Aalborg Universitet

Reliability of the conditioned pain modulation paradigm across three anatomical sites

Nuwailati, Rania; Curatolo, Michele; LeResche, Linda; Ramsay, Douglas S; Spiekerman, Charles; Drangsholt, Mark

Clinical pain research

Rania Nuwailati, Michele Curatolo, Linda LeResche, Douglas S. Ramsay, Charles Spiekerman[a] and Mark Drangsholt*

Reliability of the conditioned pain modulation paradigm across three anatomical sites

Abstract

Background and aims: Conditioned Pain Modulation (CPM) is a measure of pain inhibition-facilitation in humans that may elucidate pain mechanisms and potentially serve as a diagnostic test. In laboratory settings, the difference between two pain measures [painful test stimulus (TS) without and with the conditioning stimulus (CS) application] reflects the CPM magnitude. Before the CPM test can be used as a diagnostic tool, its reliability on the same day (intra-session) and across multiple days (inter-session) needs to be known. Furthermore, it is important to determine the most reliable anatomical sites for both the TS and the CS. This study aimed to measure the intra-session and inter-session reliability of the CPM test paradigm in healthy subjects with the TS (pressure pain threshold-PPT) applied to three test sites: the face, hand, and dorsum of the foot, and the CS (cold pressor test-CPT) applied to the contralateral hand.

Methods: Sixty healthy participants aged 18–65 were tested by the same examiner on 3 separate days, with an interval of 2–7 days. On each day, testing was comprised of two identical experimental sessions in which the PPT test was performed on each of the three dominant anatomical sites in randomized order followed by the CPM test (repeating the PPT with CPT on the non-dominant hand). CPM magnitude was calculated as the percent change in PPT. The Intraclass Correlation Coefficient (ICC), Coefficient of Variation (CV), and Bland-Altman analyses were used to assess reliability.

Results: PPT relative reliability ranged from good to excellent at all three sites; the hand showed an intra-session ICC of 0.90 (0.84, 0.94) before CPT and ICC of 0.89 (0.83, 0.92) during CPT. The PPT absolute reliability was also high, showing a low bias and small variability when performed on all three sites; for example, CV of the hand intra-session was 8.0 before CPT and 8.1 during CPT. The relative reliability of the CPM test, although only fair, was most reliable when performed during the intra-session visits on the hand; ICC of 0.57 (0.37, 0.71) vs. 0.20 (0.03, 0.39) for the face, and 0.22 (0.01, 0.46) for the foot. The inter-session reliability was lower in all three anatomical sites, with the best reliability on the hand with an ICC of 0.40 (0.23, 0.55). The pattern of absolute reliability of CPM was similar to the relative reliability findings, with the reliability best on the hand, showing lower intra-session and inter-session variability (CV% = 43.5 and 51.5, vs. 70.1 and 73.1 for the face, and 75.9 and 78.9 for the foot). The CPM test was more reliable in women than in men, and in older vs. younger participants.

Discussion: The CPM test was most reliable when the TS was applied to the dominant hand and CS performed on the contralateral hand. These data indicate that using the CS and TS in the same but contralateral dermatome in CPM testing may create the most reliable results.

Keywords: pain assessment; pain measurement; conditioned pain modulation; experimental pain; reliability; test-retest.

[a]Deceased August 9, 2019.

*Corresponding author: Mark Drangsholt, Department of Oral Health Sciences, University of Washington, Seattle, WA, USA; and Department of Oral Medicine, University of Washington, Seattle, WA, USA, E-mail: drangs@uw.edu

Rania Nuwailati: Department of Oral Health Sciences, University of Washington, Seattle, WA, USA

Michele Curatolo: Department of Anesthesiology and Pain Medicine, School of Medicine, University of Washington, Seattle, WA, USA; and Center for Sensory-Motor Interaction, Department of Health Science and Technology, Aalborg University, Aalborg, Denmark

Linda LeResche: Department of Oral Health Sciences, University of Washington, Seattle, WA, USA; and Department of Oral Medicine, University of Washington, Seattle, WA, USA

Douglas S. Ramsay: Department of Oral Health Sciences, University of Washington, Seattle, WA, USA; Department of Orthodontics, University of Washington, Seattle, WA, USA; and Department of Pediatric Dentistry, University of Washington, Seattle, WA, USA

Charles Spiekerman: Department of Oral Health Sciences, University of Washington, Seattle, WA, USA; and Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA

1 Introduction

Conditioned pain modulation (CPM) is a psychophysical phenomenon that describes endogenous pain modulation pathways in humans [1, 2]. It is a surrogate measure for the net effect of pain inhibitory and facilitatory mechanisms in the descending pain control system. In laboratory settings, the difference between two pain measures [painful test stimulus (TS) without and with the conditioning stimulus (CS) application] reflects the CPM magnitude [2]. The reduction of TS pain with the CS application is a sign of efficient CPM in healthy individuals [3]. In chronic pain populations, pain inhibition shown via CPM has been found to be deficient [3].

A clinical CPM test could be useful because patients with impaired endogenous pain systems may benefit from pharmacological agents or other methods that augment CPM [4, 5]. In addition, CPM efficiency may be a prognostic factor in chronic pain [4, 6–10]. Thus, a clinical CPM test may be possibly utilized as a clinical diagnostic test of pain inhibition efficiency in developing a more personalized pain medicine approach for chronic pain populations [4]; for example, temporomandibular disorders (TMD), irritable bowel syndrome (IBS), or fibromyalgia [11]. It is critical to optimize the test's reliability before it is used in the clinic [12].

Despite multiple investigations conducted on the reliability of CPM, results have varied greatly, ranging from excellent [13], to poor reliability [14]. In an attempt to improve the CPM test reliability, studies have investigated different modalities such as different TS and CS [8, 9]. Intervals between the tests have ranged from 15-min to 10 months [14, 15], and sample sizes from 12 and 230 subjects, including healthy subjects and chronic pain patients [16, 17]. The most common TS used was the PPT test [13, 14, 16, 18–24], and the most common CS used was the CPT test [13, 14, 16–18, 22, 23, 25–28]. Most studies applied the TS on the upper and lower extremities, with the CS on the hand [13–20, 22, 25–30].

Numerous factors affect CPM effect and also CPM test reliability, including the anatomical site of the TS and CS. In particular, extrasegmental nociceptive stimulation effect on pain modulation depends on the site of application [31–33]. Previous CPM reliability protocols examined multiple anatomical sites [14, 16, 21–24, 30], but showed limitations in identifying which sites show the highest reliability. The limitations included: small sample sizes, tested sites were in the same dermatome, findings were not reported by site, some used unstandardized protocols, and others tested only subgroups. Thus, there is a strong need to further investigate which anatomical test and conditioning sites are most reliable in CPM. Studies have shown that CPM may be influenced by demographic factors such as age [34, 35] and gender [36–40]. It has been suggested that the CPM test is more reliable in women [17, 18], while the effect of age on CPM reliability remains unclear.

The overall aim of this study was to determine if intra-session and *inter*-session reliability of CPM varied according to anatomical test sites (face, hand, and foot) in healthy subjects. Furthermore, the intra-session and *inter*-session reliability of the CPM test paradigm were stratified by age and gender.

2 Methods

All tests were conducted by the same examiner (R.N.) who is a dentist-scientist trained in pain assessment. The investigation was conducted at the Department of Oral Medicine, University of Washington (UW) over 18 months, between November 2014 and April 2016. It was approved by the UW Institutional Review Board, and informed consent was obtained from each participant. This study is classified as a research reliability study, which does not fit any particular reporting checklist, to our knowledge. Therefore, three checklists were used to standardize the reporting of the study: the Quality Appraisal for Reliability Studies (QAREL) checklist, the Strengthening of the Reporting of Observational Studies in Epidemiology (STROBE) statement, and the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) checklist. The QAREL 11-item checklist is an appraisal tool that evaluates the quality of studies of diagnostic reliability [41]. Our study met seven of the items, did not meet three items related to blinding, and one item was not applicable. The STROBE statement is a 22-item checklist of information that should be included in a cohort study [42]. Our study included 15 of the items, and the other seven were not applicable. The GRRAS 15-item checklist is a set of guidelines that are broadly useful and applicable to the vast majority of diagnostic studies. This study met all 15-items of the checklist [43]. The study outcomes were the overall, intra-session and *inter*-session reliability – both relative and absolute – of the CPM test among the three tested sites.

