Original research study - Test-Retest Reliability and Concurrent Validity of Novel Nerve Testing Device for Measuring Thermal Detection and Thermal Pain Thresholds

Test-Retest Reliability and Concurrent Validity of Novel Nerve Testing Device for Thermal Detection and Thermal Pain Thresholds

Abstract

Thermal threshold testing is important for evaluating the thermal function of small-fiber nerves types C and A-delta. This study investigated the reliability and validity of a novel nerve testing device (NNTD) in evaluating thermal detection and thermal pain thresholds. Test-retest reliability of the NNTD and its concurrent validity compared to the current technology (Medoc TSA-2, Advanced Thermosensory Stimulator, Israel) were investigated among 10 healthy participants. Each participant was tested for the warm detection threshold (WDT), cold detection threshold (CDT), hot pain threshold (HPT) and cold pain threshold (CPT) on the medial forearm with NNTD for two trials and the Medoc TSA-2 for one trial over two consecutive days. Intraclass Correlation Coefficient values, Standard Error of Measurement and Bland Altman plots were calculated for test-retest reliability. One-way ANOVA, and Bland Altman plots were calculated for validity. The test-retest reliability of the NNTD was good for CPT (ICC=0.88), moderate for WDT (ICC=0.545) and HPT (ICC=0.710). The NNTD was valid for both trials of HPT and CPT and one trial for WDT compared to the Medoc TSA-2. In conclusion, the NNTD showed good to moderate reliability and found to be valid compared to the Medoc TSA-2.

Keywords: thermal detection, pain, threshold, reliability, validity, nerve

Introduction

Small diameter nerve fibers which control thermal and pain perception can be evaluated using thermal threshold testing [1,2]. If small-fiber nerve damage is detected early, the patient can benefit by receiving treatment for the underlying cause, possibly preventing further deterioration (Hovaguimian and Gibbons, 2011, Chai et al., 2005, Bachmann et al., 2010). There are four temperature thresholds (warm detection threshold (WDT), cold detection threshold (CDT), hot pain threshold (HPT) and cold pain threshold (HPT)) which should be tested, especially at the diagnostic stage, as different pathologies affect different thresholds [3-7]. Some examples include: diagnosing small fiber neuropathy with WDT and CDT testing [4], detecting nerve damage in patients with diabetes mellitus who had normal nerve conduction studies presented with higher CDT values from baseline [8], using thermal thresholds to distinguish restless leg syndrome which is caused by small fiber neuropathy compared to primary restless leg syndrome. [3]. Unfortunately, in a clinical setting thorough assessment of somatosensory deficits, including temperature sensation, are often not addressed, or addressed inadequately, often leading to poor outcomes for patients [9].

Clinicians are aware of the importance of testing thermal thresholds, however, do not have the appropriate equipment [10,11]. There are several existing devices for thermal testing, which traditionally are limited to a laboratory or research setting due to the initial cost, large size, set-up time, requirement of a laptop and an electricity outlet, and require special training making them impractical for a clinical setting [10-16]. The use of coins and test tubes for CDT and WDT testing, ice for CPT testing and cool tuning forks for CDT testing have all been proposed as alternative thermal testing methods, however, no method is able to test all

four thresholds, the temperature of the objects is unknown and limited research is available on the reliability and validity of these methods [13-15,17-20].

Thus, there is a need to develop a testing method for use in the clinical setting, reliably and validly testing all four thermal thresholds. The goal of this research was to I) determine whether it is feasible to design and build a device which can test all four thermal thresholds and II) analyze the test-retest reliability of the new device and its concurrent validity compared to the Medoc TSA-2 (Advanced Thermosensory Stimulator, Serial Number 1554, 2001, Israel). Due to access, the Medoc TSA-2 was used as an established reliable and valid thermal threshold testing device [14,17,21-25].

Methods

This feasibility quantitative cross-sectional study, investigating test-retest reliability and concurrent validity was approved by an institutional research ethics panel (Ref: 2019-2816). After ethical approval, a convenience sample of 10 healthy participants were recruited through an email. The inclusion criteria for this study were that the participant was healthy and able to come to the university campus for maximum 30 minutes of testing on two consecutive days. To ensure the results of this study represented healthy participants without any known nerve impairments, participants were excluded if they had any known neurological conditions, acute or chronic pain, were taking analgesic medications, had loss of skin sensation or did not speak English [15,25-27].

Thermal Testing Equipment

During this study two devices, each able to test all four thermal thresholds, were tested on participants. The first device was a nerve testing device, a minimum viable prototype (MVP) device which was constructed specifically for this study. Due to pending intellectual property disclosure restrictions, further information on the prototype cannot be discussed at this time. The

second device was a commercially available thermal testing laboratory machine, the Medoc TSA-2 (Advanced Thermosensory Stimulator, Serial Number 1554, 2001, Israel). Both devices had a 30mm x 30mm thermode and a stop button for each participant to press at the appropriate threshold.

Procedure

Each participant was tested twice with the prototype and once with the TSA-2 in a randomized order, based on a computer-generated list of randomized numbers. The participants were tested in a quiet room, sitting in a comfortable chair with their eyes closed and left medial forearm (testing site) undressed and resting on the chair's arm rest.

