Test–retest reliability of thermal quantitative sensory testing on two sites within the L5 dermatome of the lumbar spine and lower extremity

Q1 I.A. Knutti, M.R. Suter, E. Opsommer

Bern University of Applied Sciences (BFH), Health, Murtenstrasse 10, 3008 Bern, Switzerland
University of Health Sciences (HESAV), University of Applied Sciences and Arts Western Switzerland (HES-SD), Av. de Beaumont 21, 1011 Lausanne, Switzerland
Pain Center, Department of Anesthesiology, University Hospital Center (CHUV) and University of Lausanne (UNIL), Rue du Bugnon 46, 1011 Lausanne, Switzerland

HIGHLIGHTS
- Quantitative sensory testing (QST) is used to assess sensory disturbances.
- These findings showed excellent intra-rater reliability of thermal QST on the feet.
- Intra-rater reliability of thermal QST on the lumbar spine was fair to excellent.

ARTICLE INFO
Article history:
Received 15 March 2014
Received in revised form 25 May 2014
Accepted 11 July 2014
Available online xxx

Keywords:
Quantitative sensory testing
Reproducibility of results
Pain threshold
Pain measurement

ABSTRACT
Introduction
Quantitative sensory testing (QST) is widely used in human research to investigate the integrity of the sensory function in patients with pain of neuropathic origin, or other causes such as low back pain. Reliability of QST has been evaluated on both sides of the face, hands and feet as well as on the trunk (Th3-L3). In order to apply these tests on other body parts such as the lower lumbar spine, it is important first to establish reliability on healthy individuals. The aim of this study was to investigate intra-rater reliability of thermal QST in healthy adults, on two sites within the L5 dermatome of the lumbar spine and lower extremity. Methods: Test–retest reliability of thermal QST was determined at the L5-level of the lumbar spine and in the same dermatome on the lower extremity in 30 healthy persons under 40 years of age. Results: Were analyzed using descriptive statistics and intraclass correlation coefficient (ICC). Values were compared to normative data, using Z-transformation. Results: Mean intrainividual differences were small for cold and warm detection thresholds but larger for pain thresholds. ICC values showed excellent reliability for warm detection and heat pain threshold, good-to-excellent reliability for cold pain threshold and fair-to-excellent reliability for cold detection threshold. ICC had large ranges of confidence interval (95%). Conclusion: In healthy adults, thermal QST on the lumbar spine and lower extremity demonstrated fair-to-excellent test–retest reliability.

© 2014 Published by Elsevier Ireland Ltd.

1. Introduction
Quantitative sensory testing (QST) is a method used to test all somatosensory submodalities (i.e. touch, vibration, temperature or pain) with different kinds of calibrated stimuli, to examine the presence of negative or positive sensory signs [1]. QST can be applied to investigate the integrity of the sensory function in order to define and classify pathologies, to analyze pathogenesis or to evaluate changes in diseases [2]. Over the past decades, there has been an increasing interest in QST in clinical and research settings, for example to determine the conduction velocity of peripheral nerve fibres [3] or to assess a treatment’s effectiveness [4]. QST is a psychophysical measurement, relying on the subjective perception of a physical stimulus [5]. Therefore, reliability needs to be assessed with a very rigorous methodology. In their systematic review, Moloney and colleagues [6] reported large variability in

Please cite this article in press as: I.A. Knutti, et al., Test–retest reliability of thermal quantitative sensory testing on two sites within the L5 dermatome of the lumbar spine and lower extremity, Neurosci. Lett. (2014), http://dx.doi.org/10.1016/j.neulet.2014.07.023
methodological quality of published studies with poor-to-excellent reliability of thermal QST. The German Research Network on Neuropathic Pain (DFNS) published a standardized QST protocol and normative data for quantitative sensory evaluations [7]. Under this protocol, that QST battery was evaluated bilaterally on the face, hands and feet using the method of limits, with reliability then assessed on these different sites [8]. The availability of this standardized protocol tends to improve reliability [9]. Indeed, before using QST on other body parts, its reliability needs to be as assessed recently shown by Pfau and colleagues [10] for the upper (Thoracic(Th)2-Th8) and the lower back (Th10-Lumbal(L)3). Among the diverse sensory submodalities described in the DFNS standardized protocol, the thermal modalities include cold and warm detection threshold (CDT, WDT), paradoxical heat sensation (PHS) as well as cold and heat pain perception threshold (CPT, HPT) [8]. All thresholds are given in °C (continuous scale) and can be tested using a thermostating device by trained investigators.