2.1 Design

2.1.1 Participants

Sixty-one healthy women and men (age-range 18–65 years) who gave informed consent were recruited for this study.

Exclusion criteria were: history of chronic pain, intake of serotonin and norepinephrine reuptake inhibitors (SNRIs), current spontaneous pain or injuries in the lower, upper limbs or face, an inability to stop analgesic use for 24-h before their visit, pregnancy, current or previous major medical conditions such as severe heart disease or respiratory diseases, or a psychiatric condition. Potential participants, who expressed interest or contacted a study team member in response to the flyer or study handout, were recruited using a standard script over the phone or in person. They were instructed to stop any analgesic medication and not to smoke cigarettes for 24-h prior to their scheduled visits. Participants were paid $150 as compensation for completing all three visits.

2.1.2 Set-up

Participants were evaluated during three visits with an interval of 2–7 days between each visit. All visits were conducted in the same clinic, by the same examiner, at the same time of the day for each participant. Participants were seated in a quiet clinic room. The examiner explained the experiment to them at the beginning of the first visit, and before each session, using a standardized set of instructions. Each visit consisted of two identical experimental sessions, with a 15-min break in between. Both PPT and CPM tests were performed in every session (Fig. 1). The recommended "pain inhibits pain" paradigm to assess CPM was adopted in this study [2, 44, 45], using

the pressure pain threshold (PPT) as a TS and the cold pressor test (CPT) as CS [46]. A training session for both tests was performed on the hand at the beginning of the first visit, before starting the experiment, until the participants were familiar with the testing procedures. The PPT test was performed at three dominant-side (determined by handedness) sites: masseter muscle below zygomatic arch (referred to as the side of the face), the middle of the thenar eminence (referred to as the hand), and the middle of extensor digitorum brevis muscle at the dorsum of the foot (referred to as the foot), while the non-dominant hand was used for the CPM test (Fig. 2). The PPT test was performed three times at a given site, followed by the CPM test (PPT + CPT); then tests were performed on the next anatomical site, thereby allowing additional time between tests. There was a minimum of a 5-min up to 8-min break between sites until subjects reported the cold sensation to have completely faded away. One session served as an intra-subject control, where lukewarm water (26.6 °C) was used as a CS instead of CPT. This session was randomly assigned using a list of test session sequences that counterbalanced for session's order by test type (control vs. experimental) and test site.

2.1.3 Tests

2.1.3.1 Pressure pain threshold test – PPT (TS)
PPT was induced using the handheld Somedic Pressure Algometer with a circular 2 cm silicon rubber tip (probe), which can create a force from 0 to 1,000 kPa. The probe
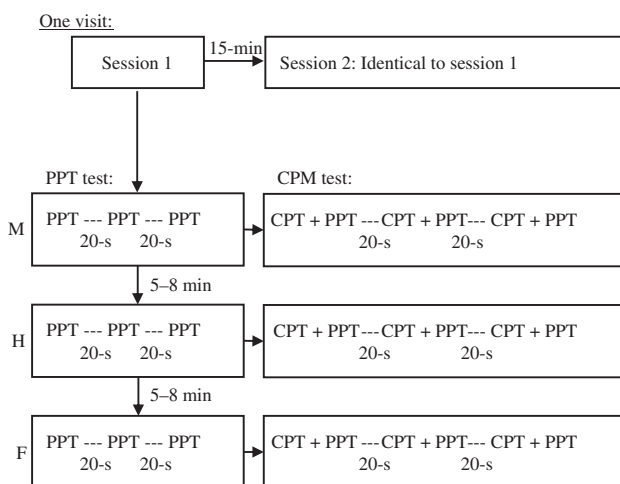

Fig. 1: Study flow chart: H = hand; F = foot; M = masseter (side of the face); PPT = pressure pain threshold; CPT = cold pressor test; CPM = conditioned pain modulation. The order of testing of the three sites assigned for each visit was counterbalanced in a random fashion for each participant.
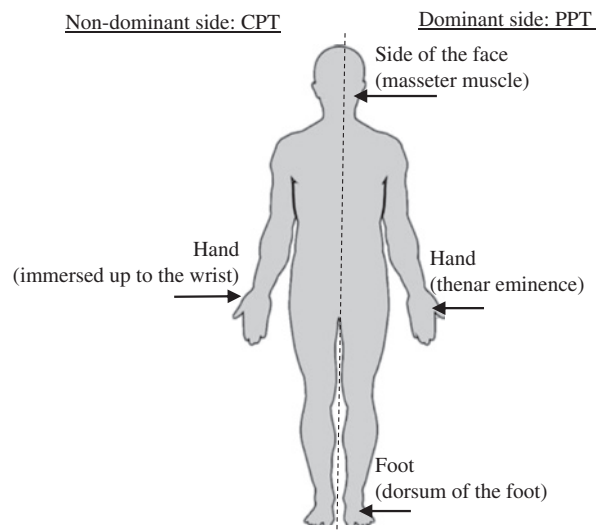

Fig. 2: Tested sites: the anatomical test sites and contralateral conditioning site.

was applied with a constant application rate of 30 kPa/s. The PPT test was performed with three applications of the algometer on a given site. A small template was used to avoid the spatial overlap of the three assessments. The pressure was increased until the participant pressed a trigger button indicating that the pressure was perceived as painful, at which point pressure was released, and the readout was recorded (PPT defined in units of pressure-kPa). Thresholds were computed as the average of the three measures taken 20-s apart. The algometer was calibrated before each experimental session.

2.1.3.2 Conditioned pain modulation test – CPM (PPT+CPT)

Participants were directed to immerse their non-dominant hand in a 5 °C cold water bath up to the wrist. In line with previous protocols [28, 45], they were asked to report their non-dominant hand pain level when it reached 7 out of 10 according to a numerical pain scale (NPS) – where 0 is no pain and 10 is severe pain. Once they reached 7/10 level of pain on a numerical pain scale, the PPT test was repeated in the same fashion as was done before the CPT, with three assessments, each with a 20-s interval break. While the nondominant hand remained in the cold water during the CPM test, participants were instructed to remove it from the cold water during the 20-s pause of the PPT. A thermometer was used to monitor the water bath temperature, and ice was added as needed to keep it at the desired level.

2.2 Statistical analysis

Descriptive statistics were used to summarize demographic data. All data are presented as means with standard deviations (SD) or frequencies and percentages. The CPM effect was reported as the percent of change in the average (Avg.) of the three PPT value during CPT and before CPT, which describes the efficacy of pain inhibition. A positive value indicates an increase in threshold, while a negative value indicates a decrease in the threshold:

$$\text{CPM} = \frac{(\text{Avg. of the 3 PPT during CPT} - \text{Avg. of the 3 PPT before CPT}) \times 100}{\text{Avg. of the 3 PPT before CPT}}$$

The absolute reliability was assessed with the standard error of measurement (SEM), the smallest detectable change (SDC), the coefficient of variation (CV) and Bland-Altman analyses, while the intraclass correlation coefficient (ICC) was used to measure the relative reliability; these are all commonly used measures for reliability assessment with continuous data [27, 28, 47, 48].