Measurements of Thermal Thresholds

The TSA-2 was strapped to the participant with the provided Velcro strap. The device used a starting temperature for WDT and CDT testing of 32˚C. For HPT and CPT, the skin temperature of the participant was measured with an infrared non-contact digital thermometer (FR-200 Thermometer, Metene, USA). The measured skin temperature was manually entered into the Medoc software as the starting/baseline temperature for the HPT and CPT tests. Between each threshold, the device was programmed to wait ten seconds at the starting temperature between tests.

The prototypes thermode was held on the participant forearm by the researcher. The device was programmed to measure the participant's skin temperature and begin all tests from this temperature (baseline). Within the cycle the prototype did not begin the next threshold test until the skin had returned to the baseline measured temperature along with a ten second delay.

Each participant's sex, age and handedness were recorded as well as the room temperature. The method of limits testing algorithm with a change of 1˚C/sec was used. Three consecutive tests for each threshold were measured starting with increasing temperature for three tests followed by decreasing temperature for 3 tests to complete one cycle. The test was stopped between WDT/CDT and WPT/CPT to record the results and the thermode was moved one thermode length proximally for HPT and CPT testing to use new skin. The mean (n=3) of each test (WDT, CDT, WPT and CPT) was used for statistical analysis.

Statistical Analysis

To determine the number of participants in this study the tables by Bujang [28] were followed which showed the sample size requirement is 9 for a power of 90%, alpha of 0.05, Intraclass Correlation Coefficient (ICC) of 0.8 and with two trials. Furthermore, recommendations in a handbook by Isaac and Michael [29] for feasibility studies is to test 10-30 participants.

The data was analyzed using the Statistical Package for the Social Sciences (SPSS) software (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.). All trials were shown to be normally distributed with the Kolmogorov-Smirnov test with 10 degrees of freedom, which accepted the null hypothesis that the results were normally distributed as $p > 0.05$ were found for each [30]. However, a comprehensive

protocol for clinical trials from 2006 recommends WDT and CDT data be logarithmically transformed so this was done with SPSS to maintain consistency [31].

Test-Retest Reliability

To begin, the differences between measurements from trial one and trial two of the prototype were analyzed using a paired sample t-test [32]. The mean differences between each trial were reported. The reliability was analyzed by calculating the intraclass correlation coefficient (ICC). The results were interpreted as per Koo and Li [33], where values less than 0.5 are indicative of poor reliability, between 0.5 and 0.75 imply moderate reliability, between 0.75 and 0.9 imply good reliability, and greater than 0.90 imply excellent reliability. ICC was calculated using a two-way mixed model with absolute agreement based on a mean rating (k=3) and a 95% confidence interval [33].

Agreement measures were calculated using Bland-Altman plots which analyze the agreement between the two different tests with the prototype. Agreement refers to the degree at which the results are identical between tests on the same subject [34]. Bland-Altman plots calculate limits of agreement (LoAs) using the mean and the standard deviation of the difference between two trials (mean difference ±1.96*SD) at a 95% confidence interval [35].

The Y axis shows the difference of each trial (Prototype Trial 1 - Prototype Trial 2) and the X axis shows the mean ((Prototype Trial 1 + Prototype Trial 2))/2).

In addition, the standard error of measurement (SEM) was calculated using the following formulas:

Sum of Squares (SS) – calculated from ANOVA table on SPSS -25

Standard Deviation (SD) = $\sqrt{\boldsymbol{SS_{Total}/(n-1)}}$

SEMs: SD$\sqrt{\boldsymbol{1-ICC}}$ [36]

The SEMs results were considered reliably if they were within 5% of the mean [37].

Concurrent Validity

The first analysis was a one-way ANOVA looking for any statistical differences between the three groups (prototype trial one, prototype trial two and the TSA-2 trial). A post-hoc analysis was performed using the equal variance Tukey test, the level of significance was set to p=0.05 and the effect size (η2) was reported [38]. The effect size was analyzed as per Cohen [39] as $\eta 2 < 0.01$ = small effect size, $0.01 > \eta^2 < 0.06$ = medium effect size and an $\eta^2 > 0.14$ = large effect size.

Bland-Altman plots were produced looking at the values from the prototype trials separately compared to the TSA-2 trial and the LoAs reported. The dependent variable was the difference between trials (Prototype trial - TSA-2) and the independent variable was the mean ((Prototype trial + TSA-2 trial)/2). The Bland Altman plots were reported with the confidence limits in hyperbolae form with a line of best fit [40].

Results

The participants included 5 male and 5 female with a mean (SD) age of 28 (4) years, and 80% were right-handed (n=8). The room temperature ranged from 19.29˚C to 25.09˚C. The mean (SD)

of all tests (n=3) from all trials (prototype trial 1, prototype trial 2 and TSA-2 trial) can be found in Table 1.

[Table 1 near here]

Reliability

There was no significant differences in results for WDT_{log} (p=0.600), CDT_{log} (p=0.939), HPT (p=0.742) and CPT (p=0.332) for the prototype trials 1 and 2 (Table 2).

[Table 2 near here]

The ICC $_{(2,1)}$ analysis showed that CPT has good reliability while HPT has moderate reliability ($p \leq 0.05$). WDT and CDT did not have statistically significant ($p \geq 0.154$) ICC values but WDT had moderate reliability and CDT poor reliability (Table 2).

