PAIN® 152 (2011) 548–556

www.elsevier.com/locate/pain

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

Christian Geber [a,*], Thomas Klein [b], Shahnaz Azad [f], Frank Birklein [a], Janne Gierthmühlen [c], Volker Huge [f], Meike Lauchart [f], Dorothee Nitzsche [e], Maike Stengel [c], Michael Valet [e], Ralf Baron [c,1], Christoph Maier [d,1], Thomas Tölle [e,1], Rolf-Detlef Treede [b,1]

[a] Klinik und Poliklinik für Neurologie, Universitätsmedizin der Johannes Gutenberg-Universität, Mainz, Germany
[b] Lehrstuhl für Neurophysiologie, CBTM, Medizinische Fakultät Mannheim der Universität Heidelberg, Heidelberg, Germany
[c] Klinik für Neurologie, Campus Kiel, Universitätsklinikum Schleswig-Holstein, Kiel, Germany
[d] Abteilung Schmerztherapie, Berufsgenossenschaftliches Universitätsklinikum Bergmannsheil, Bochum, Germany
[e] Klinik für Neurologie, Technische Universität München, Germany
[f] Klinik für Anästhesie der Ludwig-Maximilians-Universität, München, Germany

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

ARTICLE INFO

Article history:
Received 18 May 2010
Received in revised form 13 October 2010
Accepted 12 November 2010

Keywords:
Quantitative sensory testing (QST)
Reliability
Multi-centre
Neuropathic pain
Somatosensory profiles
Diagnostic assessment

ABSTRACT

Quantitative sensory testing (QST) is an instrument to assess positive and negative sensory signs, helping to identify mechanisms underlying pathologic pain conditions. In this study, we evaluated the test–retest reliability (TR-R) and the interobserver reliability (IO-R) of QST in patients with sensory disturbances of different etiologies. In 4 centres, 60 patients (37 male and 23 female, 56.4 ± 1.9 years) with lesions or diseases of the somatosensory system were included. QST comprised 13 parameters including detection and pain thresholds for thermal and mechanical stimuli. QST was performed in the clinically most affected test area and a less or unaffected control area in a morning and an afternoon session on 2 consecutive days by examiner pairs (4 QSTs/patient). For both, TR-R and IO-R, there were high correlations ($r = 0.80–0.93$) at the affected test area, except for wind-up ratio (TR-R: $r = 0.67$; IO-R: $r = 0.56$) and paradoxical heat sensations (TR-R: $r = 0.35$; IO-R: $r = 0.44$). Mean IO-R ($r = 0.83$, 31% unexplained variance) was slightly lower than TR-R ($r = 0.86$, 26% unexplained variance, $P < .05$); the difference in variance amounted to 5%. There were no differences between study centres. In a subgroup with an unaffected control area (n = 43), reliabilities were significantly better in the test area (TR-R: $r = 0.86$; IO-R: $r = 0.83$) than in the control area (TR-R: $r = 0.79$; IO-R: $r = 0.71$, each $P < .01$), suggesting that disease-related systematic variance enhances reliability of QST. We conclude that standardized QST performed by trained examiners is a valuable diagnostic instrument with good test–retest and interobserver reliability within 2 days. With standardized training, observer bias is much lower than random variance.

1. Introduction

Quantitative sensory testing (QST) is an acknowledged diagnostic tool for the assessment of somatosensory changes caused by nerve lesions of different etiologies, such as polyneuropathy [3,47], and is regarded a useful diagnostic instrument in the assessment of neuropathic pain [15,16,24,51]. Sensory changes in QST can also be found in non-neuropathic pain conditions such as arthritic, myofascial, and fibromyalgic pain [6,25,34].

The widespread use of QST has been limited so far because of different QST procedures used, restriction to local centres, and often not comprising all somatosensory modalities [6,55]. To overcome these limitations, the German Research Network on Neuropathic Pain (DFNS) selected a set of tests comprising all somatosensory submodalities mediated by different primary afferents (C-, Aδ-, Aβ-) including the assessment of deeper tissue layers. Standardized instructions were established, and a training program for examiners and a certification for laboratories were implemented [26,44]. In addition, a multi-centre reference database was generated [43]. However, as an important quality criterion, data on reliability of this testing procedure are required [6,30,47,54]. Previous QST studies had shown that psychophysical paradigms (eg, method of limits vs levels; threshold vs suprathreshold measures),

* Corresponding author.
E-mail address: geber@uni-mainz.de (C. Geber).
[1] These authors are on the DFNS steering committee.

0304-3959/$36.00
doi:10.1016/j.pain.2010.11.013

technical equipment and the way of test stimuli application may influence the reliability of QST [13,14,22,48,54,55]. A literature review by Siao and Cros [48] on reproducibility of QST studies from 1981 to 2001 identified characteristics of the study population, the sample size, the number of examiners and centres involved in the study, the chosen test areas, the baseline skin temperature, and the time intervals as additional factors influencing reproducibility.

The primary aim of the present multi-centre study was to investigate to what extent the QST results of the DFNS protocol depend on the performance of the examiner. Therefore, interobserver reliability (IO-R) was assessed by testing the same patients experiencing a wide range of neuropathy and neuropathic pain conditions by 2 trained examiners in each study centre. In addition, test–retest reliability was assessed by testing the same patients on 2 consecutive days to evaluate the influence of patient and procedural factors. The multi-centre approach of this study allowed an evaluation of the homogeneity of QST data across different centres that were trained for the same QST protocol. Interim results from one centre (Mainz) were presented at the Second International Congress on Neuropathic Pain in Berlin [23].

2. Methods

The test–retest and interobserver reliability of quantitative sensory testing (QST) according to the protocol of the German Network on Neuropathic Pain (DFNS) [44] was addressed in a multi-centre study involving 4 centres in Germany, as follows: Centre 1: Department of Neurology of the University medical centre of the Johannes Gutenberg-Universität, Mainz; Centre 2: Department of Neurology of the Technische Universität, München; Centre 3: Department of Anesthesiology of the Ludwig-Maximilians-Universität, München; and Centre 4: Department of Neurology, Division of Neurological Pain Research and Therapy, Christian-Albrechts-Universität, Kiel. The study protocol for patient assessment was approved by the ethics committees of all participating centres.

2.1. Patients

Overall, 60 inpatients and outpatients with various unilateral or bilateral, peripheral, or central lesions or diseases of the somatosensory system with or without pain (37 male, 23 female, mean age 56.4 ± 1.9 years, range 20–78 years; Table 1) and with stable medication were investigated between November 2005 and March 2008. Exclusion criteria were the following: problems with understanding the QST instructions because of aphasia, impaired vigilance, dementia/cognitive impairment, or inadequate knowledge of the German language (inability to understand the standardized instruction); Parkinson's disease (potential confounder of reaction time sensitive QST measures); or multiple sclerosis (fluctuating neurological deficits, eg, depending on temperature).