The SEM was calculated as the standard deviation of the difference divided by the square root of 2, and is considered to be a parameter for the amount of measurement error present in an instrument. The SEM can be used to provide a range around the observed value within which the true, theoretical value, lies. The related absolute reliability value, SDC, or the change in an instruments score beyond measurement error, is calculated as the 1.96 times the square root of 2, times the SEM for an individual, and these SDC values then divided by the square root of the sample size for the SDC of the group. The ICC represents the measurement error relative to the heterogeneity of the subjects [49]. The ICC parameter ranges from 0 to 1, with values closest to 1 indicating the highest reproducibility. An ICC less than 0.4 was considered poor agreement; 0.4–0.59, fair agreement; 0.6–0.75, good agreement; and greater than 0.75, excellent agreement [50]. The ICC was calculated using variance components estimated from a two-factor random effects model, with a person as one factor and visit nested within-person as the other. Variance components were estimated using REML (restricted maximum likelihood), and bootstrap 95% confidence interval for the ICC were constructed based on 10,000 bootstrap simulations [51]. The bootstrap method was also used to construct 95% confidence intervals for the difference in ICC between the three anatomical sites [52]. The CV represents the within-subject standard deviation (i.e. the standard deviation of repeated measures over the same subject) expressed as a percentage of the subjects' average threshold/rating [28, 47, 53]. The within-subject standard deviation was estimated by the square root of the mean square error from the two-factor random effects model and then divided by the mean of the outcome being evaluated to compute the CV (and expressed as a percentage). The CV shows the extent of variability in relation to the mean of a given population. The higher the CV, the greater the dispersion in the variable relative the mean.

Bland-Altman analysis, a measure of absolute reliability, is based on the evaluation of the average vs. the difference of two given measurements, from which the limits of agreement (LoA) can be derived around the average difference (bias). This method of Altman and Bland [54] was used to compute the LoA. The LoA delimits the range within which 95% of the differences between thresholds/ratings in two sessions may be expected to lie, which can be interpreted as the maximum difference that can be expected due to measurement error. Testing if the bias was equal to zero was performed using linear regression and generalized estimating equations with robust standard

errors to account for the multiple pairs of observations per subject [55], and $p < 0.05$ was considered significant.

Data from visits that included a control session were excluded from the reliability calculations. We plan to publish the control session data in a different manuscript, as they are outside the aims of this paper, and had a minimal effect on the present results. The intra-session and *inter*-session reliabilities were calculated based on all subjects, and by gender and age. In addition, intra-session reliability was calculated separately for each visit and *inter*-session reliability for each visit combination. Statistical analyses were conducted using R statistical software, version 3.5.0 [56].

2.3 Sample size

Both Morrow et al. and Olds et al. recommend that "The analysis of the reliability of a test should not be approached as a typical hypothesis-testing problem; as such, significance levels and statistical power are not a primary concern for sample size considerations" [57, 58]. Sample size estimation was based on two parameters: the randomized order of the tested sites, and adequate size for precise estimation of reliability measures. In addition, the inclusion of the analyses into four age-gender groups increased the required size of our sample. A sample size of 60 participants, each tested in three visits, generated a total of 180 visits. To avoid potential bias, all possible orders of testing of the three sites assigned for each visit were counterbalanced. The evaluation of three sites in each session created six possible orders for each visit: side of the face (masseter) (M), hand (H), foot (F) or MHF, then MFH, FMH, FHM, HFM, and HMF. These six combinations were then randomly assigned to each of the 180 visits; resulting in 30 sessions of each order. Of the total six sessions (i.e. three visits) for each subject, the CS in one of the sessions was lukewarm water instead of the CPT. At each visit, the three anatomical sites were tested in the same order in both sessions for a given participant.

3 Results

3.1 Demographics

Sixty-one healthy participants were recruited. One subject did not complete the third visit, therefore data collected from 60 subjects were included in the analysis; 38 women (age

Table 1: Characteristics of the study population.

Description	Women $n=38$	Men $n=22$
Age (mean ± SD) year	37.6 ± 15.1	34.2 ± 14.7
Race, n (%)		
White	15 (39.5)	10 (45.5)
Asian	10 (26.3)	10 (45.5)
Black	5 (13.1)	2 (9.1)
Native Hawaiian	1 (2.6)	0 (0)
N/A	4 (10.7)	0 (0)
Others/more than 1 race	3 (7.8)	0 (0)
Ethnicity, n (%)		
Non-hispanic	35 (92.1)	20 (91)
Hispanic	3 (7.8)	1 (4.5)
N/A	0 (0)	1 (4.5)
Level of education, n (%)		
High school or less	2 (5.3)	3 (13.6)
Some college or more	36 (94.7)	19 (86.4)

37.6 ± 15.1 years) and 22 men (age 34.2 ± 14.7 years) (Table 1). The mean number of days between the 1st and 2nd visit was 3.5 ± 1.4 days, and 3.1 ± 2.8 between the 2nd and 3rd visit. The mean session duration was 22 ± 4.6 min, and one visit (which included two sessions separated by a 15-min break) lasted an average of 62.4 ± 9.2 min. All participants reported that they reached a conditioning pain intensity of 7/10 on NPS before the PPT application to the tested site. The mean water bath temperature during the CPT was 5.2 ± 0.05 °C.

3.2 Detection of PPT and CPM effect

The PPT was recorded at all three tested anatomical sites in all three visits for each participant, regardless of the CPT application. All PPT values are presented as mean ± SD of three PPT readings before and during CPT application in Table 2. The foot showed the highest-pressure pain threshold, particularly during the CPT application (585.4 ± 199.7 kPa), followed by the hand, then the side of the face.

The CPM effect was induced at all three tested anatomical sites in all three visits for each participant, with no significant differences among all sites. The CPM magnitude is presented as the mean ± SD of the percent of change between the PPT value during and before CS application. Data from all CPT sessions were pooled and used for analysis (Table 3), and additional data of the CPM magnitude by each session is reported in Table S1. The highest mean CPM magnitude for all subjects was detected on the hand $(23.0 ± 21.6)$ $(p=0.34)$, followed by the side of the face $(19.2 ± 21.3)$ $(p=0.19)$, and the foot $(19.1 ± 23.1)$ $(p=0.075)$,

Table 2: Pressure pain threshold (PPT) (TS) before (baseline) and during CPT application: relative and absolute reliability by anatomical test site.

Test site	PPT mean±SD before (kPa)	PPT mean±SD during (kPa)	Session type	PPT standard deviation (SD)		Relative reliability ICC[a] (95% CI[b])		Bland-Altman analysis bias (lower LoA – upper LoA)[c]		Absolute reliability of PPT					
				PPT before	PPT during	PPT before	PPT during	PPT before	PPT during	CV[d]		Bias p-value		SEM[e]	
										PPT before	PPT during	PPT before	PPT during	PPT before	PPT during
Side of the face	243.4±80.51	287.2±91.23	Intrasession	31.5	36.7	0.85 (0.77, 0.91)	0.84 (0.77, 0.91)	−10.6 (−95.7, 74.5)	3.1 (−98.8, 105)	9.1	9.1	0.013	0.54	22.2	25.9
			Intersession	39.8	45.4	0.77 (0.66, 0.84)	0.76 (0.66, 0.82)	−0.8 (−107.9, 106.2)	9.8 (−112, 132)	11.6	11.3	0.891	0.14	28.2	32.1
Hand	432.7±153.75	522.6±176.63	Intrasession	49.3	59.5	0.90 (0.84, 0.94)	0.89 (0.83, 0.92)	11.4 (−123.9, 146.7)	12.2 (−152, 176.)	8.0	8.1	0.065	0.099	34.8	42.1
			Intersession	63.8	79.0	0.84 (0.74, 0.89)	0.81 (0.71, 0.86)	−10.1 (−186.6, 166.4)	−0.6 (−222, 220.5)	10.4	10.7	0.296	0.963	45.1	55.9
Foot	550.9±166.62	585.4±199.73	Intrasession	67.0	36.7	0.84 (0.75, 0.88)	0.86 (0.78, 0.90)	−22.0 (−203.5, 159.5)	18.6 (−194, 231)	9.5	9.4	0.011	0.069	47.4	54.8
			Intersession	85.5	45.4	0.73 (0.58, 0.83)	0.71 (0.56, 0.82)	−14.2 (−247.1, 218.8)	−8.4 (−309, 292)	12.1	13.2	0.271	0.628	60.5	77.5

[a]ICC, intraclass correlation coefficient.

[b]CI, confidence interval.

[c]Bias, mean difference; lower LoA – upper LoA, limits of agreement (lower boundary, upper boundary).

[d]CV, coefficient of variation, (CPM standard deviation by session type/√2/mean CPM×100%.