The SEM results demonstrate that all trials were less than 2˚C with HPT having the smallest value (1.22˚C) while CPT has the largest (1.74˚C) (Table 2). However, contrary to the ICC value for CPT, the SEM value is greater than 5% of the mean, indicating it is not reliable. The SEM values for WDT, CDT and HPT show good reliability.

The Bland-Altman Plot's for the prototype trials 1 and 2 can be found in Figure 1 while the LoA results are reported in Table 2. All points, except one in CPT testing, fall within the LoA. The LoAs are smallest for WDT (-4.61 to 3.84) and largest for CPT (-7.64 to 5.47). This shows that the results from all four tests (WDT, CDT, HPT, CPT) are agreeable.

[Figure 1 near here]

Validity

There were significant differences between the prototype WDT Trials and the TSA-2 trial [F (2,27)=18.373, p=0.000, $\eta 2$ = 0.276] (Table 3). Post hoc comparisons using the Tukey test indicated that the mean score for the prototype trial one was significantly different than the TSA-2 trial (M=2.117, CI=0.372 – 3.861, p=0.015). However, the prototype trial two did not significantly differ from the TSA-2 trial (M=1.730, CI=-0.014 – 3.475, p=0.052) (Table 3).

[Table 3 near here]

There were significant differences between the prototype CDT Trials and the TSA-2 trial [F (2,27) =5.133, p=0.013, $\eta 2$ = 0.576]. Post hoc comparisons using the Tukey test indicated that the mean score for the prototype trial one was significantly different than the TSA-2 trial (M=3.706, CI=1.968 – 5.444, p=0.000). The prototype trial two was also significantly different from TSA-2 trial (M=3.653, CI= 1.915 – 5.391, p=0.000) (Table 3).

There were no significant differences between the prototype HPT Trials and the TSA-2 trial [F (2,27)=1.224, p=0.310, $\eta 2$ = 0.083] and there were no significant differences between the prototype CPT Trials and the TSA-2 trial [F (2,27)=0.059, p=.943, $\eta 2$ =0.004] (Table 3).

The LoAs for the Bland-Altman are reported in Table 4. Bland-Altman plots were created to look at the agreement between the two trials of the prototype compared to the TSA-2 trial. The LoAs of WDT, HPT and CPT for both trials contain the number zero. The Bland-Altman plots show that the results from all four tests (WDT, CDT, HPT, CPT) are agreeable (Figure 2).

[Figure 2 near here]

Discussion

The aim of this study was to test a novel thermal testing device and analyze its test-retest reliability and agreement, along with its concurrent validity compared to the reference standard (TSA-2). With regard to the main aim of the study, a novel gadget called Nerve Sensory Function Device (NSFD) was developed as an innovative equipment. The NSFD works by testing the sensory nerve endings and evaluates various sensory functions of the nerve such as hot-cold sensation and hot-cold pain thresholds. The NSFD applies a sensory stimulus to the sensory nerve endings in order to evaluate the sensory function of the nerve as a response. Thus after developing the NSFD, the reliability and validity of the device was also examined in the current study.

Reliability

The prototype showed good test-retest reliability for all thresholds as no significant differences between each of the two trials based on the means was found. The good reliability findings of the CPT (0.88) compare to the findings of Knutti, Suter [41] on the L5 dermatome (ICC 0.68 – 0.90) and Wasner and Brock [25] who found an ICC of 0.781 comparing testing on day 1 vs day 21 on the dorsum of the hand using the same size of thermode. HPT had moderate reliability (0.71) which compares to Felix and Widerstrom-Noga [21] (0.55-0.79) who also tested 10 healthy subjects, however, with a smaller 16mm x 16mm probe in eight different locations on the body. It has been shown that a larger thermode ($9cm^2$) results in decreased thresholds compared to a much smaller thermode ($2.5cm^2$), possibly explained by spatial summation [42]. Other studies reported higher ICC values for HPT (good to excellent reliability) however they included larger sample sizes [25,26,41]. The prototype requires further testing with a larger sample.

Although our study found poor reliability for CDT (ICC 0.29), findings are in line with previous research. With a one week interval, testing on L4, L5 and S1, Krassioukov, Wolfe [23] found poor reliability with CDT, however the thermode size was not reported. Zwart and Sand [43] found poor reliability for both WDT and CDT when testing L4, L5 and S1 with a larger 25x50mm thermode after 1-2 hours. A moderately reliable WDT result, as found in our study (ICC 0.54), is common. Felix and Widerstrom-Noga (2009), Krassioukov et al. (1999) and. Nothnagel, Puta [26] also found a moderate WDT (ICC 0.70, 0.36–0.84 and 0.51 respectively).

The agreement is important to analyze as it can be counterintuitive to the ICC findings [30]. Nothnagel, Puta [26] found CPT LoAs as the largest out of all thermal tests (-18.22˚C to 15.20˚C) on the hand which is a very large temperature range (33˚C) while this study found a small range of only 13˚C indicating improved agreement. Bland Altman plots do not provide an analysis if the LoA's found are suitable but the smaller the bias the better [30].

Historically thermal threshold testing reliability varies significantly on healthy participants, especially for thermal detection thresholds [21,26,44]. There are many factors which could contribute to this including the small sample size of this study and other studies, the differences in testing location on the body, the different baseline temperature (measured skin temperature instead of a set temperature), non-standardized testing, reaction time of the participant which needs to be quicker for thermal detection thresholds compared to thermal pain thresholds, outside distractions or unknown room and skin temperatures [42,45-47]. The prototype design is working towards improving the reliability with improved technology including the skin temperature and room temperature sensors which could have an advantage over coins, ice and test tube testing methods as the temperature of the thermode is always known [13-15].