2.2. Study design

In each patient, 4 QST sessions performed by 2 examiners were distributed over 2 consecutive days. Within 1 day, a morning and an afternoon QST session were performed. With an individual patient, the same examiners performed either the morning or the afternoon sessions, and the assignment of morning or afternoon sessions was balanced across examiners. In total, 240 QST profiles in 60 patients were obtained. Within each centre, there were 1 to 2 pairs of examiners. All examiners in this study had received a special training session (Mainz) and performed QST regularly in the clinical routine. Before the initiation of this study, a retraining session was held at each centre by an experienced QST examiner not further involved in this study.

2.3. Test area and control area

At the first QST session the area of QST testing was marked on the skin to ensure that the same skin area was tested across all assessments. The most symptomatic skin area was referred to as the test area (TA), whereas the unaffected skin area (43 patients) or less affected skin area (17 patients) was referred to as the con-

Table 1
Demographic and clinical data of study subjects.

Characteristic			Total	CA unaffected	CA partly affected
Patients (n)			60	43	17
Age (y)			56.4 ± 1.9		
Gender (male/female)			37/23	26/17	11/6
Pain (yes/no)			47/13	35/12	12/1
Pain medication (yes/no)			34/26	27/17	7/9
Diagnosis (pain/no pain)					
Peripheral nerve lesion	Polyneuropathy		25 (21/4)	14 (12/2)	11 (9/2)
	Peripheral nerve injury		5 (2/3)	4 (2/2)	1 (1/0)
	Post Herpetic Neuralgia (PHN)		3 (3/0)	3	0
	Trigeminal neuralgia (TGN)		1 (1/0)	1	0
	Radiculopathy		4 (2/2)	4	0
	Plexus lesion		3 (2/1)	3	0
	Lumbar spinal stenosis		3 (1/2)	2 (1/1)	1
CNS lesion	Stroke (CPSP)		2 (2/0)	2	0
	Syringomyelia		1 (1/0)	1	0
	Spinal lesion (myelitis)		4 (2/2)	2 (2/0)	2 (0/2)
Other neuropathy	CRPS		3 (3/0)	3	0
	Unclear neuropathy		6 (5/1)	4 (4/0)	2 (1/1)
Areas	Test area	Face	3	3	0
		Hand	6	4	2
		Foot	39	25	14
		Other	12	11	1
	Control area	Face	4		
		Hand	21		
		Foot	24		
		Other	11		

CA, control area; CPSP, central poststroke pain; CRPS, complex regional pain syndrome.
The classifications ''unaffected'' and ''partly affected CA'' refer to the clinical assessment and not to the QST assessment.

trol area (CA). Before and after each test session, the skin temperature was assessed at the TA and the CA.

2.4. Pain ratings

Subjects rated the magnitude of ongoing pain as well as pain to mechanical test stimuli on a numeric rating scale (NRS) ranging from 0 (nonpainful) to 100 (most intense pain imaginable). Before each test session, patients were asked for their pain characteristics including the ongoing pain intensity and their present (pain) medication.

2.5. Test stimuli

Skin and muscle sensitivity was tested using a standardized test battery for QST [43,44] that provides a comprehensive profile of somatosensory functions within 30 minutes per TA and encompasses thermal as well as mechanical testing procedures (discussed below). The QST always started at the CA.

2.5.1. Thermal detection and pain thresholds; paradoxical heat sensations

Thermal thresholds were determined using a TSA 2001-II (MEDOC, Ramat Yishai, Israel) thermal sensory testing device with a thermode of Peltier elements (contact area 16 × 16 mm; 32 °C baseline temperature; ramped stimuli with 1 C/s; method of limits). First, the cold detection threshold (CDT) and warm detection threshold (WDT) were measured in triplicate. The number of paradoxical heat sensations (PHS; ie, reports of hot or burning sensations to innocuous cold stimuli) was determined during the thermal sensory limen (TSL; the difference limen for alternating warm and cold stimuli) procedure, followed by determination of the cold pain threshold (CPT) and heat pain threshold (HPT) measured in triplicate. The mean threshold temperature of the 3 consecutive measurements was calculated.

2.5.2. Vibration detection threshold

The vibration detection threshold (VDT) was assessed with a Rydel–Seiffert tuning fork (64 Hz, 8/8 scale) applied with suprathreshold vibration intensity left on the skin until the sensation of vibration had ceased. The final vibration detection threshold was the arithmetic mean of 3 consecutive measurements.

2.5.3. Mechanical detection threshold for modified von Frey filaments

The mechanical detection threshold (MDT) was assessed using a standardized set of modified von Frey hairs (Optihair2-Set, Marstock Nervtest, Germany) exerting forces between 0.25 and 512 mN. The contact area of the von Frey hairs with the skin was a rounded tip (0.5 mm in diameter) to avoid sharp edges that would facilitate nociceptor activation. Threshold determinations were made using an adaptive method of limits by series of alternating ascending and descending stimulus intensities yielding 5 just suprathreshold and 5 just subthreshold estimates. The final threshold was the geometric mean of the 10 determinations.

2.5.4. Mechanical pain threshold for pinprick stimuli

The mechanical pain threshold (MPT) was assessed using custom-made weighted pinprick stimuli with fixed stimulus intensities (8, 16, 32, 64, 128, 256, and 512 mN; flat contact area, 0.25 mm in diameter; PinPrick; MRC Systems GmbH, Germany). These punctate stimuli adequately excite cutaneous nociceptors [29,49]. Again, threshold determinations were made using an adaptive method of limits by series of alternating ascending and descending stimulus intensities yielding 5 just suprathreshold and 5 just subthreshold estimates. The final threshold was the geometric mean of the 10 estimates.

2.5.5. Mechanical pain sensitivity for pinprick stimuli and dynamic mechanical allodynia for stroking light touch

In a separate test, a stimulus–response function for mechanical pain sensitivity (MPS) was determined using the same weighted pinprick stimuli as for mechanical pain threshold. In addition, pain in response to light touch (dynamic mechanical allodynia [DMA]) was tested by gentle/light stroking with a cotton wisp (∼3 mN), a cotton wool tip fixed to an elastic strip (∼100 mN), and a brush (∼200–400 mN). Each of the 7 intensities of pinpricks and of the 3 intensities of light stroking was applied 5 times in a balanced sequence. The MPS was calculated as the geometric mean of all pain ratings for pinprick stimuli, and allodynia was quantified as the geometric mean of all pain ratings after light touch stimuli.

2.5.6. Wind-up ratio — temporal pain summation for repetitive pinprick stimuli

In this test, the perceived magnitude of pain to a series of pinprick stimuli (pinprick force: 256 mN, repeated 10 times at a rate of 1/s on separate spots within a small area of ∼1 cm^2) was compared to a single pinprick stimulus of the same force. The patient was asked to assign a pain rating for the single stimulus and for the pain reached at the end of the train. This procedure was applied 5 times at different skin sites within the marked area. The mean pain rating of trains divided by the mean pain rating to single stimuli was calculated as wind-up ratio (WUR). The WUR was not calculated if the first (single) stimulus was rated numeric rating scale 0/100 in more than 3 assessments, and in this case was handled as missing data.