[e]SEM, standard error of measurement = SD of difference/√2.

with no significant difference across all three sites (Table 3). Overall, the mean of the magnitude of CPM of all three sites in the CPT session was (20.9 ± 2.5).

3.3 CPM and PPT reliability

3.3.1 PPT reliability

3.3.1.1 Relative reliability

Intraclass correlations (ICC) values were calculated to examine the relative reliability of the PPT with 95% CI. All ICC values were consistently high when the PPT test was performed on all three sites either before or during the application of the CPT. The intra-session and *inter*-session reliability ranged from good to excellent, and the hand was slightly superior to the other two sites specifically; the intra-session reliability before the CPT application was ICC = 0.90 (Table 2).

3.3.1.2 Absolute reliability

The PPT consistently showed low intra-session and *inter*-session variability (lower CV%) when performed on all three sites, either before or during the CPT application. In line with the relative reliability findings, the intra-session reliability was the highest (lowest CV) before and during the CPT (CV = 8.0 and 8.1), respectively, when the PPT was performed on the hand followed by the foot and the side of the face (Table 2). The Bland-Altman analysis of the PPT difference between visits showed an average near zero (with 95% confidence intervals) for all three sites, indicating low bias and high agreement.

3.3.2 CPM reliability

3.3.2.1 Relative reliability

The ICC values were consistently higher when the CPM test was performed on the hand than the foot and the side of the face (Tables 3–5). The overall intra-session and *inter*-session reliability were fair when performed on the hand, ICC = 0.57 and 0.40, respectively, and poor when performed on the foot (0.22 and 0.15) and the side of the face (0.20 and 0.13) (Table 3).

The CPM test continued to show higher reliability when performed on the hand in both age groups, with fair intra-session reliability in the younger group, ICC = 0.53, but poor *inter*-session reliability, ICC = 0.30; good intra-session reliability in the older group ICC = 0.60, and fair *inter*-session reliability, ICC = 0.46 (Table 4).

Table 3: Conditioned pain modulation (CPM) test magnitude, and relative and absolute reliability by anatomical test site.

Test site	CPM mean%±SD	Session type	CPM standard deviation (SD)	Relative reliability of CPM, ICC[a] (95% CI[b])	Absolute reliability of CPM			
					Bland-Altman analysis bias (lower LoA – upper LoA)[c]	CV[d]	Bias p-value	SEM[e]
Side of the face	19.2 ±21.3	Intrasession	19.0	0.20 (0.03, 0.39)	7.3 (−43.6, 58.4)	70.1	0.002	13.5
		Intersession	19.8	0.13 (0.00, 0.29)	6.7 (−45.1, 58.5)	73.1	0.008	14.0
Hand	23.0±21.6	Intrasession	14.1	0.57 (0.37, 0.71)	−0.1 (−39.4, 39.3)	43.5	0.976	10.0
		Intersession	16.8	0.40 (0.23, 0.55)	3.9 (−42.4, 50.2)	51.5	0.104	11.8
Foot	19.1±23.1	Intrasession	20.5	0.22 (0.01, 0.46)	8.8 (−45.3, 63.1)	75.9	0.001	14.5
		Intersession	21.3	0.15 (0.00, 0.32)	2.0 (−57.7, 61.7)	78.9	0.475	15.0

[a]ICC, intraclass correlation coefficient.
[b]CI, confidence interval.
[c]Bias, mean difference; lower LoA – upper LoA, limits of agreement (lower boundary, upper boundary).
[d]CV, coefficient of variation, (CPM standard deviation by session type/$\sqrt{2}$)/mean CPM×100%.
[e]SEM, standard error of measurement = SD of difference/$\sqrt{2}$.

Table 4: Conditioned pain modulation (CPM) test magnitude, and relative and absolute reliability by anatomical test site stratified by gender and age.

Test site	CPM mean%±SD	Session type	CPM standard deviation (SD)	Relative reliability of CPM, ICC[a] (95% CI[b])	Absolute reliability of CPM			
					Bland-Altman analysis Bias (lower LoA – upper LoA)[c]	CV[d]	Bias p-value	SEM[e]
Age group 18–30 (n = 27)								
Side of the face	22.7±21.4	Intrasession	21.4	0.00 (0.00, 0.20)	12.3 (−44.3, 68.9)	66.8	0.002	15.2
		Intersession	21.4	0.00 (0.00, 0.17)	5.5 (−50.3, 61.4)	66.8	0.142	15.2
Hand	25.0±19.6	Intrasession	13.5	0.53 (0.21, 0.76)	1.1 (−36.5, 38.7)	38.1	0.694	9.5
		Intersession	16.5	0.30 (0.12, 0.45)	7.0 (−35.4, 49.4)	46.6	0.038	11.6
Foot	20.7±19.4	Intrasession	19.2	0.02 (0.00, 0.31)	9.8 (−40.2, 59.9)	65.5	0.003	13.6
		Intersession	19.3	0.01 (0.00, 0.13)	1.8 (−51.5, 55.0)	66.0	0.638	13.7
Age group 31–65 (n = 33)								
Side of the face	16.3±20.7	Intrasession	16.2	039 (0.11, 0.60)	3.4 (−41.4, 48.2)	70.2	0.245	11.5
		Intersession	18.6	0.20 (0.00, 0.43)	7.6 (−41.0, 56.1)	80.7	0.024	13.2
Hand	21.4±23.1	Intrasession	14.7	0.60 (0.34, 0.75)	−1.0 (−41.9, 39.9)	48.5	0.664	10.4
		Intersession	17.0	0.46 (0.19, 0.65)	1.3 (−47.6, 50.3)	56.2	0.686	12.0
Foot	17.7±25.8	Intrasession	21.4	0.31 (0.02, 0.59)	8.2 (−49.3, 65.9)	85.7	0.011	15.2
		Intersession	22.7	0.22 (0.00, 0.43)	2.2 (−62.6, 67.1)	90.9	0.588	16.1
Women (n = 38)								
Side of the face	16.7±21.5	Intrasession	19.8	0.15 (0.00, 0.38)	6.7 (−47.0, 60.4)	83.9	0.039	14.0
		Intersession	20.5	0.09 (0.00, 0.29)	7.9 (−45.5, 61.2)	86.9	0.014	14.5
Hand	19.5±21.9	Intrasession	13.3	0.63 (0.37, 0.79)	−0.1 (−37.3, 37.1)	48.3	0.967	9.4
		Intersession	17.2	0.38 (0.17, 0.57)	2.8 (−44.3, 50.0)	62.4	0.385	12.2
Foot	17.2±23.7	Intrasession	21.1	0.21 (0.00, 0.45)	9.3 (−47.6, 66.2)	86.7	0.001	14.9
		Intersession	21.1	0.21 (0.00, 0.41)	0.6 (−59.2, 60.4)	86.7	0.863	14.9
Men (n = 22)								
Side of the face	23.5±20.2	Intrasession	17.6	0.25 (0.02, 0.55)	8.5 (−37.7, 54.9)	53.0	0.017	12.4
		Intersession	18.6	0.16 (0.00, 0.37)	4.6 (−44.7, 53.8)	55.9	0.259	13.1
Hand	29.0±19.9	Intrasession	15.4	0.41 (0.08, 0.64)	0.0 (−43.2, 43.2)	37.6	0.995	10.9
		Intersession	15.9	0.37 (0.06, 0.53)	5.9 (−39.3, 50.7)	38.7	0.079	11.2
Foot	22.3±21.9	Intrasession	18.7	0.27 (0.00, 0.68)	8.1 (−41.7, 57.9)	59.2	0.031	13.2
		Intersession	21.9	0.00 (0.00, 0.14)	4.5 (−55.3, 64.4)	69.5	0.381	15.5

[a]ICC, intraclass correlation coefficient.
[b]CI, confidence interval.
[c]Bias, mean difference; lower LoA – upper LoA, limits of agreement (lower boundary, upper boundary).
[d]CV, coefficient of variation, (CPM standard deviation by session type/$\sqrt{2}$/mean CPM x 100%
[e]SEM, standard error or measurement = SD of difference/$\sqrt{2}$.

