

# Novice assessors demonstrate good intra-rater agreement and reliability when determining pressure pain thresholds; a cross-sectional study

Roland R. Reezigt<sup>1,2</sup>, Geranda E. C. Slager<sup>2</sup>, Michel W. Coppeters<sup>1,3,4</sup> and Gwendolyne G. M. Scholten-Peeters<sup>1</sup>

<sup>1</sup> Department of Human Movement Sciences, Faculty of Behavioural and Movement Sciences, Amsterdam Movement Sciences—Program Musculoskeletal Health, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

<sup>2</sup> Academy of Health, Department of Physiotherapy, Hanze University of Applied Sciences, Groningen, Netherlands

<sup>3</sup> Griffith University, Menzies Health Institute Queensland, Brisbane and Gold Coast, Australia

<sup>4</sup> School of Health Sciences and Social Work, Griffith University, Brisbane and Gold Coast, Australia

## ABSTRACT

**Background:** Experienced assessors show good intra-rater reproducibility (within-session and between-session agreement and reliability) when using an algometer to determine pressure pain thresholds (PPT). However, it is unknown whether novice assessors perform equally well. This study aimed to determine within and between-session agreement and reliability of PPT measurements performed by novice assessors and explored whether these parameters differed per assessor and algometer type.

**Methods:** Ten novice assessors measured PPTs over four test locations (tibialis anterior muscle, rectus femoris muscle, extensor carpi radialis brevis muscle and paraspinal muscles C5-C6) in 178 healthy participants, using either a Somedic Type II digital algometer (10 raters; 88 participants) or a Wagner Force Ten FDX 25 digital algometer (nine raters; 90 participants). Prior to the experiment, the novice assessors practiced PPTs for 3 h per algometer. Each assessor measured a different subsample of ~9 participants. For both the individual assessor and for all assessors combined (*i.e.*, the group representing novice assessors), the standard error of measurement (SEM) and coefficient of variation (CV) were calculated to reflect within and between-session agreement. Reliability was assessed using intraclass correlation coefficients ( $ICC_{1,1}$ ).

**Results:** Within-session agreement expressed as SEM ranged from 42 to 74 kPa, depending on the test location and device. Between-session agreement, expressed as SEM, ranged from 36 to 76 kPa and the CV ranged from 9–16% per body location. Individual assessors differed from the mean group results, ranging from –55 to +32 kPa or from –9.5 to +6.6 percentage points. Reliability was good to excellent ( $ICC_{1,1}$ : 0.87 to 0.95). Results were similar for both types of algometers.

**Conclusions:** Following 3 h of algometer practice, there were slight differences between assessors, but reproducibility in determining PPTs was overall good.

Submitted 25 August 2022

Accepted 22 November 2022

Published 4 January 2023

Corresponding author

Gwendolyne G. M. Scholten-Peeters,  
g.g.m.scholten-peeters@vu.nl

Academic editor

Filipe Clemente

Additional Information and  
Declarations can be found on  
page 15

DOI 10.7717/peerj.14565

© Copyright

2023 Reezigt et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

**Subjects** Anesthesiology and Pain Management, Rehabilitation

**Keywords** Quantitative sensory testing, Pain measurement, Reproducibility, Reliability, Mechanical hyperalgesia, Central sensitisation, Pressure pain threshold, Pain sensitivity, Novice rater, Novice assessor

## INTRODUCTION

Since Victorian times, attempts have been made to quantify pain perception with mechanical stimulus devices (Keele, 1954). The measurements and devices evolved into analogue and digital pressure algometers to obtain pressure pain thresholds (PPTs) (Nussbaum & Downes, 1998). PPT is defined as the minimal applied pressure that induces a painful sensation (Fischer, 1987). PPTs are often used as a static measure of pain sensitivity in various patient populations in research (Amiri et al., 2021) and clinical practice (Ohrbach & Gale, 1989; Nijs et al., 2021). More recently, PPTs also became part of dynamic measures, such as conditioned pain modulation (CPM). In CPM, a baseline PPT is compared to a PPT during or after a conditioned stimulus (Yarnitsky et al., 2015; Reezigt et al., 2021).