Validity

The results show that the prototype is valid for testing HPT and CPT. The results for the prototype compared to the TSA-2 were statistically different for WDT and CDT with large effect sizes. One way to look at the results is by comparing how far CDT and WDT are from the baseline temperature (e.g. WDT result minus measured skin temperature for the prototype or WDT result minus 32˚C for the TSA-2). The WDT results for the prototype are much farther from baseline than previous studies with participants in prototype trial one pressing the stop button +3.73˚C from skin/starting temperature, trial two +4.6˚C from skin temperature and the TSA-2 was +2.39˚C from the 32˚C starting temperature. From previous studies, on similar testing locations +1.64˚C and +1.67˚C were found for WDT from baseline [16,48]. Higher WDT with warmer baseline temperature (35˚C) has been previously reported as found in this study [12] and could be a result of starting at skin temperature.

CDT was similar to previous findings with regards to the change in temperature from the baseline temperature. The prototype trial one averaged -2.38˚C from skin temperature, trial two -1.84˚C from skin temperature and the TSA-2 was -2.12˚C from the 32˚C starting temperature. From previous studies, on similar locations -1.12˚C and -1.77˚C were found for CDT from baseline [16,48].

One of the big differences between the TSA-2 and the prototype is the baseline temperatures. The baseline temperature at which the thermode begins each cycle can influence results [45,46,49]. For HPT and CPT, both devices (prototype and TSA-2) started at the measured skin temperature, which could have contributed to the validity results. A previous study found that there was no significant difference for CPT results when the baseline temperature was 32˚C or 36˚C [50]. For the WDT and CDT tests, the TSA-2 was started at a common starting temperature of 32˚C [12],

while the mean temperature reported by the prototype was 33.96˚C. [45,46,51]. Colder thermode temperatures can lead to increased detection of decreasing temperature while a warmer starting thermode can result in improved detection of increasing temperatures [49]. Participants pressed the stop button on the TSA-2 as it warmed up to 32˚C prior to starting WDT testing, indicating that the probe felt warm. This design is similar to the alternative testing methods such as coins or test tubes, which do not change the skin temperature prior to beginning the test [13-15].

The temperature of the room, if too cold (<10˚C) or too hot (>25˚C) can influence skin temperature and change the WDT and CDT [52]. The prototype was designed to instantly display the room temperature to allow the clinician to address the issue if too warm or too cool before testing. The prototype also records the skin temperature, which may be influenced by room temperature. Therefore, the test should not be completed until the skin is 25˚C-37˚C [45,46]. Both features are advantages of the prototype, ensuring more accurate thermal testing. Participants were not acclimatized to the room temperature prior to testing, resulting in skin temperatures measured ranging from 29.1˚C to 37.71˚C, which may have influenced results [42,46].

Bland Altman plots can also be used to compare a new measurement technique (prototype) with a gold standard (TSA-2), as even a gold standard's results could have error [30]. The Bland-Altman plots show agreement between all the tests indicating both the TSA-2 and the prototype had acceptable results.

Participants were not blinded to their results, especially with the TSA-2 as it displays a visual graph on a laptop immediately after completion of the testing. Participants could see their results if they wished. This could have influenced further tests, influencing the reliability of the results [53]. The prototype, however, did not have a visual display of the results for the participants (only for the researcher) ensuring accurate reliable testing. As this was a feasibility study, the number of

participants was small. To improve the significance of the results, the number of participants will be increased in the future. The results can only be applied to the small number of participants who came from a distinct group (all physiotherapy students at one university). Furthermore, all participants were between 26 and 39 years old, which represents a small age group. Further research of the prototype is needed, including design alterations and further testing of reliable and valid for different pathologies, body locations and participant ages.

Conclusion

This study is a steppingstone for the creation of a clinically useful thermal threshold device as the test-retest reliability for HPT and CPT ranged from moderate to good respectively, and all thermal thresholds were aggregable. Furthermore, both trials of HPT and CPT and one trial for WDT were found to be valid compared to the TSA-2. However, the device can be improved even further to ensure its reliability, validity and ease of use in practice. Therefore, the prototype will be redesigned and further research towards its feasibility, reliability and validity is warranted.

Acknowledgement:

The authors wish to thank the Department of Research, Enterprise and Social Partnerships, University of Brighton, United Kingdom for all the sincere support and assistance provided to the research study. The current study is supported by Innovation Kick Start Funding awarded by the University of Brighton.