By gender, the intra-session reliability was good, ICC = 0.63, and *inter*-session reliability was poor, ICC = 0.38 when performed on the hand in women (n = 38) (Table 4). Men (n = 22) showed fair intra-session, ICC = 0.41, and poor *inter*-session reliability, ICC = 0.37, when it was performed on the hand. Both the intra-session and *inter*-session reliability continued to be poor when applied to the foot and the side of the face in both age groups and both genders (Table 4).

By session, the CPM test continued to be most reliable when performed on the hand, primarily at early sessions, intra-session ICC = 0.75. The highest *inter*-session reliability occurred on the hand from the 2nd to the 3rd visit ICC = 0.66 (Table 5).

3.3.2.2 Absolute reliability

The CPT stimuli consistently evoked a smaller intra-session and *inter*-session variability (lower CV%) CV = 43.51

and 51.5, respectively, when the CPM test was performed on the hand than foot and the side of the face (Table 3). In line with the relative reliability findings, the CPM test showed lower intra-session and *inter*-session variability when performed on the hand in both age groups, CV = 38.1 and 46.6, respectively, for younger group, CV = 48.5 and 56.2, respectively for the older group (Table 4), and both genders, CV = 48.3 and 62.4 for women, and CV = 37.6 and 38.7 for men (Table 4). The test continued to be most reliable when performed on the hand, primarily at early sessions (2nd session), lowest intra-session variability CV = 73.1. The lowest *inter*-session variability occurred on the hand from the 1st to the 2nd visit CV = 44 (Table 5).

The Bland-Altman analysis in Table 3 showed the hand demonstrating the lowest bias and range for the 95% limits of agreement (LoA) for the intra-session reliability [−0.1; (−39.4, 39.3)] and second best for the *inter*-session reliability [3.9, (−42.4, 50.2)], compared to the foot (8.8 intra-session reliability and 2.0 *inter*-session reliability,

Table 5: Conditioned pain modulation (CPM) test magnitude, and relative and absolute reliability by anatomical test site at three different time points.

Test site	Time point	CPM mean%±SD	CPM standard deviation (SD)	Relative reliability of CPM, ICC[a] (95% CI[b])	Absolute reliability of CPM		
					Bland-Altman analysis bias (lower LoA – upper LoA)[c]	CV[d]	Bias p-value	SEM[e]
Intrasession								
Side of the face	Visit 1	23.0±23.8	23.3	0.05 (0.00, 0.26)	12.4 (−47.9, 72.8)	75.6	0.009	16.4
	Visit 2	18.1±20.5	17.9	0.24 (0.00, 0.58)	5.9 (−43.0, 54.9)	70.1	0.127	12.7
	Visit 3	18.3±21.5	19.9	0.14 (0.01, 0.27)	3.6 (−37.2, 44.4)	77.1	0.278	14.1
Hand	Visit 1	26.8±22.1	15.4	0.51 (0.06, 0.75)	4.9 (−37.3, 47.2)	40.8	0.136	10.9
	Visit 2	20.9±21.1	10.9	0.75 (0.55, 0.85)	−2.0 (−32.5, 28.5)	37.1	0.417	7.7
	Visit 3	21.9±22.2	17.7	0.36 (0.23, 0.46)	−3.3 (−46.4, 39.7)	57.2	0.334	12.6
Foot	Visit 1	19.4±23.4	19.4	0.31 (0.00, 0.61)	4.4 (−49.5, 58.3)	71.0	0.300	13.8
	Visit 2	21.6±25.4	24.4	0.08 (0.00, 0.50)	12.9 (−50.6, 76.3)	79.9	0.011	17.2
	Visit 3	18.6±26.9	25.7	0.09 (0.00, 0.26)	9.6 (−33.6, 52.7)	97.5	0.006	18.2
Intersession								
Side of the face	Visits 1 and 2	19.0±19.4	19.4	0.00 (0.00, 0.13)	2.7 (−46.7, 52.1)	72.0	0.501	13.7
	Visits 1 and 3	19.7±23.4	22.1	0.11 (0.00, 0.32)	12.7 (−43.3, 68.6)	79.4	0.006	15.6
	Visits 2 and 3	18.8±21.2	17.3	0.34 (0.00, 0.72)	7.1 (−39.2, 53.3)	64.8	0.061	12.2
Hand	Visits 1 and 2	21.6±15.9	13.5	0.29 (0.03, 0.44)	2.8 (−34.7, 40.2)	44.0	0.424	9.5
	Visits 1 and 3	26.4±21.6	20.6	0.10 (0.00, 0.28)	0.1 (−39.9, 40.0)	55.1	0.982	14.6
	Visits 2 and 3	20.9±26.4	15.5	0.66 (0.41, 0.80)	−3.2 (−44.1, 37.5)	52.5	0.260	11.0
Foot	Visits 1 and 2	20.9±22.9	20.2	0.23 (0.00, 0.39)	5.9 (−45.1, 56.8)	68.3	0.111	14.3
	Visits 1 and 3	16.9±21.2	21.2	0.00 (0.00, 0.04)	8.6 (−50.0, 67.3)	88.7	0.036	15.0
	Visits 2 and 3	19.2±25.2	21.6	0.28 (0.00, 0.59)	12.5 (−41.0, 66.0)	79.5	0.002	15.3

[a]ICC, intraclass correlation coefficient.
[b]CI, confidence interval.
[c]Bias, mean difference; lower LoA – upper LoA, limits of agreement (lower boundary, upper boundary).
[d]CV, coefficient of variation, (CPM standard deviation by session type/$\sqrt{2}$)/mean CPM×100%.
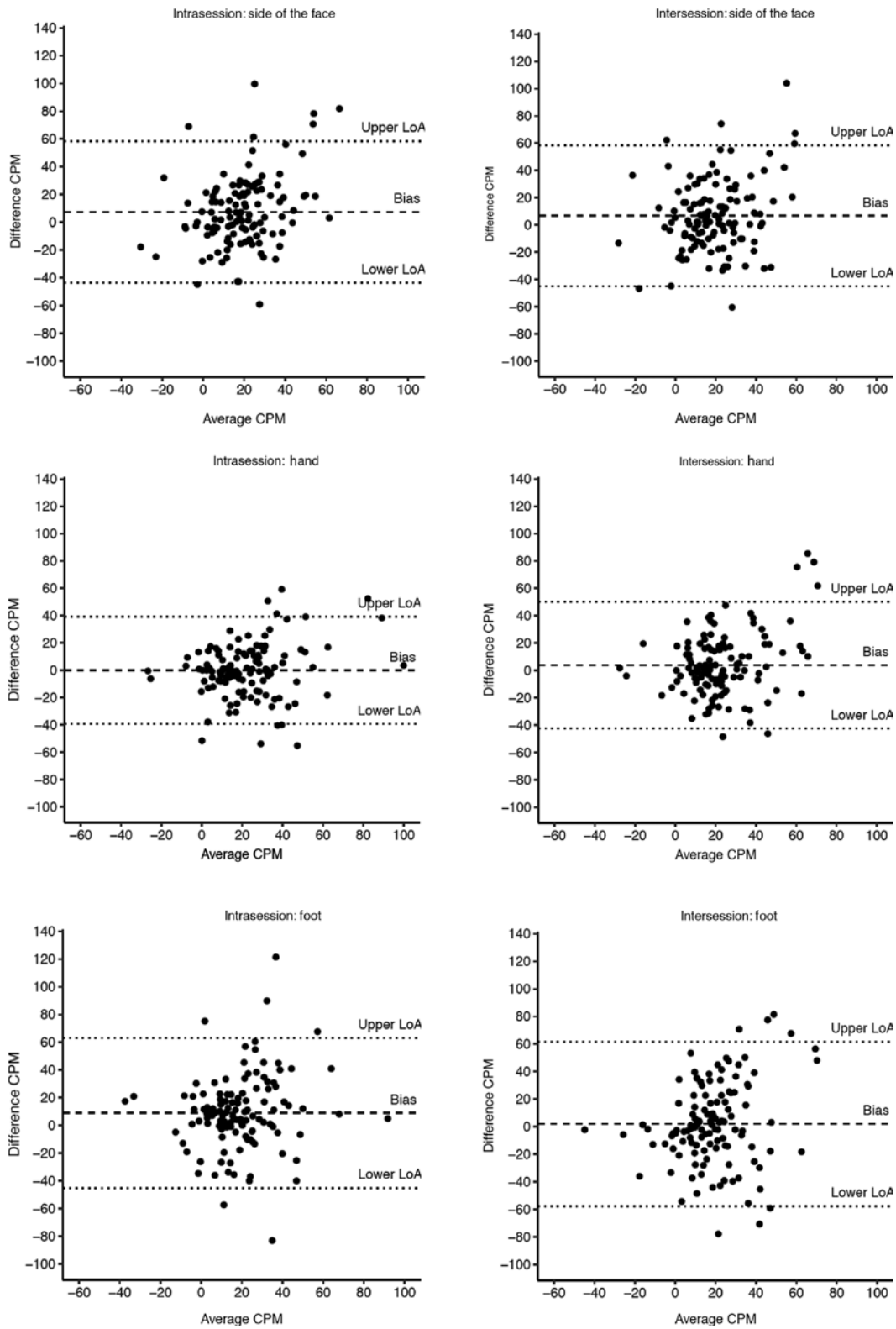[e]SEM, standard error or measurement = SD of difference/$\sqrt{2}$.