2.5.7. Pressure pain threshold

Pressure pain threshold (PPT) was assessed by using a pressure gauge device (FDN200, Wagner Greenwich, CT, USA Instruments) with a probe area of 1 cm^2 that exerts pressure up to ∼2000 kPa. The pressure pain threshold was determined by 3 series of ascending stimulus intensities with a slowly increasing stimulus ramp (50 kPa/s). The final threshold was the arithmetic mean of the 3 consecutive measurements.

2.6. Data evaluation and statistical analysis

Ongoing pain ratings and skin temperature before each QST session were compared using a one-factorial analysis of variance (ANOVA; factor "session": S1–S4).

2.6.1. Evaluation of QST data

All QST values with the exception of CPT and HPT, VDT, and PHS were transformed into decadic logarithm to achieve a (secondary) normal distribution [43]. To avoid a loss of zero values, a small constant (0.1) was added to all pain ratings (mechanical pain sensitivity, dynamic mechanical allodynia) before log-transformation [35].

2.6.2. Z-transformation of QST data to create profiles of sensory changes

To compare a patient's QST data profile with control data independent of the physical units of measurement across QST parameters, the patient data were transformed into a standard normal distribution (Z-transformation; zero mean, unit variance) for each single parameter except DMA and PHS by using the following expression:

$$Z = (\text{value}_{\text{patient}} - \text{mean}_{\text{controls}})/\text{SD}_{\text{controls}}$$

Normative data of the German Research Network on Neuropathic Pain (DFNS) were used as reference values for *z*-score calculation [43].

For clarity of data presentation, we adjusted the algebraic sign of *z*-score values for each parameter so that it reflects the individual patient's sensitivity for this parameter. *z*-scores above zero indicate a gain of function when the patient is more sensitive to the tested stimuli compared with controls, whereas *z*-scores below zero indicate a loss of function, referring to a lower sensitivity of the patient. Thus, elevations of thresholds (CDT, WDT, TSL, HPT, CPT, MDT, MPT, VDT, PPT) result in negative *z*-scores, whereas increases in ratings (MPS, WUR) result in positive *z*-scores. Dynamic mechanical allodynia and paradoxical heat sensation could not be transformed into *z*-scores because they are usually absent at the control condition (ie, mean and standard deviation are zero in the reference group). Instead, original data are shown.

2.6.3. Mean value differences

Because correlation analyses measure the strength of the relationship, but not the agreement between variables [55], the intrasubject between-session variance was analysed beforehand by calculating mean value differences. The mean value differences for test–retest reliability (TR-R), interobserver reliability (IO-R), and circadian rhythm effects (morning vs afternoon session) were calculated for each QST parameter by subtracting the QST data of the second assessment from the first assessment for each individual subject. In QST parameters that were log-transformed to achieve a secondary normal distribution, differences between the first and second assessment correspond to between-session ratios. Paired *t* tests were used to assess significant differences (Table 2).

2.6.4. Correlation analyses

Correlation analysis was performed by using original or log-transformed data.

Pearson's product–moment correlation was used for ongoing pain ratings and all QST parameters, except for DMA and PHS. For these parameters, Spearman's rank correlation was calculated.

To test for TR-R, ongoing pain ratings and the single QST measures obtained at day 1 and day 2 within the same patient by the same examiner were correlated. QST data of examiner 1 and 2 were treated as individual cases, ie, in each patient the QST measures of examiner 1 on day 1 were correlated with QST measures of examiner 1 on day 2, and, in analogy, QST measures of examiner 2 on day 1 were correlated with QST measures of examiner 2 on day 2. This procedure resulted in a total number of 120 QST data pairs (2 examiners × 60 patients).

To test for IO-R, ongoing pain ratings and the single QST measures obtained by the 2 examiners within the same patient were correlated. QST data obtained by the same examiner on days 1 and 2 were treated as individual cases, ie, in each patient the measures of the QST assessment of examiner 1 on day 1 were correlated with the measures of the QST assessment of examiner 2 on day 1, and, in analogy, the measures of the QST assessment of examiner 1 on day 2 were correlated with the measures of the QST assessment of examiner 2 on day 2. This procedure resulted in a total of 120 QST data pairs (60 patients × 2 days).

To test for circadian rhythm effects, ongoing pain ratings and the single QST measures obtained in the morning and afternoon sessions within the same patient were correlated. For this analysis, QST data pairs (morning/afternoon session) of days 1 and 2 were treated as individual cases; ie, in each patient the measures of the QST assessment in the morning of day 1 were correlated with the measures of the QST assessment in the afternoon of day 1, and, in analogy, the measures of the QST assessment in the morning of day 2 were correlated with the measures of the QST assessment in the afternoon of day 2. This procedure resulted in a total of 120 QST data pairs (60 patients × 2 days).

To calculate mean correlation coefficients across all QST parameters (except DMA and PHS), the single correlation coefficients were transformed into the arcus tangens hyperbolicus (Fisher transformation) to get a secondary normal distribution [1]. Resulting arithmetic means of the arcus tangens hyperbolicus were retransformed into correlation coefficients using the tangens hyperbolicus function.

2.6.5. Comparison of reliability across the participating centres

A one-factorial ANOVA (factor: "centre": 1–4) was applied to compare reliability (IO-R) between centres. For this analysis, data for only the clinically affected area, or test area (TA), were analysed.

P values less than .05 were generally considered statistically significant. Data are presented as mean ± standard error of the mean (SEM).

Table 2
Mean values, correlation coefficients, and paired *t* test results for all data pairs (TR-R, IO-R, and circadian rhythm effects) for ongoing pain (NRS, 0–100) and all QST parameters.

