nerve. 2006;34(2):214–24.
4. Bravenboer B, Van Dam PS, Hop J, vd Steenhoven J, Erkelens DW. Thermal threshold testing for the assessment of small fibre dysfunction: Normal values and reproducibility. Diabet Med. 1992;9(6):546–49.
5. Claus D, Hilz MJ, Neundörfer B. Thermal discrimination thresholds: A comparison of different methods. Acta Neurol Scand. 1990;81(6):533–40.
6. Claus D, Mustafa C, Vogel W, Herz M, Neundörfer B. Assessment of diabetic neuropathy: Definition of norm and discrimination of abnormal nerve function. Muscle Nerve. 1993;16(7):757–68.
7. De Neeling JN, Beks PJ, Bertelsmann FW, Heine RJ, Bouter LM. Sensory thresholds in older adults: Reproducibility and reference values. Muscle Nerve. 1994;17(4):454–61.
8. Dyck PJ, Kratz KM, Lehman KA, Karnes JL, Melton LJ 3rd, O'Brien PC, Litchy WJ, Windebank AJ, Smith BE, Low PA, et al. The Rochester Diabetic Neuropathy Study: Design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. Neurology. 1991;41(6):799–807.
9. Felix ER, Widerström-Noga EG. Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. J Rehabil Res Dev. 2009;46(1):69–84.
10. Gelber DA, Pfeifer MA, Broadstone VL, Munster EW, Peterson M, Arezzo JC, Shamoon H, Zeidler A, Clements R, Green DA, Porte D Jr, Laudadio C, Bril V. Components of variance for vibratory and thermal threshold testing in normal and diabetic subjects. J Diabetes Complications. 1995;9(3):170–76.
11. Heldestad V, Linder J, Sellersjö L, Nordh E. Reproducibility and influence of test modality order on thermal perception and thermal pain thresholds in quantitative sensory testing. Clin Neurophysiol. 2010;121(11):1878–85.
12. Kemler MA, Reulen JP, Van Kleef M, Barendse GA, Van den Wildenberg FA, Spaans F. Thermal thresholds in complex regional pain syndrome type I: Sensitivity and repeatability of the methods of limits and levels. Clin Neurophysiol. 2000;111(9):1561–68.
13. Krassioukov A, Wolfe DL, Hsieh JT, Hayes KC, Durham CE. Quantitative sensory testing in patients with incomplete spinal cord injury. Arch Phys Med Rehabil. 1999;80(10):1258–63.
14. Moravcová E, Bednaík J, Svobodník A, Dušek L. Reproducibility of thermal threshold assessment in small-fibre neuropathy patients. Scripta Medica (BRNO). 2005;78(3):177–84.
15. Peltier A, Smith AG, Russell JW, Sheikh K, Bixby B, Howard J, Goldstein J, Song Y, Wang L, Feldman EL, Singleton JR. Reliability of quantitative sudomotor axon reflex testing and quantitative sensory testing in neuropathy of impaired glucose regulation. Muscle Nerve. 2009;39(4):529–35.
16. Pigg M, Baad-Hansen L, Svensson P, Drangsholt M, List T. Reliability of intraoral quantitative sensory testing (QST). Pain. 2010;148(2):220–26.
17. Valensi P, Attali JR, Gagant S. Reproducibility of parameters for assessment of diabetic neuropathy. The French Group for Research and Study of Diabetic Neuropathy. Diabet Med. 1993;10(10):933–39.
18. Wasner GL, Brock JA. Determinants of thermal pain thresholds in normal subjects. Clin Neurophysiol. 2008;119(10):2389–95.
19. Yarnitsky D, Sprecher E. Thermal testing: Normative data and repeatability for various test algorithms. J Neurol Sci. 1994;125(1):39–45.
20. Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA. Heat pain thresholds: Normative data and repeatability. Pain. 1995;60(3):329–32.
21. Zwart JA, Sand T. Repeatability of dermatomal warm and cold sensory thresholds in patients with sciatica. Eur Spine J. 2002;11(5):441–46.

NA = not applicable.

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].