Fig. 3: Bland-Altman plots display the mean of CPM on the x-axis, and the difference between CPM tests on the y-axis, for both intra-session and *inter*-session measures for each of the anatomical sites. The mean difference in the visits (bias) is marked as the heavy dashed line, with the line at zero for no bias; the estimated 95% limits of agreement (LoA) are shown as thin dotted lines in which 95% of the data are expected to lie. The lower the limits, the closer the values between both measurements and the better the reliability.

and face (7.3 intra-session reliability and 6.7 *inter*-session reliability, respectively). The corresponding Bland-Altman plots (Fig. 3) display the relationship between the mean of CPM on the x-axis, and the difference between CPM at each visit on the y-axis, for both intra-session and *inter*-session measures for each one of the tested anatomical sites. In these plots, the mean difference between measurements (bias) as noted above is marked as the dashed line, and the estimated 95% limits of agreement (LoA) as thin dashed lines in which 95% of the data are expected to lie. The lower the limits, the closer the values between both measurements and the better the reliability. These plots in Fig. 3 show all the measured values distributed close to the mean difference of zero from one session to another, meaning there is minimal bias between measurements, which indicates high agreement i.e. good reliability. The lowest bias values and 95% CI were for the intra-session of the hand (−0.1), *inter*-session of the foot (2.0) and *inter*-session of the hand (3.9). The hand measures show a smaller LoA and fewer large outliers than the foot and face. A few subjects for each of the six measures (range −1 to 5) were either above or below the 95% limit of agreement line.

As shown in Table 2, the SEM for the CPM effect of the PPT before (34.8 kPa) and during (42.1 kPa) the CPM in the hand intra-session group means that one can be 68% confident (±1 SEM) that the "true" PPT value of a subject can be found between 398 and 467 kPa before, and between 481 kPa and 565 kPa during the CPT. The SDC for the individual for the CPM effect of the PPT for this same group would be 96.2 kPa before and 116 kPa during CPM. This would mean that the PPT from CPM would need to change at least 116 kPa before the observed change can be considered a true change in the PPT related CPM.

4 Discussion

To our knowledge, this is the first study to specifically compare the reliability of the CPM test paradigm among three anatomical test sites. The CPM test was consistently more reliable when the TS was performed on the dominant hand and the CS performed on the contralateral hand than at the side of the face or foot test sites (Table 3).

The PPT test alone showed good to excellent reliability in all anatomical sites. This indicates that the PPT measurement, by itself, is highly repeatable, but the coupling of the PPT before and during cold pressor as calculated as the CPM is more variable. This finding may be due to pain modulation as a more complex psychophysical phenomenon, influenced by many factors. In subgroup analyses, the test was more reliable in women than men, and more

reliable in older than younger subjects. Also, the study findings showed the *inter*-session reliability – over days – to be consistently lower than intra-session (same day) reliability, where the interval between tests was less than half an hour.

Previous CPM test reliability findings have been inconsistent, ranging from poor [13–15, 19, 22–24, 28, 30] reported as ICC of −0.40 [14] to excellent [14, 20] ICC of 0.85 [14]. This variation is likely due to multiple factors, such as: (1) differences in methodological approaches, including the type of stimuli used for CS and TS, and the anatomical sites for both stimuli; (2) different subject populations, e.g. healthy individuals vs. patients; (3) different reported reliability measures, e.g. ICC vs. CV; and (4) the possibility that human endogenous pain modulation itself may be highly variable from day to day.

A sample size ranging between 25 and 50 participants is usually sufficiently accurate for reliability studies [48]. This study's findings were based on data collected from 60 healthy participants to assure an adequate number of participants in subgroups of both genders and two age groups.

CPM test reliability studies have examined different protocols with time intervals between tests of 15-min to 10 months [14, 15]. The intra-session time intervals have ranged from 2 to 60-min between sessions on the same day [17, 21]. In line with Lewis et al., a 15-min break between sessions was found to be sufficient to allow the pain system to reestablish its baseline status [14]. A carry-over effect may be a factor in reliability studies. Reviewing previous studies, Cathcart et al. waited for 2-min at each location after the previous PPT was taken, and to establish their baseline for CPM they waited for 5 min after completion of the previous TS assessment [21]. Imai et al. also waited for 5 min within the same session [22]. Graven-Nielsen et al. waited for 5 min between the baseline TS and the CS [24], and Valencia et al. waited for 2-min [17]. In the current study, there was a break of 5–8 min until subjects reported that the cold sensation completely faded away before testing the next site. Therefore, the time between tests for the present study was well within the time intervals of previous work.

CPM testing recommendations were published in 2015 [44] while the current research was conducted beginning in 2014. Nevertheless, the fundamentals of CPM testing "pain inhibits pain" paradigm can be done through a variety of protocols that usually generate similar results. Granovsky et al., compared different paradigms and reported higher reliability with single test stimulus protocol [30]. Among the different types of conditioning stimuli. CPT was the most frequently used and one of the most efficient to induce CPM. In particular, when combined with

PPT as TS using manual algometry, it shows good test-retest reliability [46, 59].

Previous reliability studies – which may have not compared the reliability among different anatomical sites- have provided data that may potentially explain the superiority of the hand as shown in the present study. Graven-Nielsen et al. showed that the highest point estimates of inter-rater reliability occurred when both the TS and CS were in the same body location but on opposite sides, and generally lower when the TS and CS were in different locations [24]. Also, TS and CS (via tight cuffs) showed the highest reliability while on the lower but opposite legs, versus the TS on the right thigh and CS on the left upper arm, or when the TS was on the right upper arm and CS on left lower leg. In an earlier study, Oono et al. explored the reliability among different sites where the forearm and orofacial regions showed the smallest intra-individual and inter-individual variability [16], specifically with the CS applied to a close contralateral segment, compared to a distant site, which was the leg. Taken together with these two previous studies, our work suggests that the application of the CS to the contralateral dermatome of the TS may enhance CPM reliability.

Another explanation of why the CPM reliability was highest on the hand is that the concomitant CPM magnitude was also highest in this region (23.0%) vs. the foot (19.1%) and side of the face (19.2%). If the CPM effect is more robust, the variability from test to test may be lessened in comparison. Graven-Nielsen et al. hypothesized a similar reason – greater CPM magnitude – greater CPM reliability [24]. However, in subgroups analyses for CPM test location by age and gender, men at the hand site showed higher CPM magnitude than women (29% vs. 19.5%) yet lower relative CPM intra-session and *inter*-session reliability. Absolute reliability, however, was consistent with the high CPM magnitude (Table 4). More work is needed to determine why same dermatome TS-CS may have both higher CPM magnitude and greater reliability than other methods.