		Unit	Test–retest				Interobserver				Circadian rhythm			
			Mean values			Correlation	Mean values			Correlation	Mean values			Correlation
			Day 1	Day 2	*t*	*r*	Observer 1	Observer 2	*t*	*r*	Morning	Afternoon	*t*	*r*
Pain		NRS	25.0	22.9	1.26	0.761	24.8	23.7	0.55	0.752	25.0	23.5	0.85	0.753
QST	CDT[a]	°C	−7.53	−7.75	−0.55	0.852	−7.52	−7.77	−0.61	0.847	−7.55	−7.73	−0.43	0.847
	WDT[a]	°C	8.28	8.21	0.18	0.802	8.29	8.20	0.28	0.837	8.34	8.15	0.60	0.838
	TSL[a]	°C	18.14	17.76	0.54	0.880	17.67	18.23	−0.75	0.856	17.95	17.95	0.01	0.855
	CPT[b]	°C	11.01	10.91	0.20	0.855	11.03	10.89	0.24	0.812	10.76	11.16	−0.71	0.811
	HPT[b]	°C	46.32	46.24	0.46	0.881	46.15	46.41	−1.30	0.871	46.29	46.27	0.13	0.868
	PPT[a]	kPa	352.65	359.49	−0.66	0.881	345.74	366.68	−1.17	0.843	354.73	357.39	−0.22	0.839
	MPT[a]	mN	81.02	69.07	1.95	0.802	73.24	76.41	−0.51	0.795	78.83	70.99	1.26	0.796
	MPS[a]	NRS	0.75	0.79	−0.80	0.902	0.79	0.95	−2.06*	0.830	0.70	0.84	−1.77	0.826
	WUR[a,c]		2.87	2.68	1.00	0.671	2.63	3.12	−2.20*	0.556	3.09	2.65	1.95	0.558
	MDT[a]	mN	16.69	14.54	1.58	0.895	15.47	15.68	−0.15	0.885	16.32	14.86	1.04	0.887
	VDT[b]	×/8	4.99	4.94	0.59	0.932	5.13	4.80	2.62*	0.886	4.91	5.01	−0.78	0.849
	DMA[a]	NRS	0.14	0.12	1.36	0.870	0.22	0.24	−1.24	0.790	0.12	0.13	−0.82	0.782
	PHS[b]	×/3	0.74	0.64	0.93	0.351	0.68	0.71	−0.35	0.444	0.67	0.72	−0.52	0.445
	Mean (QST)					**0.861**				**0.832**				**0.827**

Abbreviations as in text.
[a] Geometric mean.
[b] Arithmetic mean.
[c] In case of WUR, only 88 of 120 (TR-R) or 86 of 120 (IO-R) could be analysed.
* $P < .05$.

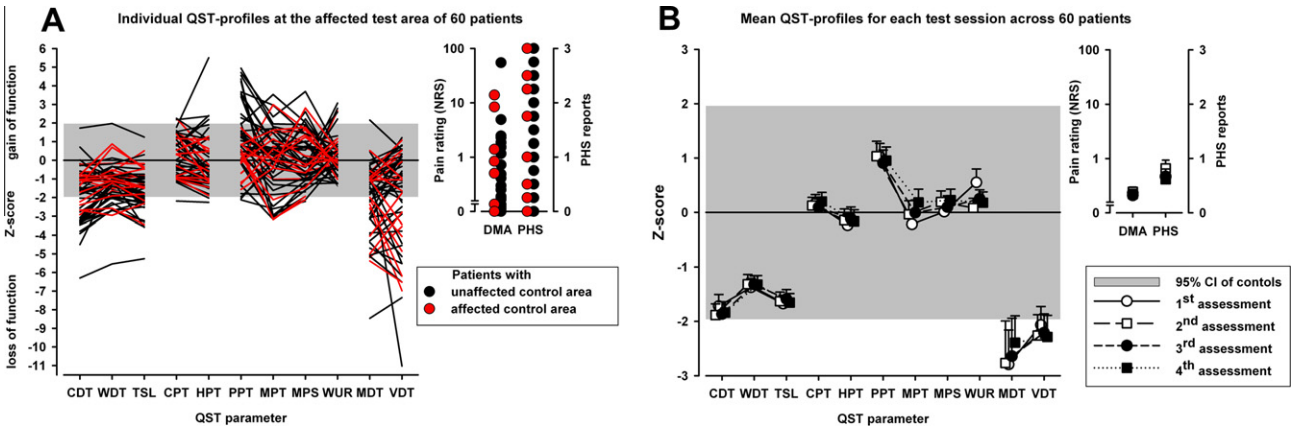

Fig. 1. Somatosensory Z-profiles of the test area. (A) Profiles per patient, averaged across 4 assessments. Black lines represent patients with unaffected control areas; red lines indicate patients with partly affected control areas. Shaded area represents 95% confidence interval of reference database. (B) Profiles per assessment, averaged across 60 patients. Open circle represents assessment 1, open square, assessment 2; filled circle, assessment 3; filled square, assessment 4. Data are mean ± SEM.

3. Results

3.1. Demographic data and clinical characteristics

Patients' demographic data and clinical characteristics are given in Table 1.

Overall, 44 of 60 patients had peripheral neuropathies, with painful (n = 21) and nonpainful (n = 4) polyneuropathies being the largest group. Six patients had lesions of the central nervous system, and 10 patients had other neuropathies. The area designated as the test area (TA) was always clinically affected. In the control area (CA), patients were either clinically unaffected (43/60 patients) or showed some weak somatosensory disturbances compared with the healthy reference group (ie, symptomatic, but less pronounced compared with the TA, 17/60 patients). Ongoing pain ratings and skin temperature were assessed immediately before each QST session and were not statistically different between sessions (pain, numeric rating scale 0–100: S1: 26.1 ± 3.7, S2: 23.6 ± 3.5, S3: 22.5 ± 3.4, S4: 23.0 ± 3.4; $F_{1,59} = 1.02$, $P = .39$; skin temperature: S1: 31.0 ± 0.3 °C, S2: 31.5 ± 0.3 °C, S3: 31.3 ± 0.3 °C, S4: 31.3 ± 0.3 °C; $F_{1,59} = 1.34$, $P = .26$).

3.2. QST profiles of somatosensory changes in all patients

Somatosensory changes in the test area of all patients (n = 60) are presented as the means of *z*-scores across the 4 assessments (QST profiles, Fig. 1A). Overall, many patients showed a sensory loss of nonpainful thermal and mechanical submodalities (negative *z*-scores) compared with normative data [43], suggesting substantial deafferentiation. In contrast, painful thermal and mechanical submodalities were either increased or decreased, indicating either hyper- or hypoalgesia. The distribution of QST profiles for patients with an unaffected CA that were used in the subgroup analysis (black lines, n = 43) did not differ from patients, whose CA could not be regarded as unaffected (red lines, n = 17).

Data analysis on an individual level revealed that *z*-scores of the non-nociceptive thermal and mechanical submodalities in the test area (n = 60) were less than −1.95 in 40% (CDT), 23% (WDT), 37% (TSL), and 43% (both MDT and VDT), indicating a pathologic loss of sensitivity. Increased sensitivity to non-nociceptive stimuli (*z*-scores >1.95) occurred only occasionally (1–2%). Concerning thermal and mechanical nociceptive submodalities, cold and heat hyperalgesia occurred more frequently (CPT and HPT, both 8%) than cold and heat hypoalgesia (2–3%). Significant hyperalgesia to pinprick-evoked pain occurred in 10% (both MPT and MPS); hypoalgesia to pinprick-evoked pain was detected in 7% (MPS) and 18% (MPT) of patients. Hyperalgesia to blunt pressure (PPT) was more frequent (25%) than hypoalgesia to blunt pressure pain (7%). The WUR was significantly increased in 25% of subjects, whereas a decrease in WUR was not observed.

Fig. 1B shows that the average QST profile across patients did not vary systematically across the 4 assessments, which were performed by 2 observers.