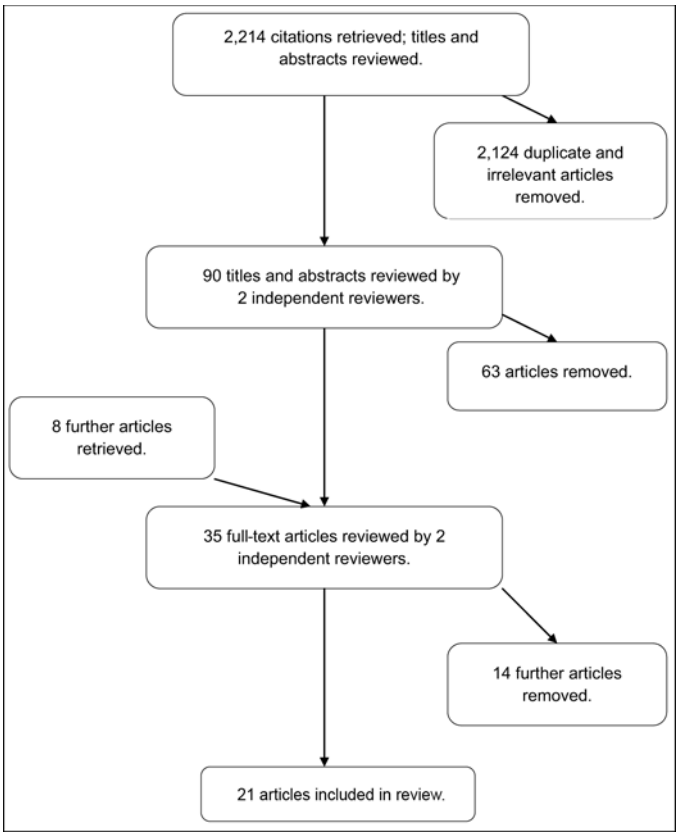

Figure 3.
Electronic search strategy results.

Methodological Issues

Modality assessed. CDT and WDT were the most commonly investigated, and all but two of the total articles reviewed examined the reliability of thermal detection thresholds. Seven studies included the assessment of thermal pain thresholds [20,22,26,29–30,37–38], while Yarnitsky et al. [32] assessed HPT alone and Wasner and Brock [31] reported on reliability of HPT and CPT. The MLI was the more common assessment method of choice studied, with nine studies having studied the MLI alone [20,22–25,29,31,37–38] while five studies assessed or compared MLI and MLE [11,21,25,27,32]. The MLE was used in seven studies [19,26,28,30,33–35].

Interval. We found the time between successive testing sessions to vary between studies. Testing intervals ranged from 2 days to 1 month, with many studies allowing some variation within this, e.g., "within a 7 day period" [24] or "3 weeks or longer" [31].

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.

Study	Method of Limits							
	Type of Reliability/ Interval	Subjects/ Testers	Modality/ Equipment	Environmental Factors/ Instructions	Measure of Repeatability	Measure of Precision/ Stability	Results	Conclusions*
Agostinho et al. (2009) [1]	Test-retest†/ 2 days	36 nondisabled; 36 non-neuropathic pain/Not specified	CDT, WDT, CPT, HPT/ TSA-II Neuro Sensory Analyzer‡	No details on environmental factors/ Standardized protocol developed by DFNS that includes verbal instructions	ANOVA	Analysis of absolute variables/Bland-Altman analysis	ANOVA: 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.	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.
Becser et al. (1998) [2]	Intrarater and interrater/Within 7 day limit	20 nondisabled/ Becser and Zwart (no other details)	CDT, WDT/ Somedic thermotest equipment§	Quiet room 22°C–23°C/ Brief details on instructions provided	CR; Bland-Altman analysis, presented as °C, ICC	95% reference limits presented as "retest as percent of first test"	Intrarater reliability— ICC average: CDT = 0.63, WDT = 0.66. CR average (°C): CDT = 1.04, WDT = 1.56. 95% reference limits (lower): CDT = 39–56, WDT = 35–68. 95% reference limits (upper): CDT = 174–244, WDT = 137–220. Interrater reliability— ICC average: CDT = 0.53, WDT = 0.60. CR average (°C): CDT = 1.17, WDT = 1.45.	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.¶
Claus et al. (1993) [3]	Test-retest/ 2 consecutive days within 1 week	30 nondisabled; 12 diabetes/ Not specified	CDT, WDT/ Modified Marstock thermode**	Not specified	Linear correlation coefficient	CV, 90th percentile day to day differences of absolute values/ Magnitude of day to day differences; % mean value	*r*-Value: CDT = 0.66, WDT = 0.77. CV (%): CDT = 65, WDT = 32. 90th percentile difference: CDT = 3.1, WDT = 2.3. Magnitude day to day difference: CDT = 19.0, WDT = 2.0.	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.
Felix and Widerström-Noga (2009) [4]	Test-retest/ 1 week	10 SCI and neuropathic pain; 10 nondisabled/Not specified	CDT, WDT, CPT, HPT/ TSA-II Neuro Sensory Analyzer‡	Quiet room with temperature controlled; skin temperature recorded/No details on instructions	ICC (one-way random effects model)	95% CI	ICC (95% CI)— SCI: CDT = 0.90 (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).	Conclusions— CDT and WDT: Excellent reliability in participants with SCI and good reliability in nondisabled participants. CPT and HPT: Fair reliability in participants with SCI and fair to good reliability in nondisabled participants.

Table 3. (cont).
Testing methods, statistical analysis, and results found in electronic search results.