A standardised protocol to determine PPTs is available within a Quantitative Sensory Testing (QST) framework to optimise reproducibility (Rolke et al., 2006; Mücke et al., 2016). Reproducibility relates to the degree to which repeated measurements in stable individuals provide similar results (de Vet et al., 2006). Intra-rater reproducibility, where test results of the same rater are analysed, can be further defined into an absolute test-retest parameter called agreement (or measurement error), and a relative test-retest parameter called reliability (Stratford & Goldsmith, 1997; Bruton, Conway & Holgate, 2000; Mokkink et al., 2010). Agreement can be further specified into within-session agreement and between-session agreement. Each PPT is determined as the average of multiple repeated ratings (typically two to four measurements) within one session. Within-session agreement indicates how equal these ratings within the measurement session are, reflecting the precision of the measurement. Between-session agreement represents how close each averaged score for repeated measurements is (*i.e.*, the measurement error), which is important for evaluation purposes. Reliability, on the other hand, accounts for between-subject variability within the tested sample. It shows how well the test can distinguish people from each other, which is essential for diagnostic purposes (de Vet et al., 2006).

Generally, PPTs measured with a digital algometer show good to excellent intra-rater agreement and reliability for most body locations when an experienced rater performs the measurements. For example, for agreement, a standard error of measurement (SEM) of 63 to 88 kPa on larger muscles associated with a higher PPTs, such as the tibialis anterior muscle, and 41 to 53 kPa on smaller muscles with lower PPTs, such as the paraspinal muscles in the neck region, and good coefficients of variations (CVs) (*e.g.*, 8% to 19% for tibialis anterior muscle, and 14% to 18% for the neck region) have been reported. Furthermore, the reported reliability is high (*e.g.*, intraclass correlation coefficients (ICCs) of 0.79 to 0.94 for the tibialis anterior muscle (Sterling et al., 2002; Jørgensen et al., 2014;

*Srimurugan Pratheep, Madeleine & Arendt-Nielsen, 2018*), and 0.82 to 0.91 for the neck region (*Sterling et al., 2002; Ylinen et al., 2007; Knapstad et al., 2018*)).

Agreement and reliability are generally reported as moderate to good in both larger and smaller muscles. However, the reported parameters vary between studies. This variability can partly be attributed to rater characteristics as PPT measurements seem to be operator dependent. Determining PPTs requires practical and psychomotor skills of the individual rater. For example, variability in application rate of the pressure increase negatively impacts on reliability and agreement (*Linde et al., 2017*). Acquiring competence in such skills can be achieved through guided practice, repetition and reinforcement, going through the cognitive, associative and autonomous phases of motor learning (*Schmidt & Lee, 2005; Tynjälä & Gijbels, 2012; Oermann, Muckler & Morgan, 2016; Reilly, Beran-Shepler & Paschal, 2020*). Novice raters may still be developing their psychomotor skills, so it could be assumed that they have not yet achieved comparable skills to experienced raters (*Sattelmayer et al., 2016*). In studies examining reliability of PPTs, only one study included novice raters (*Walton et al., 2011*). This study showed lower reliability values (ICC 0.79) compared to studies with experienced raters (ICC 0.92) (*Waller et al., 2016*). As individual psychomotor skills differ between people (*Schmidt & Lee, 2005; van Duijn, Swanick & Donald, 2014*), this may lead to different reproducibility parameters per individual. As such, it remains unclear whether novice raters are able to determine PPT measurements with an acceptable reproducibility, following limited hours of training. Insight into reproducibility is important as QST measures are often performed by novice raters in student projects and higher degree research projects (*Geber et al., 2009*). Simply extrapolating reproducibility parameters from experienced raters is debatable.

To summarise, research about agreement and reliability parameters in novice raters when performing PPTs is scarce despite the fact that novice raters often perform these measurements. Therefore, this study aimed to determine intra-rater reproducibility parameters (within-session and between-session agreement and reliability) of PPTs measured by novice (*i.e.*, inexperienced) raters. Additionally, this study explored the extent of individual differences in intra-rater reproducibility parameters.