3.3. Test–retest and interobserver reliability of QST assessments

As an example, Fig. 2 illustrates the correlation analysis for one QST parameter, the CDT, which is a sensitive parameter for small fiber function or spinothalamic tract function deficits. Some patients consistently detected cooling by as little as 0.5 °C; others did not detect cooling down to 0 °C (cut-off for CDT: −32 °C). Data points for test–retest on 2 days (Fig. 2A), testing by 2 different observers within a day (Fig. 2B), and testing in the morning or afternoon (Fig. 2C) all scattered symmetrically around the unity–slope line, indicating absence of bias. All 3 correlation coefficients were highly significant and indicated that more than 70% of the data variance was systematic, ie, presumably because of systematic differences between patients.

Table 2 shows mean values, correlation coefficients, and paired *t* tests for all data pairs and all QST parameters. TR-R of the vibration threshold was best ($r = 0.93$), followed by mechanical detection threshold and mechanical pain sensitivity ($r = 0.90$ both). The IO-R vibration detection threshold and mechanical detection threshold were similarly correlated ($r = 0.89$) between the assessments. The WUR was the parameter with the least Pearson correlation coefficient in TR-R ($r = 0.67$) and IO-R ($r = 0.56$). It was not assessable in all patients (TR-R: n = 88/120, IO-R: n = 86/120), because the single 256-mN pinprick stimulus was perceived as not painful by some patients. Mean correlations between computer-applied parameters (CDT, WDT, TSL, CPT, and HPT) and hand-held parameters (PPT, MPT, MPS, WUR, MDT, and VDT) were not different (IO-R: $r = 0.85$ vs $r = 0.82$; TR-R: $r = 0.86$ vs $r = 0.87$, both NS). Among the parameters assessed by the Spearman correlation, we found consistently high correlations for dynamic mechanical allodynia (TR-R: $r = 0.87$; IO-R: $r = 0.79$) and low correlations for paradoxical heat sensations (TR-R: 0.39; IO-R: 0.44).

The overall mean test–retest and interobserver reliability (without DMA and PHS) in the affected area was $r = 0.86$ (TR-R; range 0.67–0.93) and $r = 0.83$ (IO-R; range 0.56–0.89). The difference between TR-R and IO-R was small but statistically significant ($P < .01$) contributing about 5% of additional variance due to the effect of the different examiners in IO-R (mean $r^2 = 0.74$

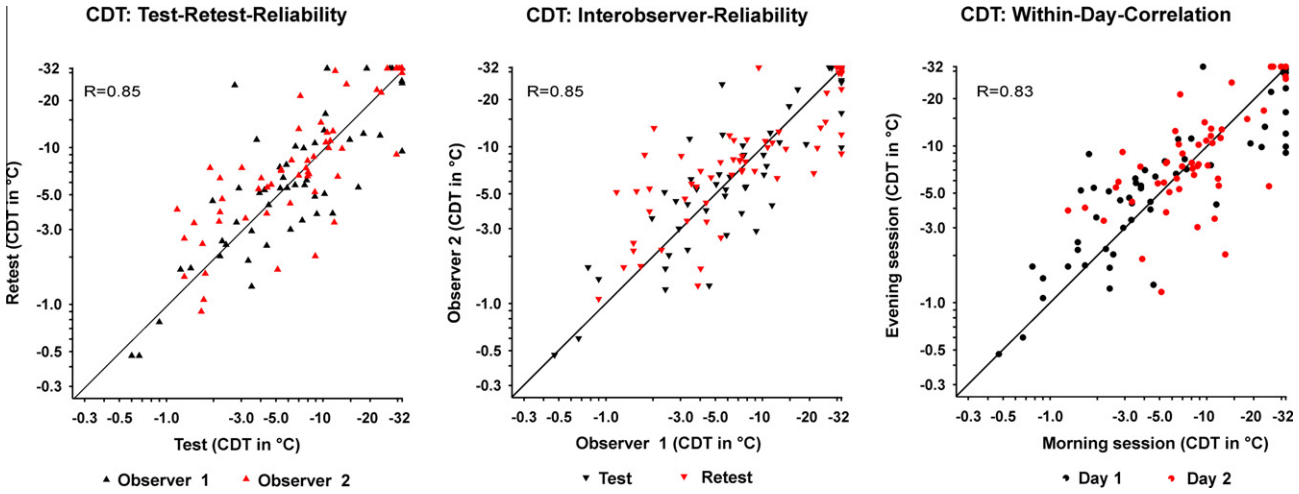

Fig. 2. Test–retest, interobserver, and within-day correlations for cold detection threshold (CDT) in the affected area (ie, test area). CDT is expressed as difference from baseline temperature (32 °C). Solid line represents a slope of 1.

vs 0.69). Observer bias had no influence on most QST parameters. The significant difference in mean values of MPS, WUR, and VDT between observer 1 and 2 (Table 2) across all test centres is difficult to interpret, as the assignment of numbers to observers across the 4 centres was entirely arbitrary. The observed differences cannot be a result of test sequence, because each observer was the first examiner for half of the patients and the second examiner for the other half.

3.4. Circadian rhythm effects (morning vs afternoon session)

Paired *t* tests revealed no mean value differences for any of the 13 QST parameters for test–retest assessment. Mean reliability ($r = 0.83$) was similar to the IO-R (Table 2).

3.5. Comparison of reliability between centres

Mean IO-R across all QST parameters was not different among the 4 participating centres (ANOVA: $F_{1,3} = 0.12$, $P = .946$). The mean correlations of the QST parameters of each centre ranged from 0.828 to 0.848 and are presented in Fig. 3.

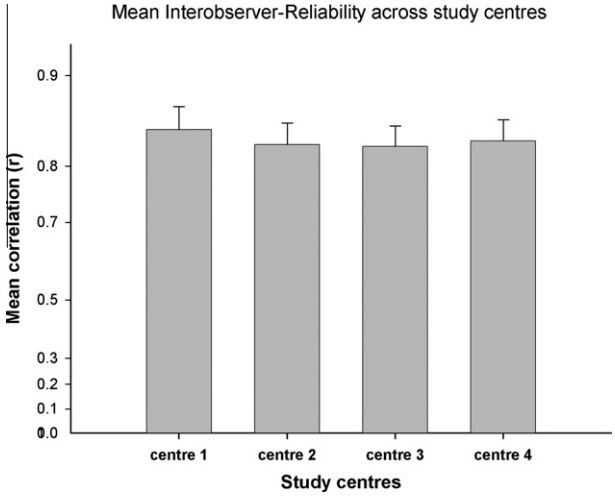

Fig. 3. Comparability of interobserver reliability among study centres.