Study	Method of Limits							
	Type of Reliability/ Interval	Subjects/ Testers	Modality/ Equipment	Environmental Factors/ Instructions	Measure of Repeatability	Measure of Precision/ Stability	Results	Conclusions*
Heldestad et al. (2010) [5]	Test-retest/ Between 1–8 days	38 nondisabled/Not specified	CDT, WDT, CPT, HPT/ Somedic thermotest equipment§	Quiet room; no other details provided/Standard instructions used	Repeatability between days and test sessions; CR Bland-Altman analysis; correlations between data at repeated testing in different days and different sessions within same day (Spearman rank correlation); variations in thresholds between days and test sessions (CV)	Inter- and intrasubject reproducibility; Analyses of inter- and intraindividual differences	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 °C)— Absolute values (mean): CDT = 3.10, WDT = 1.80, CPT = 0.63, HPT = 1.60. Δ (%): CDT = 65.80, WDT = 35.30, CPT = 0.29, HPT = 5.80. Within days repeatability (CV%) absolute values— CDT = 0.89–6.07 (first test), WDT = 0.33–4.10 (first test), CPT = 0.46–1.70, HPT = 0.63–8.10.	High degree of reproducibility for all measures. Detection thresholds more reliable when assessed before pain thresholds. Conclusions— CDT = good reliability. WDT = excellent reliability. CPT = excellent reliability. HPT = good reliability.
Krasioukov et al. (1999) [6]	Test-retest/ SCI: 3 weeks; Nondisabled controls: 1 week	21 SCI; 14 nondisabled/Not specified	CDT, WDT, CPT/TSA-II Neuro-Sensory Analyzer‡	Not specified	ICC	SD; CV	Nondisabled— ICC: CDT = 0.75–0.90, WDT = 0.36–0.84, CPT = 0.91 0.95. CV%: CDT = 4.00–10.80, WDT = 1.20–8.30, 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.	No systematic differences across repeated days. Very large SD/CV for CDT and CPT in participants with SCI and CPT in nondisabled participants. Conclusions— CDT = excellent reliability in nondisabled participants; fair reliability in participants with SCI. WDT = fair to excellent reliability in nondisabled participants; fair to good reliability in participants with SCI. CPT = high CV scores indicate poor to fair reliability in both groups.
Pigg et al. (2010) [7]	Intra- and interrater/ Twice day 1 by 2 examiners, again 1–2 weeks later by 1 examiner	21 nondisabled/2 authors trained by DFNS	CDT, WDT, CPT, HPT/ MSA Thermal Stimulator§	Followed DFNS protocol (DFNS provides standardized instructions but no information on environmental factors)	ICC	Mean ± SD/ MID	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.	Reliability varied although authors suggest that reliability is suitable to clinical use. MID values low for most measures with low ICC values. Conclusions— Intrarater: CDT = fair to excellent. WDT = poor to good. CPT = fair to excellent. HPT = good to excellent. Interrater: CDT = poor to good. WDT = poor to good. CPT = fair to excellent. HPT = fair to excellent.

Table 3. (cont).
Testing methods, statistical analysis, and results found in electronic search results.

Study	Type of Reliability/ Interval	Subjects/ Testers	Modality/ Equipment	Environmental Factors/ Instructions	Measure of Repeatability	Measure of Precision/ Stability	Results	Conclusions*
Method of Limits								
Wasner and Brock (2008) [8]	Test-retest/ 3 times over 3 weeks (days 0, 1, and 21)	20 nondisabled/ Wasner	CPT, HPT/ TSA-II Neuro Sensory Analyzer‡	Room held at 22°C–23°C with relative humidity of 50%–60%/ Used DFNS protocol instructions; no information on noise control	ICC	No	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.	Conclusions— CPT and HPT = good to excellent reliability but limited information.¶
Zwart and Sand (2002) [9]	Test-retest/ Tested twice 1 to 2 hours between tests	19 lumbosacral radiculopathy/ Zwart	CDT, WDT/ Somedic thermotest equipment§	Not specified	CR; ICC for between variation, repeated measures of ANOVA	Not applicable as statistical analysis sufficient	Symptomatic side— CR (%): CDT = 42–51, WDT = 39–57. ICC: 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.	Mixed results with ICC values varying from poor to excellent. Coefficients of repeatability were high throughout. Conclusions— CDT = fair reliability. WDT = fair reliability.
Mixed Method of Limits and Method of Levels								
Claus et al. (1990) [10]	Test-retest†/3 consecutive days	55 nondisabled/Not specified	CDT, WDT/ Modified Marstock thermode** (MLI, MLE)	Not specified	Reliability coefficient (Rtt)	Confidence limits mentioned but values not provided	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.	Rtt values indicate good reliability, but analyses of retest values as percentage of initial assessment reveals marked variability between tests. Conclusions— CDT and WDT = fair reliability but limited information.¶
Kemler et al. (2000) [11]	Intrarater/ 1 month	53 CRPS/ Not specified	CDT, WDT/ TSA-II NeuroSensory Analyzer‡ (MLE vs MLI)	Temperature-controlled laboratory (22°C–24°C); no visual access to computer; no visual/auditory cues/ Instructions described	CR	Bland-Altman analysis	CR— MLE: CDT unaffected 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 = 5.4, 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.	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.
Moravcová et al. (2005) [12]	Intrarater/ Twice over 1 week	58 small-fiber neuropathy; 30 nondisabled/Moravcová	CDT, WDT/ Nicolet Viking IV electrodiagnostic unit††, TSA-II NeuroSensory Analyzer‡ (MLI [random and nonrandom], MLE)	Protocol description for thermal QST very brief; authors state that "conditions were standardized" but detail insufficient/ Standardized instructions used	CR	No	Thenar 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. Thenar 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.	Reliability better for MLE than MLI. Reliability better for patient group than nondisabled participants. Conclusions— CDT and WDT = good reliability but limited information.¶