QST has mostly been used in patients with neuropathic pain [2] but also in different musculoskeletal pathologies such as low back pain (LBP) [1]. People with LBP may have sensory disturbances in the back with or without radicular symptoms in the lower limb. QST has been used in some studies to assess these disturbances [11–13]. Nevertheless, stimulation sites were not combining two sites within the same dermatome of the lumbar spine and the lower extremity in these studies and reliability of thermal QST has, to our knowledge, never been assessed on the lumbar paravertebral area, corresponding to the L5-dermatome.

Given the increased use of QST to assess sensitivity in the lumbar spine in patients with LBP [13], reliability of QST–measurement in this region needs to be assessed. Furthermore, testing a second site on the same dermatome will allow further studies to be performed on patients. Therefore, the aim of this study was to determine test–retest reliability of thermal QST on two sites within the same dermatome (L5) of the lumbar spine and the lower extremity in healthy adults younger than 40 years of age, according to the protocol of the DFNS [8]. After examining reliability, data obtained on the lower extremity were compared to the normative values of the DFNS for the same site [7] and then to the data obtained in our study on the lumbar spine within the same dermatome.

2. Methods

Healthy volunteers were recruited from the staff and student population of the University of Health Sciences, Physical Therapy Department and the University Hospital Centre (CHUV), Lausanne, through an e-mail campaign. Inclusion criteria were: (1) good health status with no lower-back pain or lower-limb pain, (2) age less than 40 years, and (3) ability to read and speak French. Participants had to fill out the French–translated “Delphi definitions of Lower–Back Pain Prevalence” (DOLBaPP) – questionnaire [14] to ensure that they were not suffering from lower-back or lower quadrant pain. Other exclusion criteria were the following: diabetes, endocrine dysfunction, cognitive disorders; spinal pain, neurological or rheumatologic disorders and known pregnancy. The first 30 volunteers meeting all inclusion criteria were included in the study (non-probabilistic voluntary sample). Subjects all confirmed not being on any pain medication. All volunteers signed a written informed consent form. The study was approved by the local Ethics Committee, and consistent with the Declaration of Helsinki.

2.1. Experimental procedure

Tests were conducted in a quiet room at the CHUV, with an ambient temperature between 20 and 22 °C. Subjects were assessed on two occasions at the same time of day, within a one-week interval. Tests were performed bilaterally on the dorsum of both feet and on the lumbar paravertebral area (L5 dermatome). The stimulation sites were randomized (Microsoft Excel 2008, version 12.3.6) for each participant and each measurement session, first to define the site (lumbar or foot) and then the side (left or right). Prior to both sessions, a demonstration of the procedure was performed on the left hand. Tests were done using a Neuro Sensory Analyzer TSA-II (Medoc, Israel). A Peltier thermode (16 × 16 mm) was attached directly on the skin of the tested area. CDT and WDT were measured first. The number of PHS was then determined during the thermal sensory limen (TSL) protocol of alternating cold and warm stimuli. CPT and HPT were finally recorded. The baseline temperature was set at 32 °C and increased or decreased at a rate of 1 °C/s. To prevent thermal injury, cut-off levels were set at 50 °C and 0 °C. Standardized instructions were read out to each subject before testing [8]. Tests were conducted under the same conditions as in clinical practice.

All measurements were performed in the same protocolled way in both sessions, by the same trained observer (physiotherapist), using the same equipment. Prior to the beginning of the study, the investigator used exactly the same protocol as that used for participants on 10 persons who were not included in the study. Volunteers were blinded to their own prior results. The Quality Appraisal for Reliability Studies (QAREL)-Checklist [15] was used to guarantee optimal methodological rigour. Guidelines for reporting reliability and agreement studies (GRRAS) in the medical field have been followed [16].