Word Count – 3413 not including abstract

Declaration of Interest Statement

The authors report no conflict of interest

References

1. Heldestad Lillieskold V, Nordh E. Method-of-limits; Cold and warm perception thresholds at proximal and distal body regions. Clin Neurophysiol Pract. 2018;3:134-140.
2. Jimenez-Cohl P, Grekin C, Leyton C, et al. Thermal threshold: research study on small fiber dysfunction in distal diabetic polyneuropathy. J Diabetes Sci Technol. 2012 Jan 1;6(1):177-83.
3. Bachmann CG, Rolke R, Scheidt U, et al. Thermal hypoaesthesia differentiates secondary restless legs syndrome associated with small fibre neuropathy from primary restless legs syndrome. Brain. 2010 Mar;133(Pt 3):762-70.
4. Bakkers M, Faber CG, Reulen JP, et al. Optimizing temperature threshold testing in small-fiber neuropathy. Muscle Nerve. 2015 Jun;51(6):870-6.
5. Farooqi MA, Lovblom LE, Lysy Z, et al. Validation of cooling detection threshold as a marker of sensorimotor polyneuropathy in type 2 diabetes. J Diabetes Complications. 2016 May-Jun;30(4):716-22.
6. Maixner W, Fillingim R, Booker D, et al. Sensitivity of patients with painful temporomandibular disorders to experimentally evoked pain. Pain. 1995 Dec;63(3):341-51.
7. Malmstrom EM, Stjerna J, Hogestatt ED, et al. Quantitative sensory testing of temperature thresholds: Possible biomarkers for persistent pain? J Rehabil Med. 2016 Jan;48(1):43-7.
8. Loseth S, Stalberg E, Jorde R, et al. Early diabetic neuropathy: thermal thresholds and intraepidermal nerve fibre density in patients with normal nerve conduction studies. J Neurol. 2008 Aug;255(8):1197-202.
9. Cahill LS, Lannin NA, Mak-Yuen YYK, et al. Changing practice in the assessment and treatment of somatosensory loss in stroke survivors: protocol for a knowledge translation study. BMC Health Serv Res. 2018 Jan 23;18(1):34.
10. Cruccu G, Truini A. Neuropathic pain and its assessment. Surg Oncol. 2010 Sep;19(3):149-54.
11. Sterling M. Testing for sensory hypersensitivity or central hyperexcitability associated with cervical spine pain. J Manipulative Physiol Ther. 2008 Sep;31(7):534-9.

12. Bakkers M, Faber CG, Peters MJ, et al. Temperature threshold testing: a systematic review. J Peripher Nerv Syst. 2013 Mar;18(1):7-18.
13. Ridehalgh C, Sandy-Hindmarch OP, Schmid AB. Validity of Clinical Small-Fiber Sensory Testing to Detect Small-Nerve Fiber Degeneration. J Orthop Sports Phys Ther. 2018 Oct;48(10):767-774.
14. Tilley P, Bisset L. The Reliability and Validity of Using Ice to Measure Cold Pain Threshold. Biomed Res Int. 2017;2017:7640649.
15. Zhu GC, Bottger K, Slater H, et al. Concurrent validity of a low-cost and time-efficient clinical sensory test battery to evaluate somatosensory dysfunction. Eur J Pain. 2019 Jul 20.
16. Rolke R, Baron R, Maier C, et al. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain. 2006 Aug;123(3):231-43.
17. Agostinho CM, Scherens A, Richter H, et al. Habituation and short-term repeatability of thermal testing in healthy human subjects and patients with chronic non-neuropathic pain. Eur J Pain. 2009 Sep;13(8):779-85.
18. Connell LA, Tyson SF. Measures of sensation in neurological conditions: a systematic review. Clin Rehabil. 2012 Jan;26(1):68-80.
19. Winward CE, Halligan, P. W., & Wade, D. T. . Somatosensory Assessment after Central Nerve Damage: the Need for Standardized Clinical Measures. Physical Therapy Reviews. 1999;4:21-28.
20. Stolk-Hornsveld F, Crow JL, Hendriks EP, et al. The Erasmus MC modifications to the (revised) Nottingham Sensory Assessment: a reliable somatosensory assessment measure for patients with intracranial disorders. Clin Rehabil. 2006 Feb;20(2):160-72.
21. Felix ER, Widerstrom-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):69-83.
22. Kemler MA, Reulen JP, van Kleef M, et al. Thermal thresholds in complex regional pain syndrome type I: sensitivity and repeatability of the methods of limits and levels. Clin Neurophysiol. 2000 Sep;111(9):1561-8.

23. Krassioukov A, Wolfe DL, Hsieh JT, et al. Quantitative sensory testing in patients with incomplete spinal cord injury. Arch Phys Med Rehabil. 1999 Oct;80(10):1258-63.
24. Moravcová E, Bednarik, J., Svobodník, A. and Dušek, L. Reproducibility of thermal threshold assessment in small-fibre neuropathy patients. Scr Med(Brno). 2005;78(3):177-184.
25. Wasner GL, Brock JA. Determinants of thermal pain thresholds in normal subjects. Clin Neurophysiol. 2008 Oct;119(10):2389-95.
26. Nothnagel H, Puta C, Lehmann T, et al. How stable are quantitative sensory testing measurements over time? Report on 10-week reliability and agreement of results in healthy volunteers. J Pain Res. 2017;10:2067-2078.
27. Pavlakovic G, Klinke I, Pavlakovic H, et al. Effect of thermode application pressure on thermal threshold detection. Muscle Nerve. 2008 Nov;38(5):1498-1505.
28. Bujang MAB, N. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. Archives of Orofacial Science. 2017;12.
29. Isaac S, Michael WB. Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences, 3rd ed. San Diego, CA, US: EdITS Publishers; 1995. (Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences, 3rd ed.).
30. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141-51.
31. Rolke R, Magerl W, Campbell KA, et al. Quantitative sensory testing: a comprehensive protocol for clinical trials. Eur J Pain. 2006 Jan;10(1):77-88.
32. Werner MU, Petersen MA, Bischoff JM. Test-retest studies in quantitative sensory testing: a critical review. Acta Anaesthesiol Scand. 2013 Sep;57(8):957-63.
33. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155-63.
34. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud. 2011 Jun;48(6):661-71.