Table 3. (cont).
Testing methods, statistical analysis, and results found in electronic search results.

Study	Type of Reliability/ Interval	Subjects/ Testers	Modality/ Equipment	Environmental Factors/ Instructions	Measure of Repeatability	Measure of Precision/ Stability	Results	Conclusions*
Method of Limits								
Yarnitsky and Sprecher (1994) [13]	Intrarater/2 weeks	106 nondisabled/Not specified ("single technician")	CDT, WDT/ TSA-II Neuro Sensory Analyzer‡ (MLI, MLE, SC)	Soundproof air-conditioned room with distractions minimized/ Standard instructions used	Repeatability *r*	MISD	*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 CDT = 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.	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.
Yarnitsky et al. (1995) [14]	Test-retest/ 2 weeks	72 nondisabled/Not specified ("single technician")	HPT/TSA-II NeuroSensory Analyzer‡ (MLI, MLE, SC)	Soundproof air-conditioned room with distractions minimized/ Standard instructions used	Repeatability *r*	MISD	Thenar eminence: *r* = 5.85. Foot: *r* = 4.47	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.¶
Method of Levels								
Bird et al. (2006) [15]	Test-retest†/3 separate days within 4-week period	1,100 clinically stable diabetes with mild neuropathy/CRCC-trained neurologists and technologists	CDT/CASE IV system‡‡ (MLE: 4-2-1 stepping algorithm)	Temperature controlled room/No details on instructions or noise	Total variance (SD); ICC	CV	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%.	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.
Bravenboer et al. (1992) [16]	Test-retest†/ 2 weeks	39 diabetes without known neuropathy/Not specified	CDT, WDT/ Triple T Thermal Threshold Tester 2§§	Not specified	CR Bland-Altman analysis	No	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.	Conclusions— CDT and WDT = fair reliability in hand and poor reliability in foot but limited information.¶
De Neeling et al. (1994) [17]	Test-retest/ 13–24 days	19 nondisabled; 20 with without non-insulin dependent diabetes/Not specified ("one of three observers")	TDT (combination of CDT and WDT)	Quiet ambience with constant room temperature of 18°C–22°C/ No details on instructions	Reliability coefficient	95% CI, SD diff, CV	*r* (95% CI): 0.54 (0.26–0.73). SD diff (95% CI): 0.49 (0.39–0.61). CV: 0.72.	Fair estimate of reliability with large variance. Conclusions— TDT = poor reliability.

Table 3. (cont).
Testing methods, statistical analysis, and results found in electronic search results.

Study	Method of Levels							
	Type of Reliability/ Interval	Subjects/ Testers	Modality/ Equipment	Environmental Factors/ Instructions	Measure of Repeatability	Measure of Precision/ Stability	Results	Conclusions*
Dyck et al. (1991) [18]	Intra- and interrater/ 3–5 days	20 diabetes with and without neuropathy/ 3 neurologists	CDT, WDT/ CASE III/ IV systems‡‡	Not specified	ICC	CI on graph but specific measures not provided	ICC: CDT > 0.9, WDT > 0.8. 95% CI: CDT = 0.95–0.99 (approx), WDT = 0.55–0.90 (approx).	Results very briefly described. Conclusions— CDT and WDT = excellent reliability but limited information.¶
Gelber et al. (1995) [19]	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)	10 nondisabled for intratester reliability; compared 140 nondisabled at 6 centers/Not specified	CDT/Thermal sensitivity tester¶¶	Quiet room free from visual distractions; skin temperature recorded/ Standardized instructions used	ANOVA linear regression	CV	CV (%)— Finger: Day–day = 41, technician–technician = 60, within day (same technician) = 80, center–center = 47. Toe: Day–day = 95, technician–technician = 145, within day (same technician) = 114, center–center = 87.	Statistical methods briefly described. CV high for all measures. Conclusions— CDT = poor reliability.
Peltier et al. (2009) [20]	Test-retest/ Twice over 30 days	19 impaired glucose regulation and peripheral neuropathy/Not specified	CDT/CASE IV system‡‡ ("previously published methodology")	"Conditions of the testing were standardized"/ Standardized instructions used	ICC	No	ICC— Test 1 vs test 4: 0.80. Trial 1 vs trial 2: 0.83.	No measure of precision provided. Mean values appear similar across tests, but SD appears large. Range of CDT values look large on graph but actual values not presented. Conclusions— CDT = excellent reliability but limited information.¶
Valensi et al. (1993) [21]	Intrarater (intercenter)/ 4 weeks	132 diabetes with peripheral neuropathy/1 neurophysiologist in each center	CDT, WDT/ Thermal testing system (no additional information)	No information provided	CV; percentages of total variance	No	Total CV (%): 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 intersubject variability. Small intrasubject or intercenter variability. Conclusions— CDT and WDT = fair reliability but limited information.¶