## METHODS

### Design

Using a cross-sectional, observational test-retest study design, ten novice raters determined PPTs using two types of digital algometers to assess intra-rater reproducibility parameters. For agreement, the SEM was used as absolute value for the measurement error and the CV was used as relative value for the measurement error compared to the mean PPT of the muscle (*Mokkink et al., 2010*). For reliability, the ICC<sub>1,1</sub> was used as ratio between the measurement error (*Mokkink et al., 2010*). All intra-rater reproducibility parameters were measured per algometer.

Data were acquired between February and June 2019. The study was approved by the Medical Ethical Committee of the University Medical Center in Groningen, The Netherlands (METc 2016.613; M17.207169). All participants and raters signed an informed consent prior to the measurements. Reported reproducibility parameters are

based upon the COSMIN group recommendations for studies on psychometrics (*Gagnier et al., 2021; Mokkink et al., 2010; de Vet et al., 2016*). The Guideline for Reporting Reliability and Agreement Studies (GRRAS) was used (*Kottner et al., 2011; Gerke et al., 2018*).

### **Raters**

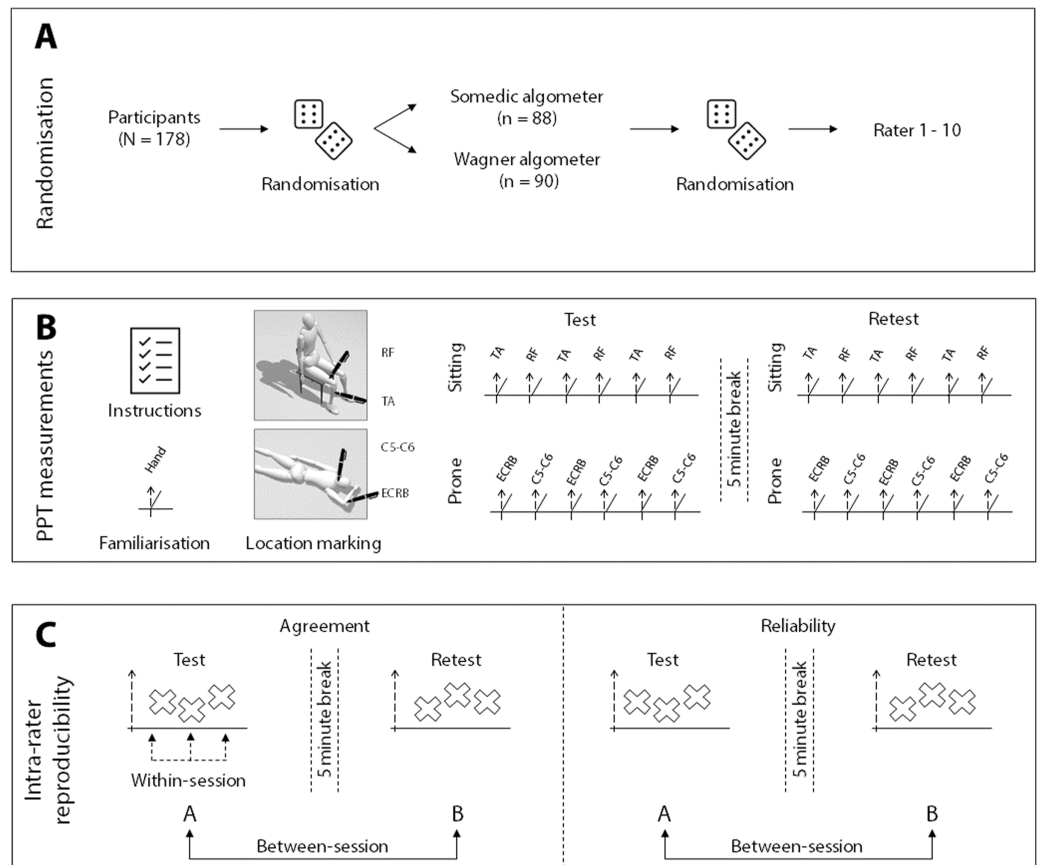
Novice raters were students in the last year of their undergraduate physiotherapy program. Ten raters (six females, four males) were recruited from the Hanze University of Applied Sciences in Groningen, The Netherlands. Physiotherapy students without experience in determining PPTs were eligible to participate. Exclusion criteria were being physically unable to perform PPTs on the selected muscles (in terms of positioning and strength) or not being able to communicate in the Dutch language. Their mean (SD) age, weight and height were 23.0 (3.0) years, 75 (11) kg and 179 (6) cm. For this project, all raters attended two four-hour training sessions to perform PPT measurements with a Somedic and a Wagner digital algometer. The training included an explanation of the study procedures, standardised participant instructions, palpating and marking of the body locations (approximately 2 h) and the use of the two digital algometers (for approximately 3 h per algometer). Practising assessment of PPTs was performed on healthy participants who did not participate in the main study.