35. Pavlakovic G, Zuchner K, Zapf A, et al. Influence of intrinsic noise generated by a thermotesting device on thermal sensory detection and thermal pain detection thresholds. Muscle Nerve. 2009 Aug;40(2):257-63.
36. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res. 2005 Feb;19(1):231-40.
37. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med. 1998 Oct;26(4):217-38.
38. Stoline MR. The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. The American Statistician. 1981;35(3):134-141.
39. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988.
40. Ludbrook J. Confidence in Altman-Bland plots: a critical review of the method of differences. Clin Exp Pharmacol Physiol. 2010 Feb;37(2):143-9.
41. Knutti IA, Suter MR, Opsommer E. 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 Sep 5;579:157-62.
42. Defrin R, Petrini L, Arendt-Nielsen L. Spatial summation of thermal sensations depends on skin type and skin sensitivity. Exp Brain Res. 2009 Sep;198(1):29-36.
43. Zwart JA, Sand T. Repeatability of dermatomal warm and cold sensory thresholds in patients with sciatica. Eur Spine J. 2002 Oct;11(5):441-6.
44. Moloney NA, Hall TM, Doody CM. Reliability of thermal quantitative sensory testing: a systematic review. J Rehabil Res Dev. 2012;49(2):191-207.
45. Hagander LG, Midani HA, Kuskowski MA, et al. Quantitative sensory testing: effect of site and skin temperature on thermal thresholds. Clin Neurophysiol. 2000 Jan;111(1):17-22.
46. Pertovaara A, Kauppila T, Hamalainen MM. Influence of skin temperature on heat pain threshold in humans. Exp Brain Res. 1996;107(3):497-503.
47. Harrison JL, Davis KD. Cold-evoked pain varies with skin type and cooling rate: a psychophysical study in humans. Pain. 1999 Nov;83(2):123-35.

48. Yarnitsky D, Sprecher E. Thermal testing: normative data and repeatability for various test algorithms. J Neurol Sci. 1994 Aug;125(1):39-45.
49. Hilz MJ, Glorius S, Beric A. Thermal perception thresholds: influence of determination paradigm and reference temperature. J Neurol Sci. 1995 Apr;129(2):135-40.
50. Kim HK, Kim KS, Kim ME. Influence of test site and baseline temperature on orofacial thermal thresholds. J Orofac Pain. 2013 Summer;27(3):263-70.
51. Leffler AS, Hansson P. Painful traumatic peripheral partial nerve injury-sensory dysfunction profiles comparing outcomes of bedside examination and quantitative sensory testing. Eur J Pain. 2008 May;12(4):397-402.
52. Hirosawa I, Dodo H, Hosokawa M, et al. Physiological variations of warm and cool sense with shift of environmental temperature. Int J Neurosci. 1984 Nov;24(3-4):281-8.
53. Lucas NP, Macaskill P, Irwig L, et al. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol. 2010 Aug;63(8):854-61.

Funding

This project was partially funded by the University of Brighton Innovation Kick Start Award. The funds received went directly towards the purchasing of components for the prototype.

Table 1 – Results mean (SD) in ˚C of all Trials

Test	Prototype T1	Prototype T2	TSA-2 Trial
WDT	32.27 (1.68)	32.65 (2.05)	34.39 (0.63)
CD	26.18 (1.86)	26.23 (1.77)	29.88 (0.87)
HPT	40.71 (2.19)	40.95 (2.44)	42.14 (1.85)
CPT	16.57 (5.07)	17.66 (5.26)	16.79 (10.72)

Notes: T1, prototype testing trial one; T2, prototype testing trial two; TSA, Medoc TSA-2, Advanced Thermosensory Stimulator, Israel; WDT, warm detection threshold; CDT, cold detection threshold; HPT, heat pain threshold; CPT, Cold Pain Threshold;

Table 2 - Test-Retest Reliability Results

Parameter	Difference (T1 – T2)			ICC			SEM (%) (˚C)	LoA (°C)
	Mean Difference ± SD (°C)	95% CI of mean (°C)	*p*	ICC	*p*	Lower LoA – Upper LoA		Lower LoA – Upper LoA
WDT_{log}	0.00±0.03	-0.03-0.02	0.600	0.54	0.140	-1.09–0.88	0.02 (4%)	-0.06 - +0.05
CDT_{log}	0.00±0.04	-0.03-0.03	0.939	0.29	0.322	-3.11–0.82	0.03 (5%)	-0.08 - +0.08
HPT	-0.24±2.25	-1.85-1.37	0.742	0.71	0.047*	-0.26–0.93	1.22 (3%)	-4.67 - +4.18
CPT	-1.08±3.35	-3.48-1.31	0.332	0.88	0.002*	0.56–0.97	1.74 (10%)	-7.64 - +5.47

Notes: Level of significance: *p≤0.05; T1, Trial one with prototype; T2, Trial two with prototype; WDT, warm detection threshold; CDT, cold detection threshold; HPT, heat pain threshold; CPT, Cold Pain Threshold; SD, standard deviation; CI, confidence interval; ICC, intraclass correlation coefficient; LoA, limits of agreement according to Bland and Altman [28]; SEM, standard error of measurement