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].
*In absence of consensus on interpretation of reliability scores and/or measures of precision, conclusions outlined are based on authors' interpretation of statistical analysis.
†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.
§§Medelc, Ltd; Old Woking, United Kingdom.
¶¶Sensortek, Inc; Clifton, New Jersey.

1. Agostinho CM, Scherens A, Richter H, Schaub C, Rolke R, Treede RD, Maier C. Habituation and short-term repeatability of thermal testing in healthy human subjects and patients with chronic non-neuropathic pain. Eur J Pain. 2009;13(8):779–85.
2. Becser N, Sand T, Zwart JA. Reliability of cephalic thermal thresholds in healthy subjects. Cephalalgia. 1998;18(8):574–82.
3. Claus D, Mustafa C, Vogel W, Herz M, Neundörfer B. Assessment of diabetic neuropathy: Definition of norm and discrimination of abnormal nerve function. Muscle Nerve. 1993;16(7):757–68.
4. Felix ER, Widerström-Noga EG. Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. J Rehabil Res Dev. 2009;46(1):69–84.
5. Heldestad V, Linder J, Sellersjö L, Nordh E. Reproducibility and influence of test modality order on thermal perception and thermal pain thresholds in quantitative sensory testing. Clin Neurophysiol. 2010;121(11):1878–85.
6. Krassioukov A, Wolfe DL, Hsieh JT, Hayes KC, Durham CE. Quantitative sensory testing in patients with incomplete spinal cord injury. Arch Phys Med Rehabil. 1999;80(10):1258–63.
7. Pigg M, Baad-Hansen L, Svensson P, Drangsholt M, List T. Reliability of intraoral quantitative sensory testing (QST). Pain. 2010;148(2):220–26.

Table 3. (cont).
Testing methods, statistical analysis, and results found in electronic search results.

8. Wasner GL, Brock JA. Determinants of thermal pain thresholds in normal subjects. Clin Neurophysiol. 2008;119(10):2389–95.
9. Zwart JA, Sand T. Repeatability of dermatomal warm and cold sensory thresholds in patients with sciatica. Eur Spine J. 2002;11(5):441–46.
10. Claus D, Hilz MJ, Neundörfer B. Thermal discrimination thresholds: A comparison of different methods. Acta Neurol Scand. 1990;81(6):533–40.
11. Kemler MA, Reulen JP, Van Kleef M, Barendse GA, Van den Wildenberg FA, Spaans F. Thermal thresholds in complex regional pain syndrome type I: Sensitivity and repeatability of the methods of limits and levels. Clin Neurophysiol. 2000;111(9):1561–68.
12. Moravcová E, Bednaík J, Svobodník A, Dušek L. Reproducibility of thermal threshold assessment in small-fibre neuropathy patients. Scripta Medica (BRNO). 2005;78(3):177–84.
13. Yarnitsky D, Sprecher E. Thermal testing: Normative data and repeatability for various test algorithms. J Neurol Sci. 1994;125(1):39–45.
14. Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA. Heat pain thresholds: Normative data and repeatability. Pain. 1995;60(3):329–32.
15. Bird SJ, Brown MJ, Spino C, Watling S, Foyt HL. Value of repeated measures of nerve conduction and quantitative sensory testing in a diabetic neuropathy trial. Muscle Nerve. 2006;34(2):214–24.
16. Bravenboer B, Van Dam PS, Hop J, vd Steenhoven J, Erkelens DW. Thermal threshold testing for the assessment of small fibre dysfunction: Normal values and reproducibility. Diabet Med. 1992;9(6):546–49.
17. De Neeling JN, Beks PJ, Bertelsmann FW, Heine RJ, Bouter LM. Sensory thresholds in older adults: Reproducibility and reference values. Muscle Nerve. 1994;17(4):454–61.
18. Dyck PJ, Kratz KM, Lehman KA, Karnes JL, Melton LJ 3rd, O'Brien PC, Litchy WJ, Windebank AJ, Smith BE, Low PA, et al. The Rochester Diabetic Neuropathy Study: Design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. Neurology. 1991;41(6):799–807.
19. Gelber DA, Pfeifer MA, Broadstone VL, Munster EW, Peterson M, Arezzo JC, Shamoon H, Zeidler A, Clements R, Green DA, Porte D Jr, Laudadio C, Bril V. Components of variance for vibratory and thermal threshold testing in normal and diabetic subjects. J Diabetes Complications. 1995;9(3):170–76.
20. Peltier A, Smith AG, Russell JW, Sheikh K, Bixby B, Howard J, Goldstein J, Song Y, Wang L, Feldman EL, Singleton JR. Reliability of quantitative sudomotor axon reflex testing and quantitative sensory testing in neuropathy of impaired glucose regulation. Muscle Nerve. 2009;39(4):529–35.
21. Valensi P, Attali JR, Gagant S. Reproducibility of parameters for assessment of diabetic neuropathy. The French Group for Research and Study of Diabetic Neuropathy. Diabet Med. 1993;10(10):933–39.
22. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–28.