### **Participants**

Pain-free, healthy participants were recruited from the staff and student populations of the Hanze University of Applied Sciences, Academy of Health Care, in Groningen, The Netherlands, and the local community. The selection of healthy participants was based on the aim to determine rater-dependence in PPT measurements and as such to eliminate potential confounding variables of dysfunction or disease. Convenience sampling took place *via* an announcement on the university's intranet and an email was sent to all staff and students. All participants underwent a short screening to determine eligibility (*Gierthmühlen et al., 2015*). Inclusion criteria were being pain-free, between 18 and 65 years old and an adequate understanding of the Dutch language. Participants were not permitted to use alcohol, caffeine or nicotine-containing products 2 h before testing, or using analgesics 24 h prior to testing (*Girdler et al., 2005; Baratloo et al., 2016; Bagot et al., 2017*). Exclusion criteria were the presence of acute or chronic pain, neurological, orthopedic or cognitive disorders, pregnancy or if participants had undergone surgery to the legs, arms or neck.

### **Sample size**

Sample size estimation was performed for both agreement and reliability parameters for the group of raters. For agreement, using an upper bound of the 95% CI of +15% of the SEM, a minimum of 81 participants was required (per algometer) (*Stratford & Goldsmith, 1997*). For reliability, based on  $\alpha = 0.05$ ,  $\beta = 0.8$ , an expected ICC<sub>1,1</sub> of >0.7 with the lower bound of the 95% CI of 0.5, a minimum of 63 participants was required to calculate reliability per algometer (*Walter, Eliasziw & Donner, 1998*).



**Figure 1** Visual representation of the methods. (A) The randomisation procedure for algometer type and rater allocation; (B) shows the measurement procedure; (C) illustrates which measurements are used for the within-session and the between-session calculations. The mean of the three measurements were used for the between-session calculations. RF, rectus femoris muscle; TA, tibialis anterior muscle; C5-C6, paraspinal muscles at C5-C6; ECRB, extensor carpi radialis brevis muscle.

Full-size DOI: 10.7717/peerj.14565/fig-1

## PPT measurements

After collecting the demographic, anthropometric, lifestyle and psychosocial data, participants were randomly allocated to a rater and went into one of four rooms where they received standardised instructions. Then, every participant was familiarised with the measurement on the dorsal aspect of the hand by the allocated rater. [Figure 1](#) shows the procedure and methods used in the study. Four rooms were used to simultaneously measure all participants. The rooms were comparable regarding size, temperature and noise levels.

PPTs were measured using either a Somedic digital algometer (Type II; Somedic AB, Stockholm, Sweden) or Wagner digital algometer (Force Ten FDX 25; Wagner, Greenwich, CT, USA). Both digital algometers were used with a 1 cm<sup>2</sup> rubber tip with an application rate of ~50 kPa/s. The Somedic algometer provides visual feedback regarding the application rate and has a hand-held switch for the participant to press when the feeling of pressure changes into painful pressure. The Wagner device does not provide feedback

regarding the application rate. Consequently, the raters had to estimate the rate themselves. Because visual feedback about the slope might assist raters in obtaining more stable PPTs, the two different algometers were used to assess whether the intra-rater reproducibility parameters are influenced by the algometer properties. Furthermore, the Wagner algometer has no patient switch, so participants verbally indicated when their pain threshold was reached.