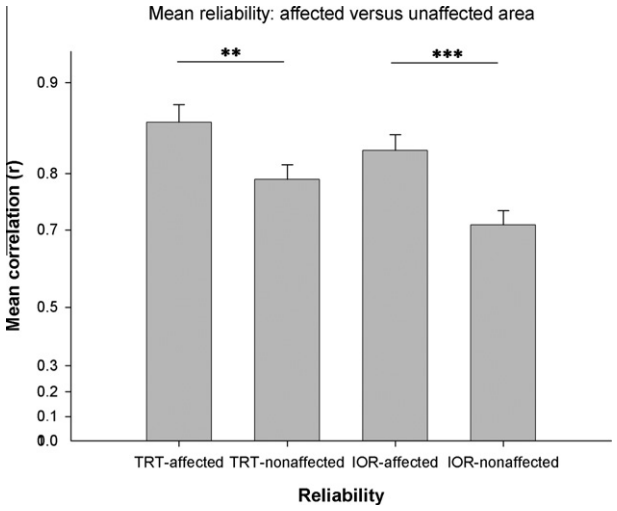

Fig. 4. Comparison of interobserver reliability and test–retest reliability in the affected vs the nonaffected area. N = 43 patients with unaffected control area. $^{**}P < .01$, $^{***}P < .001$.

3.6. Comparison of reliability in the affected TA vs the unaffected CA

QST data obtained at the affected TA was highly variable across patients but was consistent between the 2 assessments (high TR-R) and observers (high IO-R, see above). To test the influence of disease-related variance, we compared TR-R and IO-R in the TA vs the CA in the subgroup with unaffected CA (n = 43). Of interest, both IO-R and TR-R displayed significantly better correlation in the clinically affected TA compared to the unaffected CA (TR-R: TA: $r = 0.86 \pm 0.07$ vs CA: $r = 0.79 \pm 0.06$; IO-R: TA: $r = 0.83 \pm 0.07$ vs CA: $r = 0.71 \pm 0.06$, $P < .01$ each of 86 data pairs) (Fig. 4).

4. Discussion

We performed a multi-centre study on reliability of QST findings. We recruited a study population that was heterogeneous with respect to underlying diseases, test areas, and sensory abnormalities (including 47 patients with and 13 patients without ongoing pain). The main findings of the present study are as follows: (a)

Test–retest reliability was high except for WUR and PHS, and there was no systematic difference across days. (b) Interobserver reliability was high but significantly lower than test–retest reliability; however, observer bias contributed only 5% of variance. (c) Mean reliability over all QST parameters did not differ among centres, all of which had been trained in using the same protocol. (d) Reliability was higher in the clinically affected TA compared to the nonaffected CA. (e) There was no difference between assessments in the morning vs afternoon.

4.1. Test–retest reliability

The analysis of mean value difference between test sessions revealed that there were no significant differences in QST measures. This means that the correlation coefficients in fact can be interpreted as measures of QST reliability. The high correlations, explaining on average 74% of the variance, reflect systematic differences between patients because of their disease states; hence, these correlations were higher in the affected area than the unaffected control area. Reporting bias by the patient and any systematic effect of the observer would also have contributed to the systematic variance. The 26% unexplained variance may be due to random errors by observer and patient, as well as fluctuations in disease states. Healthy subject data may overestimate test–retest reliability of QST in the clinical setting because of lack of disease fluctuation, as suggested in a study on vibration perception thresholds in normal subjects and patients with diabetic polyneuropathy [8]. However, healthy subject data may also underestimate test–retest reliability of QST because of a lack of systematic variance by disease. Our data suggest that the latter effect prevails, as we found higher reliability in clinically affected than in unaffected areas. In addition, although not explicitly tested, an attentional shift to painful or deafferented body regions might also contribute to the higher reliability in clinically affected areas [46].

Both WUR and PHS appear to be fragile phenomena, significantly less stable than other sensory signs (both, loss or gain). Notably, DMA, although mostly mild in our study, was highly reproducible (see also Samuelsson et al. [45]). For WUR, there simply may not have been enough systematic variance because of missing data. In very sensitive patients, WUR data were missing when the procedure had to be discontinued because of too strong pain. On the other hand, in patients with severe deafferentiation, the initial stimulus was not experienced as painful or was not detected at all and thus, it was not possible to calculate the WUR (not allowed to divide by zero). Our findings of poor reliability for this parameter are consistent with a recent study using the same QST protocol (DFNS) in the orofacial region in healthy subjects [41].

Independent of the mode of assessment, high TR-reliability has been consistently reported in the literature for mechanical QST measures such as pressure pain thresholds (eg, intraclass correlation [ICC] >0.7) and vibration detection thresholds (eg, ICC >0.55) for both healthy individuals (PPT: [2,7,11,12,18,38,42,50]; VDT: [8,14,17,19,28,39]) and patient populations (PPT: [7], VDT: [8,10,14,17,19,20,31]). Mechanical detection (MDT) was also found to be a reliable measure in the assessment of sensory function (eg, in diabetic neuropathy) using Semmes–Weinstein monofilaments [36,53]. The use of hand held filaments has, however, been criticized, mainly because of potential variability of the application procedure (degree of filament indention or accidental movement of the hand) and concerns that the mechanical filament characteristics may change over time [3]. The good TR-R for MDT assessed with von Frey filaments in the present study suggests that training and standardization of the testing procedure substantially enhance the performance in QST. In a recent study on QST in the orofacial region in healthy subjects using the same protocol as in the present study, TR-R was, however, considerably lower (ICC <0.4), which may be partly explained by a floor effect because of the higher sensitivity in the face, and by limiting the study to healthy subjects [41].

Studies on reliability of measures of mechanical pinprick-evoked pain (MPT and MPS) or DMA capable of indicating central sensitization processes showed consistently high intersession reliability in patients [37,45] and healthy controls [37,41].

More heterogeneous results were found concerning reliability of thermal detection and pain thresholds (see Siao and Cros [48] for review). Reliability of thermal testing was regarded to vary considerably depending on the applied algorithm: good TR reliability was found using reaction time exclusive methods in patients and healthy subjects [4,9,32,33], whereas studies using a reaction-time-sensitive method of limits procedure reported poor reliability in normal subjects and patients [21,56]. This contrasts to the present study, in which generally high correlations for thermal parameters were found using the methods of limits, suggesting that this test algorithm is not a principal limiting factor concerning the reliability of thermal testing.

4.2. Interobserver reliability

Compared with TR-R, in IO-R, the different examiner as a further possible source of systematic variance has to be considered. Nevertheless, in our study, IO-R was only marginally less across all parameters than TR-R, with just adding 5% of additional variance. In addition, we found no differences in reliability in computer-assisted (thermal) parameters compared to hand-held stimuli altogether suggesting that QST-examiners do not introduce relevant additional variance. This is an important prerequisite for the use of QST as a diagnostic tool in multi-centre studies.

However, in other studies, IO-R has been described as substantially lower than TR-R for different modalities such as pressure pain [38,50] and vibration [2,39], suggesting that a high level of standardization as well as an extended training session by experienced examiners may be essential to minimize interobserver effects. This is supported by a study on pressure pain were standardization of stimulus application and training of the examiners markedly raised interobserver correlation so that it was comparable to intraobserver correlation [18].