ANOVA = analysis of variance, approx = approximately, CDT = cold detection threshold, CI = confidence interval, CPT = cold pain threshold, CR = coefficient of reliability, CRCC = Central Reading and Coordinating Center, CRPS = complex regional pain syndrome, CV = coefficient of variation, DFNS = German Research Network on Neuropathic Pain, HPT = heat pain threshold, ICC = intraclass correlation coefficient, MID = mean intraindividual difference, MISD = mean intersession difference, MLE = method of levels, MLI = method of limits, QST = quantitative sensory testing, Rtt = reliability, SC = staircase method, SCI = spinal cord injury, SD = standard deviation, TDT = thermal detection threshold, WDT = warm detection threshold.

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.

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]; Heldestad et al.: good for CDT and excellent for WDT [20]; Kemler 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]; Heldestad 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

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

1. Chien A, Eliav E, Sterling M. Hypoesthesia occurs in acute whiplash irrespective of pain and disability levels and the presence of sensory hypersensitivity. Clin J Pain. 2008; 24(9):759–66. [PMID: 18936593] http://dx.doi.org/10.1097/AJP.0b013e3181773b95
2. Chien A, Eliav E, Sterling M. Whiplash (grade II) and cervical radiculopathy share a similar sensory presentation: An investigation using quantitative sensory testing. Clin J Pain. 2008;24(7):595–603. [PMID: 18716498] http://dx.doi.org/10.1097/AJP.0b013e31816ed4fc
3. Chien A, Eliav E, Sterling M. Hypoaesthesia occurs with sensory hypersensitivity in chronic whiplash—Further evidence of a neuropathic condition. Man Ther. 2009;14(2): 138–46. [PMID: 18294899] http://dx.doi.org/10.1016/j.math.2007.12.004
4. Freynhagan R, Rolke R, Baron R, Tölle TR, Rutjes AK, Schu S, Treede RD. Pseudoradicular and radicular low-back pain—A disease continuum rather than different entities? Answers from quantitative sensory testing. Pain. 2008; 135(1–2):65–74. [PMID: 17570589] http://dx.doi.org/10.1016/j.pain.2007.05.004
5. Maier C, Baron R, Tölle TR, Binder A, Birbaumer N, Birklein F, Gierthmühlen J, Flor H, Geber C, Huge V, Krumova EK, Landwehrmeyer GB, Magerl W, Maihöfner C, Richter H, Rolke R, Scherens A, Schwarz A, Sommer C, Tronnier V, Uçeyler N, Valet M, Wasner G, Treede RD. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Somatosensory abnormalities in 1236 patients with difference neuropathic pain syndromes. Pain. 2010;150(3):439–50. [PMID: 20627413] http://dx.doi.org/10.1016/j.pain.2010.05.002
6. George SZ, Bishop MD, Bialosky JE, Zeppieri G Jr, Robinson ME. Immediate effects of spinal manipulation on thermal pain sensitivity: An experimental study. BMC Musculoskelet Disord. 2006;7:68. [PMID: 16911795] http://dx.doi.org/10.1186/1471-2474-7-68
7. Sterling M, Pedler A, Chan C, Puglisi M, Vuvan V, Vicenzino B. Cervical lateral glide increases nociceptive flexion reflex threshold but not pressure or thermal pain thresholds in chronic whiplash associated disorders: A pilot randomised controlled trial. Man Ther. 2010;15(2):149–53.