2.2. Statistical analysis

All statistical calculations were performed using SPSS v.21 (IBM, SPSS, Inc., Chicago, IL, USA), except t-test statistics that were done using the internet-based statistical software Simple Interactive Statistical Analysis (SISA) URL: http://www.quantitativeskills.com/sisa/ (accessed 2014, May 10), as proposed by Magee and colleagues [17] and Bland–Altman plots that were realized with MedCalc Software v.13.2 (Ostend, Belgium). Reliability of measures was assessed with the intraclass correlation coefficient for average measurement (ICCagreement) with 95% confidence interval (CI) for each modality. Sample size was estimated as follows: with a sample size of 30 and an expected ICC of 0.85, the lower boundary of the 95% CI would still be above 0.7 for at least 80% of all ICC calculated. This is accepted as a sufficiently high level of reliability [18], Kolmogorov–Smirnov test was used to assess normality of data distribution.

2.3. Test–retest reliability

Means and standard deviations (SD) of three consecutive measurements were calculated for CDT, WDT, CPT and HPT for each tested area in both sessions. In addition, mean intra-individual differences (MID) were calculated for each pair of data sets. PHS was made to follow the DFNS protocol [8] but was not further developed in this study. Calculations of relative reliability were done using a two-way random effects analysis of variance (ANOVA) model, with absolute agreement. Obtained ICC can range from 0 (no correlation) to 1 (perfect correlation). Strength of the correlation was interpreted as follows: ICC ≤ 0.40 is considered as poor, 0.40–0.59 as fair, 0.60–0.75 as good and >0.75 as excellent [19]. Bland–Altman plots were created to determine absolute reliability.

2.4. Comparison with normative data

Data obtained from the lower extremity were compared to corresponding normative data for the feet (L5 dermatome) from the DFNS [7] using a Z-transformation for each parameter.

Please cite this article in press as: LA. Knutti et al., Test–retest reliability of thermal quantitative sensory testing on two sites within the L5 dermatome of the lumbar spine and lower extremity, Neurosci. Lett. (2014), http://dx.doi.org/10.1016/j.neulet.2014.07.023
In the absence of side difference in healthy individuals, both right and left of the second measurement session were used for Z-transformations [7]. We selected the second session, because in the first one, there was an outlier (according to the Grubbs’ test) who did not feel cold sensation before the cut-off temperature (0°C).

For CDT and WDT, difference to baseline was calculated before log-transformation, as it was performed by the German group [7]. For CPT and HPT, absolute values were used. Z-scores were calculated as follow: \( Z = \frac{\text{value}_{\text{participant}} - \text{mean}_{\text{reference}}}{\text{SD}_{\text{reference}}} \). Z-transformations were done separately for men and women [2]. Plus and minus signs were adjusted so that a plus sign signified a gain of sensory function and a minus sign a loss of sensory function. Mean Z-values were compared to the expected mean values of 0 ± 0.25 and to the expected SD of 1 ± 0.1. Comparison between Z-transformed data of this study and reference data was performed by a t-test statistic [17].

2.5. Comparison within the same dermatome

Data obtained in this study on the two sites within the L5 dermatome were Z-transformed and compared to each other to investigate differences. For CDT and WDT, difference to baseline was calculated before log-transformation. Z-Scores were calculated as follow: \( Z = \frac{\text{value}_{\text{lumbar spine}} - \text{mean}_{\text{lower extremity}}}{\text{SD}_{\text{lower extremity}}} \).

3. Results

Thirty healthy volunteers (10 men and 20 women; mean age: 27.1 years; SD: 5.4 years; range: 20–39) took part in this study.

Most of the samples were normally distributed but some only after log transformation. Moreover, it has been shown that QST-parameters were normally (or log-normal) distributed in the normal population [7]. Using parametric tests allowed us to compare our results with those in the literature [17].

Table 1 shows the mean of the three consecutive measurements for each modality and each of both measurement sessions, with SD, range of data for each session, MID, ICC and bias with 95% limits of agreement for CDT, WDT, CPT and HPT.

For every modality, ICC was good-to-excellent on all sites, except for CDT on the right lumbar spine, where ICC showed fair reliability. MID were small for CDT and WDT but larger for pain thresholds, predominantly for CPT with the largest SD. For CPT, ICC indicated good-to-excellent reliability. However, lower bounds showed a large variability in data. For CDT, ICC was excellent for all tested area except for the right lumbar spine, where reliability was interpreted as fair. On that site, one participant did not feel cold sensation before the 0°C cut-off in the first measurement session.