Our findings also suggest that the CPM test may lack reliability over the 1–2-week time period, either between visits or even between sessions, especially when performed multiple times in a short time frame. Martel et al. found the CPM test to be more stable in female patients over a short time period (10 days) vs. months [18]. In addition, CPM has previously been shown to lack reliability and stability over long periods of time, such as 7–10 months [15]. Graven-Nielsen et al. showed that the TS reliability over 1-month was high; while the CPM reliability was low [24]. These results support the concept that CPM may be an unstable trait in some people [27] and that numerous factors may contribute to the test's lack of temporal stability over extended periods [60].

Personal factors such as age have been known to affect CPM efficiency in healthy individuals. The CPM was reported to be significantly higher in younger than older individuals [34, 60, 61]. In terms of CPM reliability, our results showed the CPM test to be more reliable in middle-age/older (31–65 years) than in younger (18–30 years) individuals. This may be related to the finding that older adults showed less inhibition of cold stimulation [34, 62], allowing older people to possibly have a less aversive pain experience and thereby allowing more precise interpretation of that experience. Also, the foot was the least sensitive site to pain in older adults, which may be due to lower extremities in older adults being less sensitive to pain [63, 64], especially to a PPT test [65].

This study's findings appear to have more important implications for CPM testing methodology in general, rather than clinical implications. CPM may be assessed using different protocols fundamentally based on the "pain inhibits pain" concept. Therefore, incorporating others' findings (e.g. testing the hand in this case) in additional research on CPM test reliability is needed before the potential use of CPM as a diagnostic test. Also, the lack of CPM test reliability in men may limit its use in clinical settings, and the rather large SDC in PPT may limit the overall utility of CPM use in the clinic.

The strengths of this study should be noted. First, the larger-than-usual sample size allowed for more precise estimation of both absolute and relative reliability and a robust comparison between the anatomical sites. Second, a wide age range of both women and men were tested, increasing the potential generalizability and allowing for intra-study comparison of age on reliability. Third, subjects were blinded as to how CPM works, minimizing the chance they could influence the outcome of the tests.

This study has limitations. All healthy subjects had multiple tests performed within a relatively short time period – three tests at each anatomical site without the CS, then three more tests with the CS, followed by a short break, and then a repeat of these tests. This was repeated 2–7 days later, and then once more after another 2–7 days. Thus, the large number of tests completed in each subject in a relatively short amount of time may have caused pain adaptation [66, 67]. Other effects such as fatigue of testing, lack of attention to the detailed instructions may all have played a role to alter the repeatability of the findings, and make them less generalizable. In addition, further stratifying each anatomical site by gender and two age groups created smaller subgroups thereby inducing lower statistical power.

5 Conclusion

The CPM test showed superior reliability – both relative and absolute – when the PPT (TS) was performed on the hand and the CPT (CS) done on the contralateral hand, when compared to the foot and face sites. Additional work to reproduce these findings and clarify the specific role of the anatomical site of stimuli in the reliability of CPM appears warranted.

References

[1] Le Bars D, Dickenson AH, Besson JM. Diffuse noxious inhibitory controls (DNIC). I. Effects on dorsal horn convergent neurones in the rat. Pain 1979;6:283–304.

[2] Yarnitsky D, Arendt-Nielsen L, Bouhassira D, Edwards RR, Fillingim RB, Granot M, Hansson P, Lautenbacher S, Marchand S, Wilder-Smith O. Recommendations on terminology and practice of psychophysical DNIC testing. Eur J Pain 2010;14:339.

[3] Lewis GN, Rice DA, McNair PJ. Conditioned pain modulation in populations with chronic pain: a systematic review and meta-analysis. J Pain 2012;13:936–44.

[4] Granovsky Y, Yarnitsky D. Personalized pain medicine: the clinical value of psychophysical assessment of pain modulation profile. Rambam Maimonides Med J 2013;4:e0024.

[5] Reimer M, Hüllemann P, Hukauf M, Keller T, Binder A, Gierthmühlen J, Baron R. Prediction of response to tapentadol in chronic low back pain. Eur J Pain 2017;21:322–33.

[6] Landau R, Kraft JC, Flint LY, Landau R, Kraft JC, Flint LY, Carvalho B, Richebé P, Cardoso M, Lavand'homme P, Granot M, Yarnitsky D, Cahana A. An experimental paradigm for the prediction of Post-Operative Pain (PPOP). J Vis Exp 2010;35:1671.

[7] Wilder-Smith OH, Schreyer T, Scheffer GJ, Arendt-Nielsen L. Patients with chronic pain after abdominal surgery show less preoperative endogenous pain inhibition and more postoperative hyperalgesia: a pilot study. J Pain Palliat Care Pharmacother 2010;24:119–28.

[8] Yarnitsky D, Crispel Y, Eisenberg E, Granovsky Y, Ben-Nun A, Sprecher E, Best LA, Granot M. Prediction of chronic post-operative pain: pre-operative DNIC testing identifies patients at risk. Pain 2008;138:22–8.

[9] Eisenberg E, Midbari A, Haddad M, Pud D. Predicting the analgesic effect to oxycodone by 'static' and 'dynamic' quantitative sensory testing in healthy subjects. Pain 2010;151:104–9.

[10] Granot M, Lowenstein L, Yarnitsky D, Tamir A, Zimmer EZ. Post-cesarean section pain prediction by preoperative experimental pain assessment. Anesthesiology 2003;98:1422–6.

[11] Nir RR, Yarnitsky D. Conditioned pain modulation. Curr Opin Support Palliat Care 2015;9:131–7.

[12] Dunn WN. Measuring knowledge use. Knowledge: Creation, Diffusion, Utilization 1983;5:120–33.

[13] Olesen SS, van Goor H, Bouwense SA, Wilder-Smith OH, Drewes AM. Reliability of static and dynamic quantitative sensory testing in patients with painful chronic pancreatitis. Reg Anesth Pain Med 2012;37:530–6.

[14] Lewis GN, Heales L, Rice DA, Rome K, McNair PJ. Reliability of the conditioned pain modulation paradigm to assess endogenous inhibitory pain pathways. Pain Res Manag 2012;17:98–102.

[15] Wilson H, Carvalho B, Granot M, Landau R. Temporal stability of conditioned pain modulation in healthy women over four menstrual cycles at the follicular and luteal phases. Pain 2013;154:2633–8.

[16] Oono Y, Nie H, Lima Matos R, Wang K, Arendt-Nielsen L. The inter- and intra-individual variance in descending pain modulation evoked by different conditioning stimuli in healthy men. Scand J Pain 2011;2:162–9.

[17] Valencia C, Kindler LL, Fillingim RB, George SZ. Stability of conditioned pain modulation in two musculoskeletal pain models: investigating the influence of shoulder pain intensity and gender. BMC Musculoskelet Disord 2013;14:182.

[18] Martel MO, Wasan AD, Edwards RR. Sex differences in the stability of conditioned pain modulation (CPM) among patients with chronic pain. Pain Med 2013;14:1757–68.

[19] Costa YM, Morita-Neto O, de Araújo-Júnior EN, Sampaio FA, Conti PC, Bonjardim LR. Test-retest reliability of quantitative sensory testing for mechanical somatosensory and pain modulation assessment of masticatory structures. J Oral Rehabil 2017;44:197–204.

[20] Bossmann T, Brauner T, Lowak H, Anton F, Forster C, Horstmann T. Reliability of conditioned pain modulation for the assessment of endogenous pain control pathways. Neurol Psychiatry Brain Res 2016;22:155–61.

[21] Cathcart S, Winefield AH, Rolan P, Lushington K. Reliability of temporal summation and diffuse noxious inhibitory control. Pain Res Manag 2009;14:433–8.

[22] Imai Y, Petersen KK, Mørch CD, Arendt Nielsen L. Comparing test-retest reliability and magnitude of conditioned pain modulation using different combinations of test and conditioning stimuli. Somatosens Mot Res 2016;33:169–77.

[23] Marcuzzi A, Wrigley PJ, Dean CM, Adams R, Hush JM. The long-term reliability of static and dynamic quantitative sensory testing in healthy individuals. Pain 2017;158:1217–23.