[PMID: 19884037]
http://dx.doi.org/10.1016/j.math.2009.09.004

8. Rolke R, Baron R, Maier C, Tölle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Bötefür IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W, Maihöfner C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Standardized protocol and reference values. Pain. 2006;123(3): 231–43. [PMID: 16697110]
http://dx.doi.org/10.1016/j.pain.2006.01.041
9. Hansson P, Backonja M, Bouhassira D. Usefulness and limitations of quantitative sensory testing: Clinical and research application in neuropathic pain states. Pain. 2007;129(3): 256–59. [PMID: 17451879]
http://dx.doi.org/10.1016/j.pain.2007.03.030
10. Chong PS, Cros DP. Technology literature review: Quantitative sensory testing. Muscle Nerve. 2004;29(5):734–47. [PMID: 15116380]
http://dx.doi.org/10.1002/mus.20053
11. Yarnitsky D, Sprecher E. Thermal testing: Normative data and repeatability for various test algorithms. J Neurol Sci. 1994;125(1):39–45. [PMID: 7964887]
http://dx.doi.org/10.1016/0022-510X(94)90239-9
12. Portney LG, Watkins MP. Foundations of clinical research: Applications to practice. 2nd ed. Upper Saddle River (NJ): Prentice Hall Health; 2000.
13. Costa LO, Maher CG, Latimer J, Smeets RJ. Reproducibility of rehabilitative ultrasound imaging for the measurement of abdominal muscle activity: A systematic review. Phys Ther. 2009;89(8):756–69. [PMID: 19520732]
http://dx.doi.org/10.2522/ptj.20080331
14. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol. 2010;63(8):854–61. [PMID: 20056381]
http://dx.doi.org/10.1016/j.jclinepi.2009.10.002
15. Van Trijffel E, Anderegg Q, Boussuyt PM, Lucas C. Interexaminer reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: A systematic review. Man Ther. 2005;10(4):256–69. [PMID: 15994114]
http://dx.doi.org/10.1016/j.math.2005.04.008
16. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–28. [PMID: 18839484]
http://dx.doi.org/10.1037/0033-2909.86.2.420
17. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998;26(4):217–38. [PMID: 9820922]
http://dx.doi.org/10.2165/00007256-199826040-00002
18. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005;19(1):231–40. [PMID: 15705040]
19. Dyck PJ, Kratz KM, Lehman KA, Karnes JL, Melton LJ 3rd, O'Brien PC, Litchy WJ, Windebank AJ, Smith BE, Low PA, et al. The Rochester Diabetic Neuropathy Study: Design, criteria for types of neuropathy, selection bias, and reproducibility of neuropathic tests. Neurology. 1991;41(6):799–807. [PMID: 2046920]
20. Heldestad V, Linder J, Sellersjö L, Nordh E. Reproducibility and influence of test modality order on thermal perception and thermal pain thresholds in quantitative sensory testing. Clin Neurophysiol. 2010;121(11):1878–85. [PMID: 20478739]
http://dx.doi.org/10.1016/j.clinph.2010.03.055
21. Kemler MA, Reulen JP, Van Kleef M, Barendse GA, Van den Wildenberg FA, Spaans F. Thermal thresholds in complex regional pain syndrome type I: Sensitivity and repeatability of the methods of limits and levels. Clin Neurophysiol. 2000; 111(9):1561–68. [PMID: 10964065]
http://dx.doi.org/10.1016/S1388-2457(00)00358-8
22. Pigg M, Baad-Hansen L, Svensson P, Drangsholt M, List T. Reliability of intraoral quantitative sensory testing (QST). Pain. 2010;148(2):220–26. [PMID: 20022428]
http://dx.doi.org/10.1016/j.pain.2009.10.024
23. Zwart JA, Sand T. Repeatability of dermatomal warm and cold sensory thresholds in patients with sciatica. Eur Spine J. 2002;11(5):441–46. [PMID: 12384751]
http://dx.doi.org/10.1007/s005860100333
24. Becser N, Sand T, Zwart JA. Reliability of cephalic thermal thresholds in healthy subjects. Cephalalgia. 1998;18(8): 574–82. [PMID: 9827251]
http://dx.doi.org/10.1046/j.1468-2982.1998.1808574.x
25. Claus D, Hilz MJ, Neundörfer B. Thermal discrimination thresholds: A comparison of different methods. Acta Neurol Scand. 1990;81(6):533–40. [PMID: 2220312]
http://dx.doi.org/10.1111/j.1600-0404.1990.tb01015.x
26. Gelber DA, Pfeifer MA, Broadstone VL, Munster EW, Peterson M, Arezzo JC, Shamoon H, Zeidler A, Clements R, Green DA, Porte D Jr, Laudadio C, Bril V. Components of variance for vibratory and thermal threshold testing in normal and diabetic subjects. J Diabetes Complications. 1995;9(3):170–76. [PMID: 7548981]
http://dx.doi.org/10.1016/1056-8727(94)00042-M
27. Moravcová E, Bednaík J, Svobodník A, Dušek L. Reproducibility of thermal threshold assessment in small-fibre neuropathy patients. Scripta Medica (BRNO). 2005;78(3): 177–84.
28. Valensi P, Attali JR, Gagant S. Reproducibility of parameters for assessment of diabetic neuropathy. The French Group for Research and Study of Diabetic Neuropathy.