Table 3 – Oneway ANOVA and Post Hoc Analysis Results for Prototype Trials Compared to TSA-2 Trial

Test (TSA-2 trials)	Group	Post Hoc Tukey Test TSA-2			ANOVA	Effect Size
		Mean Difference, °C	95% CI, °C	*p*	*p*	*η2*
WDT	**T1**	2.12	0.37 – 3.86	0.015*		
	T2	1.73	-0.01 – 3.47	0.052		
	Between Groups				0.013*	0.276
CDT	**T1**	3.71	1.97 – 5.44	0.000**		
	T2	3.65	1.91 – 5.39	0.000**		
	Between Groups				0.000**	0.576
HPT	**T1**	1.42	-0.99 – 3.84	0.324		
	T2	1.18	-1.23 – 3.60	0.456		
	Between Groups				0.310	0.083
CPT	**T1**	0.22	-8.01 – 8.52	0.998		
	T2	-0.87	-9.17 – 7.44	0.964		
	Between Groups				0.943	0.004

Notes: level of significance: *p≤0.05, **p≤0.001; T1, Trial one with prototype; T2, Trial two with prototype; WDT, warm detection threshold; CDT, cold detection threshold; HPT, heat pain threshold; CPT, Cold Pain Threshold; CI, confidence interval

Table 4 – Limits of Agreement for Prototype Trials with TSA-2

Trial	Parameter	LoA
		Lower LoA –Upper LoA (°C)
Prototype T1 compared to TSA-2 Trial	**WDT_{log}**	-0.08 – +0.02
	CDT_{log}	-0.12 – 0.00
	HPT	-6.25 – +3.40
	CPT	-20.33 - +19.90
Prototype T2 compared to TSA-2 Trial	**WDT_{log}**	-0.07 – +0.03
	CDT_{log}	-0.11 - 0.00
	HPT	-7.58 – +5.21
	CPT	-19.17 - +20.91

Notes: T1, Trial one with prototype; T2, Trial two with prototype; WDT, warm detection threshold; CDT, cold detection threshold; HPT, heat pain threshold; CPT, Cold Pain Threshold; LoA, limits of agreement according to Bland and Altman [28];

Figure 1 - Bland-Altman Plot for prototype trials.
Figure 2 - Bland Altman plot of differences against averages for prototype trials and TSA-2 trial with line of best fit.

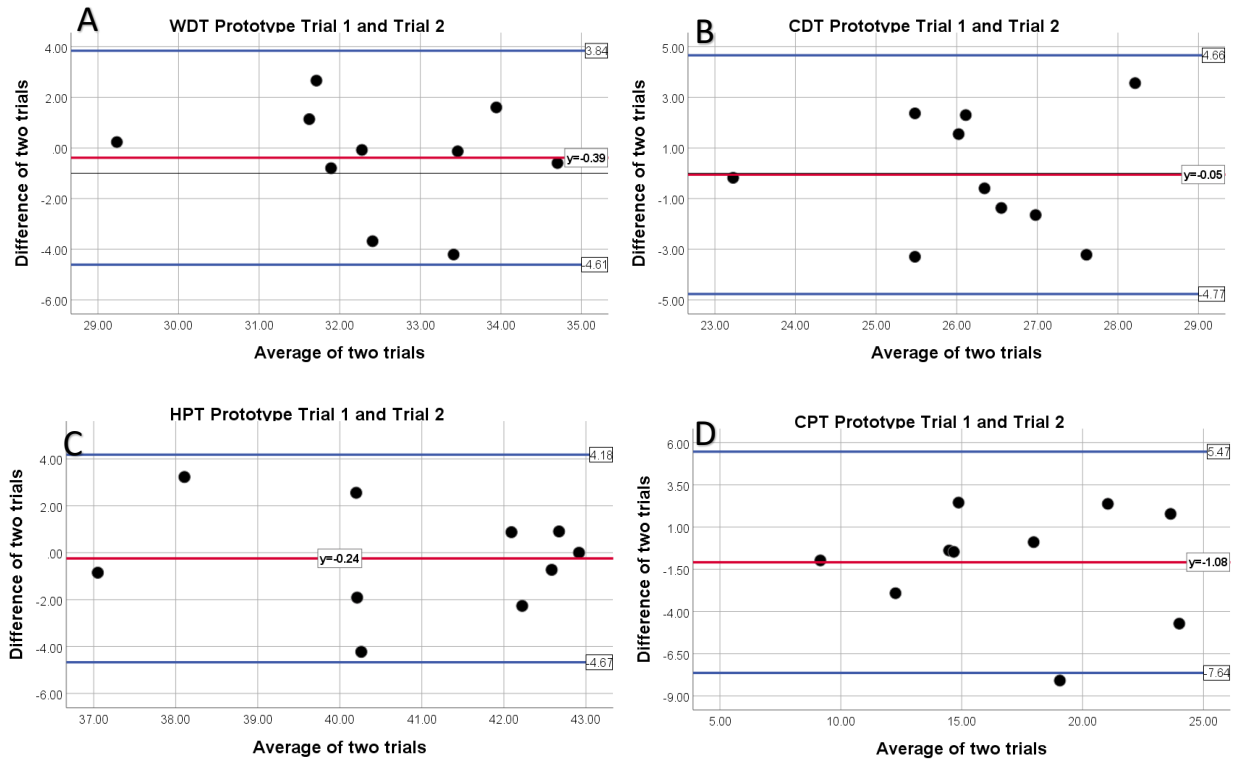

Figure 1 - Bland-Altman Plot for prototype trials.