[24] Graven-Nielsen T, Izumi M, Petersen KK, Arendt-Nielsen L. User-independent assessment of conditioning pain modulation by cuff pressure algometry. Eur J Pain 2017;21:552–61.

[25] Gehling J, Mainka T, Vollert J, Pogatzki-Zahn EM, Maier C, Enax-Krumova EK. Short-term test-retest-reliability of conditioned pain modulation using the cold-heat-pain method in healthy subjects and its correlation to parameters of standardized quantitative sensory testing. BMC Neurol 2016;16:125.

[26] Vuilleumier PH, Biurrun Manresa JA, Ghamri Y, Mlekusch S, Siegenthaler A, Arendt-Nielsen L, Curatolo M. Reliability of quantitative sensory test in a low back pain population. Reg Anesth Pain Med 2015;40:665–73.

[27] Lie MU, Matre D, Hansson P, Stubhaug A, Zwart JA, Nilsen KB. A tonic heat test stimulus yields a larger and more reliable conditioned pain modulation effect compared to a phasic heat test stimulus. Pain Rep 2017;2:e626.

[28] Biurrun Manresa JA, Fritsche R, Vuilleumier PH, Oehler C, Mørch CD, Arendt-Nielsen L, Andersen OK, Curatolo M. Is the conditioned pain modulation paradigm reliable? A test-retest assessment using the nociceptive withdrawal reflex. PLoS One 2014;9:e100241.

[29] Jurth C, Rehberg B, von Dincklage F. Reliability of subjective pain ratings and nociceptive flexion reflex responses as measures of conditioned pain modulation. Pain Res Manag 2014;19:93–6.

[30] Granovsky Y, Miller-Barmak A, Goldstein O, Sprecher E, Yarnitsky D. CPM test-retest reliability: "standard" vs "single test-stimulus" protocols. Pain Med 2016;17:521–9.

[31] Terkelsen AJ, Andersen OK, Hansen PO, Jensen TS. Effects of heterotopic- and segmental counter-stimulation on the nociceptive withdrawal reflex in humans. Acta Physiol Scand 2001;172:211–7.

[32] Haefeli J, Kramer JL, Blum J, Curt A. Heterotopic and homotopic nociceptive conditioning stimulation: distinct effects of pain modulation. Eur J Pain 2014;18:1112–9.

[33] Rosen A, Feldreich A, Dabirian N, Ernberg M. Effect of heterotopic noxious conditioning stimulation on electrical and pressure pain thresholds in two different anatomical regions. Acta Odontol Scand 2008;66:181–8.

[34] Edwards RR, Fillingim RB, Ness TJ. Age-related differences in endogenous pain modulation: a comparison of diffuse noxious inhibitory controls in healthy older and younger adults. Pain 2003;101:155–65.

[35] Riley 3rd JL, King CD, Wong F, Fillingim RB, Mauderli AP. Lack of endogenous modulation and reduced decay of prolonged heat pain in older adults. Pain 2010;150:153–60.

[36] Ge HY, Madeleine P, Arendt-Nielsen L. Sex differences in temporal characteristics of descending inhibitory control: an evaluation using repeated bilateral experimental induction of muscle pain. Pain 2004;110:72–8.

[37] Goodin BR, McGuire L, Allshouse M, Stapleton L, Haythornthwaite JA, Burns N, Mayes LA, Edwards RR. Associations between catastrophizing and endogenous pain-inhibitory processes: sex differences. J Pain 2009;10:180–90.

[38] Serrao M, Rossi P, Mandrini G, Parisi L, Amabile GA, Nappi G, Pierelli F. Effects of diffuse noxious inhibitory controls on temporal summation of the RIII reflex in humans. Pain 2004;112:353–60.

[39] Staud R, Robinson ME, Vierck CJ, Price DD. Diffuse noxious inhibitory controls (DNIC) attenuate temporal summation of second pain in normal males but not in normal females or fibromyalgia patients. Pain 2003;101:167–74.

[40] Weissman-Fogel I, Sprecher E, Pud D. Effects of catastrophizing on pain perception and pain modulation. Exp Brain Res 2008;186:79–85.

[41] Lucas N, Macaskill P, Irwig L, Moran R, Rickards L, Turner R, Bogduk N. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). BMC Med Res Methodol 2013;13:111.

[42] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Jan P. Vandenbroucke[f] STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Int J Surg 2014;12:1495–9.

[43] Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud 2011;48:661–71.

[44] Yarnitsky D, Bouhassira D, Drewes AM, Fillingim RB, Granot M, Hansson P, Landau R, Marchand S, Matre D, Nilsen KB, Stubhaug A, Treede RD, Wilder-Smith OH. Recommendations on practice of conditioned pain modulation (CPM) testing. Eur J Pain 2015;19:805–6.

[45] Moont R, Pud D, Sprecher E, Sharvit G, Yarnitsky D. 'Pain inhibits pain' mechanisms: is pain modulation simply due to distraction? Pain 2010;150:113–20.

[46] Locke D, Gibson W, Moss P, Munyard K, Mamotte C, Wright A. Analysis of meaningful conditioned pain modulation effect in a pain-free adult population. J Pain 2014;15:1190–8.

[47] Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998;26:217–38.

[48] Hopkins WG. Measures of reliability in sports medicine and science. Sports Med 2000;30:1–15.

[49] Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005;19:231–40.

[50] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

[51] Bradley E. Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall, 1993.

[52] Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods, 2nd ed. Thousand Oaks: Sage Publications, 2002.

[53] Dransfield RD, Brightwell R. How to get on top of statistics: design & analysis for biologists, with R. United Kingdom: influential points 2012. Available at: http://influentialpoints.com/Training/coefficient_of_variation.htm.

[54] Altman DG, Bland JM. Commentary on quantifying agreement between two methods of measurement. Clin Chem 2002;48:801–2.

[55] Hardin JW, Hilbe JM. Generalized estimating equations. New York: Chapman Hall/CRC, 2003.

[56] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2018. Available at: https://www.R-project.org/.

[57] Morrow Jr JR, Jackson AW. How "significant" is your reliability? Res Q Exerc Sport 1993;64:352–5.

[58] Olds T. Invited editorial: five errors about error. J Sci Med Sport 2002;5:336–40.

[59] Mutlu EK, Ozdincler AR. Reliability and responsiveness of algometry for measuring pressure pain threshold in patients with knee osteoarthritis. J Phys Ther Sci 2015;27:1961–5.

[60] Hermans L, Van Oosterwijck J, Goubert D, Goudman L, Crombez G, Calders P, Meeus M. Inventory of personal factors influencing conditioned pain modulation in healthy people: a systematic literature review. Pain Pract 2016;16:758–69.

[61] Washington LL, Gibson SJ, Helme RD. Age-related differences in the endogenous analgesic response to repeated cold water immersion in human volunteers. Pain 2000;89:89–96.

[62] Larivière M, Goffaux P, Marchand S, Julien N. Changes in pain perception and descending inhibitory controls start at middle age in healthy adults. Clin J Pain 2007;23:506–10.

[63] Riley 3rd JL, Cruz-Almeida Y, Glover TL, King CD, Goodin BR, Sibille KT, Bartley EJ, Herbert MS, Sotolongo A, Fessler BJ, Redden DT, Staud R, Bradley LA, Fillingim RB. Age and race effects on pain sensitivity and modulation among middle-aged and older adults. J Pain 2014;15:272–82.

[64] Stevens JC, Choo KK. Temperature sensitivity of the body surface over the life span. Somatosens Mot Res 1998;15: 13–28.

[65] Lautenbacher S, Kunz M, Strate P, Nielsen J, Arendt-Nielsen L. Age effects on pain thresholds, temporal summation and spatial summation of heat and pressure pain. Pain 2005;115:410–8.

[66] Greffrath W, Baumgartner U, Treede RD. Peripheral and central components of habituation of heat pain perception and evoked potentials in humans. Pain 2007;132:301–11.

[67] Hollins M, Harper D, Maixner W. Changes in pain from a repetitive thermal stimulus: the roles of adaptation and sensitization. Pain 2011;152:1583–90.