Bland–Altman plots showed a low bias for CDT and WDT and a larger bias for pain thresholds, mainly for CPT. Limits of agreements were also the highest for CPT, independently of the tested sites. Fig. 1 illustrates a Bland–Altman plot for CDT on the right lumbar spine, which showed fair reliability. In Fig. 1A, each participant is represented; in Fig. 1B, the same Bland–Altman plot is illustrated without the outlier who did not feel cold sensation before the cut-off temperature (0°C) during the first measurement session. According to the Grubbs’ test, the outlier could be excluded, as the chance of finding such a value in a sample of population with normal distribution is very small.

3.2. Comparison with normative data

Using Z-scores, we compared our data of the lower extremity with the normative data of the DFNS [7]. Women were less sensitive in sensing cold and warm (mean Z-score ± SD for CDT: −0.85 ± 0.10, p < 0.001; mean Z-score ± SD for WDT: −1.10 ± 1.04, p < 0.0001), and detected cold pain earlier in our sample than in the reference population (mean Z-score ± SD for CPT: 0.45 ± 0.88, p < 0.05). HPT was similar to reference data (mean Z-score ± SD for HPT: −0.31 ± 0.73, p = 0.11). For men, all modalities were similar to normative values (p > 0.05). Mean Z-scores ± SD were for CDT: 0.10 ± 1.10; for WDT: 0.06 ± 1.16; for CPT: −0.06 ± 1.06 and for HPT: 0.38 ± 1.03.

3.3. Comparison within the same dermatome

For women, comparison between the data obtained in this study from the lumbar spine and data obtained from the lower
with other studies concerning reliability of thermal QST [6]. Best ICC and MID were assessed for WDT, where reliability was interpreted as excellent.

Reliability of lumbar spine data was similar to that of the lower extremity, except for CDT of the right lumbar spine, which had the lowest ICC with the largest CI. It showed only small improvements in MID and ICC with the removal of the outlier. In Fig. 1 we can observe that the 95% CI is much smaller without the outlier and test–retest differences tend to increase as the mean temperature goes away from the normal range as observed by Geber and colleagues [20]. Reliability results obtained in this study are consistent, however, with the existing literature, which states a poor-to-excellent reliability for CDT [6].

There are significant disparities in published studies about reliability of QST and variation in statistical tests used [6]. As explained in Weir [21], ICC is a relative measurement: this means that it depends on the between-subjects variance. If subjects differ a lot from each other, ICC can be high, even if the test–retest variability is large. In our study, between-subjects variance is quite large and ICC results are high. Indeed, reliability could be overestimated. Bland–Altman plots recognize this limitation with the calculation of the limits of agreement. In our study, we can observe a large variability in bias and limits of agreement, mainly for CPT. This limits the utility of CPT despite its good-to-excellent reliability measured by ICC. Furthermore, in the second session, pain thresholds showed less sensitive values, which could suggest some learning effect. This also shows limits of CPT and HPT testing in the clinic in an individual patient due to its very large variability in absolute value and large range.

4.2. Comparison with normative data

The DFNS developed a standardized protocol for the use of QST in clinical or research settings. A database with 180 healthy subjects was created. The obtained data were separated by age and gender since sensitivity thresholds differ by gender and rise with age. In this study, we chose to focus on young adults under 40 years, which corresponds to an age category of the DFNS. One of the main objectives of the DFNS-database was to obtain normative values [7]. This allows us to establish if a value in a patient is inside or outside a defined normal range (95% CI). Z-scores obtained in the present study showed a variation with the normative data of the DFNS, for women. Although participants were described as healthy, it raises the question of validity and usefulness of a norm for all measured sensory modalities, as normative values might vary in different populations. Therefore, in clinical practice, it may be interesting to compare side-to-side results in the same person or assess the evolution of the results in the same person on the same area over time. Indeed, in the present study, the excellent reliability of thermal QST on the lower extremity shows that in the same person, results are consistent over time. To assess local changes using a comparison with the unaffected contralateral area in the same person might be more useful, but generalized changes can only be judged using normative values [8]. In this case, narrower age categories might be considered as proposed by Magerl and associates [17], or local values to have normative values more adapted to each case.