4.3. Multi-centre approach

There are only a few multi-centre studies on QST in patients [5,27,52] or healthy controls [27], consistently showing poor reproducibility of thermal [27,52] and vibratory [5,27,52] perception thresholds. This result was attributed to the multi examiner effect. In the present study, however, there are no differences in mean correlations among the 4 study centres (IO-R), again suggesting that the strictly standardized protocol and training of the examiners may overcome potential limitations of the multi-centre approaches.

4.4. Technical considerations

As there is no consensus on how reproducibility in QST should be assessed [47], different approaches comprise correlation techniques (Pearson correlation coefficient, intraclass correlation coefficient), coefficients of variation, and the calculation of repeatability factors [12,27,48,55].

Because of its widespread use, we used correlation analyses (Pearson or Spearman) to describe the strength of the relationship between measurements [55]. However, because these correlative approach analyses do not describe the agreement between measurements, we also analysed mean differences and found no systematic differences in TR-R, IO-R, and daytime effects.

The present study investigated only short-term reliability within consecutive days. Additional studies are needed on reproducibility of QST in the longer time range such as weeks or months, a time interval relevant for clinical studies or patient follow-up. However, with a standardized procedure and only marginal variance between different examiners, it is likely that sensory changes in long-term studies can be interpreted as actual clinical change rather than variability of the QST procedure. This hypothesis is supported by a recent study on the natural course of sensory function after herpes zoster, in which thermal detection and pain thresholds showed consistently low within- and between-subject coefficients of variation over an interval of 6 months [40].

5. Summary and conclusions

In conclusion, QST according to the DFNS protocol is a diagnostic instrument with good interobserver and test–retest reliability. To the extent that our results were better than in previously published studies, they support the introduction of strict quality criteria and certification procedures [26]. The higher correlation over the symptomatic test area is most likely due to disease-related systematic variance, indicating that QST reliability should be studied in patients rather than in healthy subjects.

Conflict of interest statement

There are no financial or other relationships that might lead to a conflict of interest.

Acknowledgments

We thank A. Binder, M. Freyer, W. Magerl, R. Rolke for the supervision of observers.

This work was supported by the DFNS (BMBF Grants 01EM0506, 01EM0101, 01EM0118, 01EM0504, and 01EM0502).

References

[1] Anderson TW. R.A. Fisher and multivariate analysis. Stat Sci 1996;11:20–34.
[2] Antonaci F, Sand T, Lucas GA. Pressure algometry in healthy subjects: inter-examiner variability. Scand J Rehabil Med 1998;30:3–8.
[3] Arezzo J, Bolton C, Boulton A, Casey K, Dyck P, Cornblath D, Fields H, Hansson P, Greenspan J, Harati Y, Jamal G, Kennedy W, Kenshalo D, Light A, Lindblom U, O'Brien P, Ochoa J, Zimmermann I. Quantitative sensory testing: a consensus report from the Peripheral Neuropathy Association. Neurology 1993;43:1050–2.
[4] Arezzo JC, Schaumburg HH, Laudadio C. Thermal sensitivity tester. Device for quantitative assessment of thermal sense in diabetic neuropathy. Diabetes 1986;35:590–2.
[5] Armstrong FM, Bradbury JE, Ellis SH, Owens DR, Rosen I, Sonksen P, Sundkvist G. A study of peripheral diabetic neuropathy. The application of age-related reference values. Diabet Med 1991;8(Spec. No.):S94–9.
[6] Backonja MM, Walk D, Edwards RR, Sehgal N, Moeller-Bertram T, Wasan A, Irving G, Argoff C, Wallace M. Quantitative sensory testing in measurement of neuropathic pain phenomena and other sensory abnormalities. Clin J Pain 2009;25:641–7.
[7] Bernhardt O, Schiffman EL, Look JO. Reliability and validity of a new fingertip-shaped pressure algometer for assessing pressure pain thresholds in the temporomandibular joint and masticatory muscles. J Orofac Pain 2007;21:29–38.
[8] Bertelsmann FW, Heimans JJ, van Rooy JC, Heine RJ, van der Veen EA. Reproducibility of vibratory perception thresholds in patients with diabetic neuropathy. Diabetes Res 1986;3:463–6.
[9] Bertelsmann FW, Heimans JJ, Weber EJ, van der Veen EA, Schouten JA. Thermal discrimination thresholds in normal subjects and in patients with diabetic neuropathy. J Neurol Neurosurg Psychiatry 1985;48:686–90.
[10] Botez SA, Liu G, Logigian E, Herrmann DN. Is the bedside timed vibration test reliable? Muscle Nerve 2009;39:221–3.
[11] Brennum J, Kjeldsen M, Jensen K, Jensen TS. Measurements of human pressure-pain thresholds on fingers and toes. Pain 1989;38:211–7.
[12] Cathcart S, Pritchard D. Reliability of pain threshold measurement in young adults. J Headache Pain 2006;7:21–6.
[13] Claus D, Hilz MJ, Neundorfer B. Thermal discrimination thresholds: a comparison of different methods. Acta Neurol Scand 1990;81:533–40.
[14] Claus D, Mustafa C, Vogel W, Herz M, Neundorfer B. Assessment of diabetic neuropathy: definition of norm and discrimination of abnormal nerve function. Muscle Nerve 1993;16:757–68.
[15] Cruccu G, Sommer C, Anand P, Attal N, Baron R, Garcia-Larrea L, Haanpaa M, Jensen TS, Serra J, Treede RD. EFNS guidelines on neuropathic pain assessment: revised 2009. Eur J Neurol 2010;17:1010–8.
[16] Cruccu G, Truini A. Tools for assessing neuropathic pain. PLoS Med 2009;6:e1000045.
[17] de Neeling JN, Beks PJ, Bertelsmann FW, Heine RJ, Bouter LM. Sensory thresholds in older adults: reproducibility and reference values. Muscle Nerve 1994;17:454–61.
[18] Delaney GA, McKee AC. Inter- and intra-rater reliability of the pressure threshold meter in measurement of myofascial trigger point sensitivity. Am J Phys Med Rehabil 1993;72:136–9.
[19] Dyck PJ, Karnes JL, Gillen DA, O'Brien PC, Zimmerman IR, Johnson DM. Comparison of algorithms of testing for use in automated evaluation of sensation. Neurology 1990;40:1607–13.
[20] Dyck PJ, Kratz KM, Lehman KA, Karnes JL, Melton 3rd LJ, 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;41:799–807.
[21] Fagius J, Wahren LK. Variability of sensory threshold determination in clinical use. J Neurol Sci 1981;51:11–27.
[22] Fillingim RB, Maddux V, Shackelford JA. Sex differences in heat pain thresholds as a function of assessment method and rate of rise. Somatosens Mot Res 1999;16:57–62.
[23] Geber C, Klein T, Rolke R, Magerl W, Dieterich M, Birklein F, Treede RD. Test/retest- and interobserver-reliability in quantitative sensory testing according to the protocol of the German Network on Neuropathic Pain (DFNS). Eur J Pain 2007;11:S87.
[24] Geber C, Baumgartner U, Schwab R, Muller H, Stoeter P, Dieterich M, Sommer C, Birklein F, Treede RD. Revised definition of neuropathic pain and its grading system: an open case series illustrating its use in clinical practice. Am J Med 2009;122(10 Suppl.):S3–S12.
[25] Geber C, Magerl W, Fondel R, Fechir M, Rolke R, Vogt T, Treede RD, Birklein F. Numbness in clinical and experimental pain—a cross-sectional study exploring the mechanisms of reduced tactile function. Pain 2008;139:73–81.
[26] Geber C, Scherens A, Pfau D, Nestler N, Zenz M, Tolle T, Baron R, Treede RD, Maier C. Procedure for certification of QST laboratories.. Schmerz 2009;23:65–9.
[27] Gelber DA, Pfeifer MA, Broadstone VL, Munster EW, Peterson M, Arezzo JC, Shamoon H, Zeidler A, Clements R, Greene DA, Porte D, Laudadio C, Bril V. Components of variance for vibratory and thermal threshold testing in normal and diabetic subjects. J Diabetes Complications 1995;9:170–6.
[28] Goldberg JM, Lindblom U. Standardised method of determining vibratory perception thresholds for diagnosis and screening in neurological investigation. J Neurol Neurosurg Psychiatry 1979;42:793–803.
[29] Greenspan JD, McGillis SLB. Stimulus features relevant to the perception of sharpness and mechanically evoked cutaneous pain. Somatosens Mot Res 1991;8:137–47.
[30] Hansson P, Backonja M, Bouhassira D. Usefulness and limitations of quantitative sensory testing: clinical and research application in neuropathic pain states. Pain 2007;129:256–9.
[31] Hilz MJ, Axelrod FB, Hermann K, Haertl U, Duetsch M, Neundorfer B. Normative values of vibratory perception in 530 children, juveniles and adults aged 3–79 years. J Neurol Sci 1998;159:219–25.
[32] Jamal GA, Hansen S, Weir AI, Ballantyne JP. An improved automated method for the measurement of thermal thresholds. 1. Normal subjects. J Neurol Neurosurg Psychiatry 1985;48:354–60.
[33] Jamal GA, Weir AI, Hansen S, Ballantyne JP. An improved automated method for the measurement of thermal thresholds. 2. Patients with peripheral neuropathy. J Neurol Neurosurg Psychiatry 1985;48:361–6.
[34] Klauenberg S, Maier C, Assion HJ, Hoffmann A, Krumova EK, Magerl W, Scherens A, Treede RD, Juckel G. Depression and changed pain perception: hints for a central disinhibition mechanism. Pain 2008;140:332–43.
[35] Magerl W, Wilk SH, Treede RD. Secondary hyperalgesia and perceptual wind-up following intradermal injection of capsaicin in humans. Pain 1998;74:257–68.
[36] Maser RE, Nielsen VK, Bass EB, Manjoo Q, Dorman JS, Kelsey SF, Becker DJ, Orchard TJ. Measuring diabetic neuropathy. Assessment and comparison of clinical examination and quantitative sensory testing. Diabetes Care 1989;12:270–5.
[37] Nixdorf D, Hemmaty A, Look J, Schiffman E, John M. Electric toothbrush application is a reliable and valid test for differentiating temporomandibular disorders pain patients from controls. BMC Musculoskel Disord 2009;10:94. doi:10.1186/1471-2474-10-94.
[38] Nussbaum EL, Downes L. Reliability of clinical pressure-pain algometric measurements obtained on consecutive days. Phys Ther 1998;78:160–9.
[39] Peters EW, Bienfait HM, de Visser M, de Haan RJ. The reliability of assessment of vibration sense. Acta Neurol Scand 2003;107:293–8.
[40] Petersen KL, Rowbotham MC. Natural history of sensory function after herpes zoster. Pain 2010;150:83–92.
[41] Pigg M, Baad-Hansen L, Svensson P, Drangsholt M, List T. Reliability of intraoral quantitative sensory testing (QST). Pain 2010;148:220–6.
[42] Reeves JL, Jaeger B, Graff-Radford SB. Reliability of the pressure algometer as a measure of myofascial trigger point sensitivity. Pain 1986;24:313–21.