Diabet Med. 1993;10(10):933–39. [PMID: 8306589] http://dx.doi.org/10.1111/j.1464-5491.1993.tb00009.x

29. Felix ER, Widerström-Noga EG. Reliability and validity of quantitative sensory testing in persons with spinal cord injury and neuropathic pain. J Rehabil Res Dev. 2009;46(1): 69–84. [PMID: 19533521] http://dx.doi.org/10.1682/JRRD.2008.04.0058
30. Peltier A, Smith AG, Russell JW, Sheikh K, Bixby B, Howard J, Goldstein J, Song Y, Wang L, Feldman EL, Singleton JR. Reliability of quantitative sudomotor axon reflex testing and quantitative sensory testing in neuropathy of impaired glucose regulation. Muscle Nerve. 2009;39(4): 529–35. [PMID: 19260066] http://dx.doi.org/10.1002/mus.21210
31. Wasner GL, Brock JA. Determinants of thermal pain thresholds in normal subjects. Clin Neurophysiol. 2008;119(10): 2389–95. [PMID: 18778969] http://dx.doi.org/10.1016/j.clinph.2008.07.223
32. Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA. Heat pain thresholds: Normative data and repeatability. Pain. 1995; 60(3):329–32. [PMID: 7596629] http://dx.doi.org/10.1016/0304-3959(94)00132-X
33. De Neeling JN, Beks PJ, Bertelsmann FW, Heine RJ, Bouter LM. Sensory thresholds in older adults: Reproducibility and reference values. Muscle Nerve. 1994;17(4): 454–61. [PMID: 8170493] http://dx.doi.org/10.1002/mus.880170414
34. Bird SJ, Brown MJ, Spino C, Watling S, Foyt HL. Value of repeated measures of nerve conduction and quantitative sensory testing in a diabetic neuropathy trial. Muscle Nerve. 2006;34(2):214–24. [PMID: 16708368] http://dx.doi.org/10.1002/mus.20577
35. Bravenboer B, Van Dam PS, Hop J, vd Steenhoven J, Erkelens DW. Thermal threshold testing for the assessment of small fibre dysfunction: Normal values and reproducibility. Diabet Med. 1992;9(6):546–49. [PMID: 1643803] http://dx.doi.org/10.1111/j.1464-5491.1992.tb01836.x
36. Claus D, Mustafa C, Vogel W, Herz M, Neundörfer B. Assessment of diabetic neuropathy: Definition of norm and discrimination of abnormal nerve function. Muscle Nerve. 1993;16(7):757–68. [PMID: 8505932] http://dx.doi.org/10.1002/mus.880160711
37. Krassioukov A, Wolfe DL, Hsieh JT, Hayes KC, Durham CE. Quantitative sensory testing in patients with incomplete spinal cord injury. Arch Phys Med Rehabil. 1999;80(10): 1258–63. [PMID: 10527084] http://dx.doi.org/10.1016/S0003-9993(99)90026-6
38. Agostinho CM, Scherens A, Richter H, Schaub C, Rolke R, Treede RD, Maier C. Habituation and short-term repeatability of thermal testing in healthy human subjects and patients with chronic non-neuropathic pain. Eur J Pain. 2009;13(8):779–85. [PMID: 19019713] http://dx.doi.org/10.1016/j.ejpain.2008.10.002
39. Jensen R, Hystad T, Kvale A, Baerheim A. Quantitative sensory testing of patients with long lasting patellofemoral pain syndrome. Eur J Pain. 2007;11(6):665–76. [PMID: 16268247] http://dx.doi.org/10.1016/j.ejpain.2006.10.007
40. El-Masry Ivanovic S, Faris H, Baad-Hansen L, Svensson P, List T, Dransgholt M. Preliminary report of reliability of quantitative sensory testing (QST) in patients with persistent idiopathic dentoalveolar pain. 13th World Congress on Pain; 2010 Aug 29–Sep 2; Montreal, Canada. Seattle (WA): IASP; 2010.
41. Geber C, Klein T, Azad S, Baron R, Birklein F, Gierthmühlen J, Huge V, Lauchart M, Nitzsche D, Stengel M, Valet M, Baron R, Maier C, Tölle T, Treede R. Test-retest and interobserver reliability of quantitative sensory testing according to the protocol of the German Research Network on Neuropathic Pain (DFNS): A multi-centre study. Pain. 2011;152(3):548–56. [PMID: 21237569] http://dx.doi.org/10.1016/j.pain.2010.11.013

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

This article and any complementary material should be cited as follows:
Moloney NA, Hall TM, Doody CM. Reliability of thermal quantitative sensory testing: A systematic review. J Rehabil Res Dev. 2012;49(2):191–208.
http://dx.doi.org/10.1682/JRRD.2011.03.0044