4.3. Comparison within the same dermatome

The data from the lumbar spine were also referenced to the data obtained from the lower extremity using a Z-transformation. Although the two measurement sites follow the same dermatome, changes in thermal sensitivity were observed. Results showed significant differences for CDT in women and a mean SD out of the expected SD of 1 ± 0.1. Women were less sensitive on the lumbar

Fig. 1. Bland–Altman plots for cold detection threshold (CDT) on the right lumbar spine. Limits of agreements (upper and lower 95% confidence interval) are represented by the thin lines and mean differences between the two assessment sessions are represented by the thick line. The upper panel (A) is a plot with all participants and the lower panel (B) without the outlier.

4. Discussion

We here demonstrated fair-to-excellent reliability in thermal QST in the L5 dermatome in the lumbar spine and in the lower extremity. Limits and applicability of the results are discussed below.

4.1. Test–retest reliability of thermal QST

Data of thermal QST showed fair-to-excellent reliability, despite large SD and large variations in the 95% CI. Reliability of CDT on the right lumbar spine was the only one considered as fair. CPT had good reliability on both sides of the lumbar spine. All other modalities had excellent reliability. These results are in accordance
spine for CDT than on the lower extremity. In men, no significant difference within the same dermatome for any modality was observed. However, mean Z-scores were also out of the expected mean values of 0 ± 0.25 for thermal thresholds and CPT and mean SD were out of the expected SD of 1 ± 0.1 for all modalities.

One conclusion from Pfau et al. [10] was that only relying on normative data from head, upper body and lower body, as proposed originally by Rolke et al. [7], is not adequate. Although our differences were not significant for most of the modalities, we observed that mean Z-scores and SD were mostly out of the expected mean of 0 ± 0.25 and SD of 1 ± 0.1, described by Magerl and colleagues [17] to claim that measures are similar. Our power was calculated for the first aim of the study and we might lack the appropriate amount of volunteers to reliably answer the comparison between sites question. We therefore agree with Pfau et al. [10] that relying on data from only three sites (head, upper and lower body) is probably not sufficient and has to be investigated further. The clinical applicability of QST testing on these two different sites could be to differentiate if an abnormal value observed in the lower extremity is due to a peripheral cause (lumbar spine with normal value) or to a more central process (lumbar spine also abnormal).

4.4. Limitations and future research

The whole parameters of the DFNS protocol were not explored [8] and only one type of reliability (intra-rater) was assessed. Other aspects of reliability should be assessed on the lumbar spine, including the whole QST-battery tests following the DFNS-protocol, with a larger sample size and participants of various ages. For comparison with normative values, we had to separate men and women. Our small sample size calculated for the reliability paradigm could partly explain deviations from the reference data. Moreover, our sample might be younger than the <40 sample from Rolke et al. [7] and this emphasizes that normative values might need to be more specific for age and gender and maybe also for other criteria such as ethnicity. Further research in different centres should investigate if similar deviation can be found. This emphasizes that each centre testing a new site should also perform a test at a site where normative values are proposed to validate the new site.

5. Conclusions

Based on our results, the test–retest reliability of thermal QST in healthy people younger than 40 years of age on two sites within the L5 dermatome of the lumbar spine and the lower extremity according to the protocol of the DFNS, showed excellent reliability on each body–part for WDT and HPT. CPT had good–to–excellent reliability. ICC values for CDT were high except for the right lumbar spine where reliability has been shown to be fair. Given the increased interest to assess sensitivity in the lumbar spine in patients with LBP, it is important to have a validated assessment tool. Results of this study were obtained from healthy subjects and encourage further research studies on LBP patients.

Conflicts of interest

There are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Contributors

All authors have substantially contributed to (i) the study conception and design, acquisition data (I.A. Knutti), analysis and interpretation of data; (ii) drafting the article or revising it critically for important intellectual content; and (iii) final approval of the version to be published.

Acknowledgements

We are pleased to thank all participants who agreed to be volunteers for this study. We also thank the staff of the Pain Center, Department of Anesthesiology, CHUV, for providing premises and equipment.

References