Notes: (A)WDT Prototype Trail 1 compared to Prototype Trial 2; (B) CDT Prototype Trail 1 compared to Prototype Trial 2; (C) HPT Prototype Trail 1 compared to Prototype Trial 2; (D) CPT Prototype Trail 1 compared to Prototype Trial 2; ; The middle horizontal red line represents the mean difference between prototype trial 1 and prototype trial 2; upper and lower blue lines indicate upper and lower limits of agreement, mean difference ±1.96× SD. CDT, cold detection threshold; WDT, warm detection threshold; HPT, heat pain threshold; CPT, cold pain threshold;

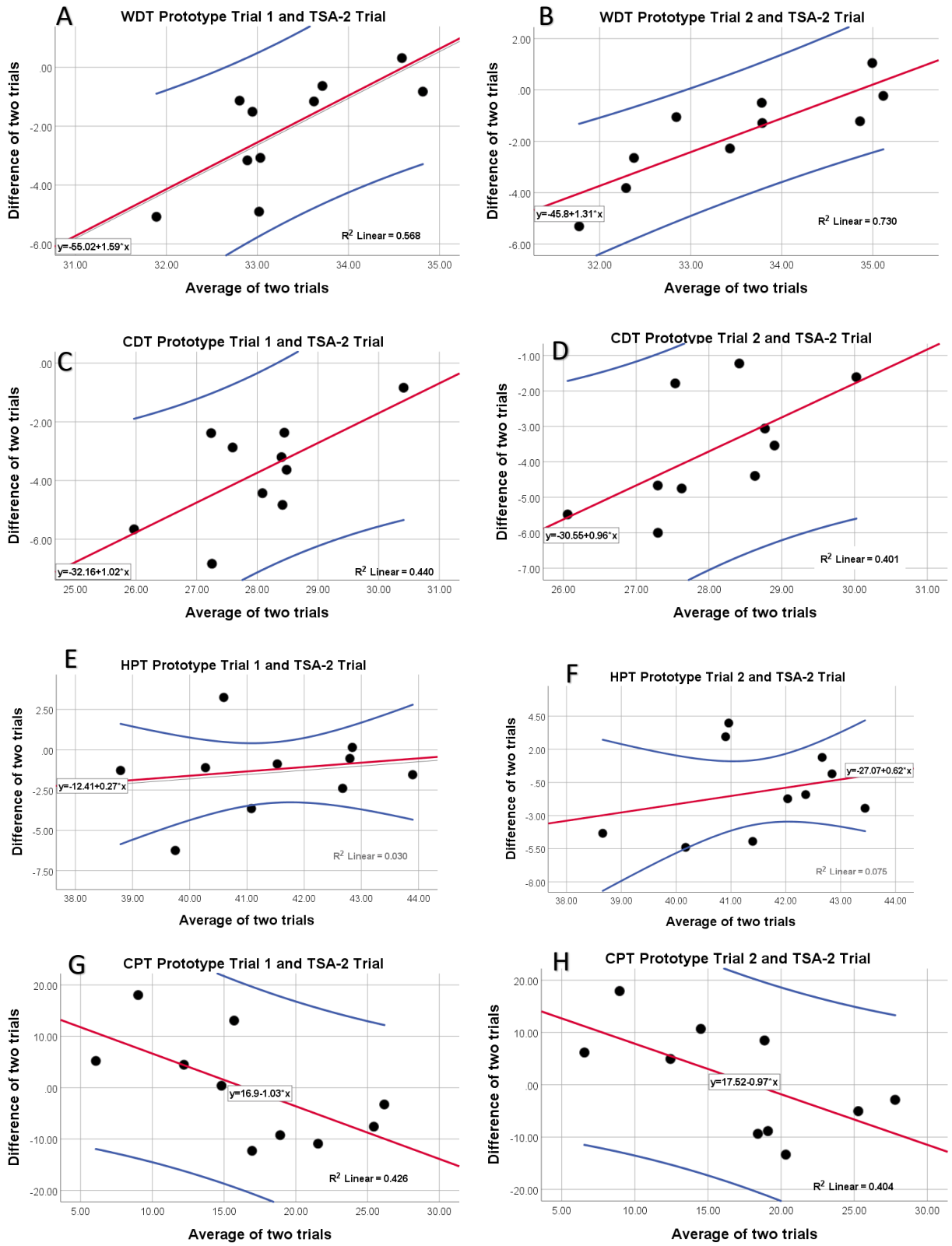

Figure 2 - Bland Altman plot of differences against averages for prototype trials and TSA-2 trial with line of best fit.

Notes: (A)WDT Prototype Trial 1; (B) WDT Protype Trial 2; (C) CDT Prototype Trial 1; (D) CDT Prototype Trial 2; (E) HPT Prototype Trial 1; (F) HPT Prototype Trial 2; (G) CPT Prototype Trial 1; (H) CPT Prototype Trial 2; The upper and lower confidence (prediction) limits for an individual at 95% confidence intervals in blue with ordinary least squares line of best fit in red. CDT, cold detection threshold; WDT, warm detection threshold; HPT, heat pain threshold; CPT, cold pain threshold;