[43] Rolke R, Baron R, Maier C, Tolle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Botefur IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W, Maihofner 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;123:231–43.
[44] Rolke R, Magerl W, Campbell KA, Schalber C, Caspari S, Birklein F, Treede RD. Quantitative sensory testing: a comprehensive protocol for clinical trials. Eur J Pain 2006;10:77–88.
[45] Samuelsson M, Leffler AS, Johansson B, Hansson P. On the repeatability of brush-evoked allodynia using a novel semi-quantitative method in patients with peripheral neuropathic pain. Pain 2007;130:40–6.
[46] Seminowicz DA, Davis KD. A re-examination of pain-cognition interactions: implications for neuroimaging. Pain 2007;130:8–13.
[47] Shy ME, Frohman EM, So YT, Arezzo JC, Cornblath DR, Giuliani MJ, Kincaid JC, Ochoa JL, Parry GJ, Weimer LH. Quantitative sensory testing: report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. Neurology 2003;60:898–904.
[48] Siao P, Cros DP. Quantitative sensory testing. Phys Med Rehabil Clin North Am 2003;14:261–86.
[49] Slugg RM, Meyer RA, Campbell JN. Response of cutaneous A- and C-fiber nociceptors in the monkey to controlled-force stimuli. J Neurophysiol 2000;83:2179–91.
[50] Takala EP. Pressure pain threshold on upper trapezius and levator scapulae muscles. Repeatability and relation to subjective symptoms in a working population. Scand J Rehabil Med 1990;22:63–8.
[51] Treede RD, Jensen TS, Campbell JN, Cruccu G, Dostrovsky JO, Griffin JW, Hansson P, Hughes R, Nurmikko T, Serra J. Neuropathic pain: redefinition and a grading system for clinical and research purposes. Neurology 2008;70:1630–5.
[52] 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;10:933–9.
[53] Valk GD, de Sonnaville JJ, van Houtum WH, Heine RJ, van Eijk JT, Bouter LM, Bertelsmann FW. The assessment of diabetic polyneuropathy in daily clinical practice: reproducibility and validity of Semmes Weinstein monofilaments examination and clinical neurological examination. Muscle Nerve 1997;20:116–8.
[54] Walk D, Sehgal N, Moeller-Bertram T, Edwards RR, Wasan A, Wallace M, Irving G, Argoff C, Backonja MM. Quantitative sensory testing and mapping: a review of nonautomated quantitative methods for examination of the patient with neuropathic pain. Clin J Pain 2009;25:632–40.
[55] Yarnitsky D. Quantitative sensory testing. Muscle Nerve 1997;20:198–204.
[56] Yarnitsky D, Sprecher E. Thermal testing: normative data and repeatability for various test algorithms. J Neurol Sci 1994;125:39–45.