After marking the locations with a skin pencil, raters measured the PPTs three times in each location, with an inter-stimulus interval of 20 s to avoid sensitisation of pain (Reezigt *et al.*, 2021; Nussbaum & Downes, 1998). The measurements were taken on the dominant side, in a fixed order in two circuits. In the first circuit, with the participant in sitting, PPTs were determined on the tibialis anterior muscle (~5 cm distal to the tibial tuberosity) and rectus femoris muscle (~15 cm proximal to the patella base) on the dominant side, alternating between the two locations (Fig. 1B). In the second circuit, with the participant in prone and the dominant arm above the head, PPTs were determined between extensor carpi radialis brevis muscle (~5 cm distal to the lateral epicondyle of the humerus) and the paraspinal muscles at the C5-C6 level of the neck, again in an alternating manner (Fig. 1B). The test locations were selected based on the different levels of complexity of the measurements and height of the PPT. Furthermore, these muscles are often selected in PPT studies and their reproducibility parameters are available for experienced raters (Sterling *et al.*, 2002; Ylinen *et al.*, 2007; Jones, Kilgour & Comtois, 2007; Jørgensen *et al.*, 2014; Waller *et al.*, 2016; Jakorinne, Haanpää & Arokoski, 2018; Srimurugan Pratheep, Madeleine & Arendt-Nielsen, 2018; Knapstad *et al.*, 2018; Middlebrook *et al.*, 2020).

After removal of the skin markings and a 5-min wash-out period, the rater re-marked the locations and retested the PPTs.

## Randomisation and Blinding

Participants were randomly allocated to either being measured (1) with the Somedic or Wagner algometer, and (2) by one of the ten raters. Computer-generated randomisation and opaque envelopes were used. To blind the raters for the measurement results, they handed the algometer to a research assistant after each measurement. The research assistant read the value on the display, recorded the result, reset the algometer and handed the device back to the rater. Participants were also blinded for their test values by shielding the display of the algometer during the measurements.

## Statistical analysis

### *Descriptive statistics*

Descriptive analyses were used to report participant characteristics containing demographic, anthropometric, lifestyle and psychosocial data. The normality assumption was visually inspected and tested using the Shapiro–Wilk test. Normally distributed data were described as mean (SD). All statistical analyses were performed in SPSS version 28 (IBM, Armonk, NY, USA). Participant's characteristics measured in the Somedic algometer group and Wagner algometer group were analysed to investigate comparability of the two samples and to be able to indirectly compare the reproducibility parameters

between the two algometers. Continuous data were tested using either an independent t-test (in normal distribution) or non-parametric Mann–Whitney U Test. The Chi-Square Test was used for categorical data. All PPT data of the Somedic algometer were expressed in kPa. All PPT data of the Wagner algometer were converted from Newton/cm<sup>2</sup> to kPa.

### **Reproducibility parameters**

All reproducibility parameters were intra-rater parameters and calculated per algometer. Group reproducibility parameters reflect novice raters as a group, whereas individual reproducibility parameters reflect each novice rater individually. Intra-rater absolute (*i.e.*, SEM) and relative (*i.e.*, CV) parameters were calculated for both within-session and between-session agreement. The ICC<sub>1,1</sub> was calculated for reliability. For all analyses, per algometer and per location, outlier or influential case detection and removal were based on average PPT and difference scores greater than three SDs (reducing an overrepresentation of extreme cases in a normal distribution) (Parrinello *et al.*, 2016).

### **Within-session agreement**

Within-session agreement reflects the precision within a measurement session; *i.e.*, how close the three ratings of the first test session are (Fig. 1C). First, for both the group and each individual rater, the within-session agreement was calculated as an absolute value in the form of the SEM. To eliminate the bias of subject variability, the calculation was based on the error mean square term (EMS) of the Analysis of Variance (ANOVA) used for the ICC<sub>1,1</sub> calculations, according to the formula:  $SEM = \sqrt{EMS}$  (Weir, 2005). Confidence intervals were calculated based on the sample error variance;  $\left[ \sqrt{\frac{SSE}{\chi^2_{\alpha, dfe}}}, \sqrt{\frac{SSE}{\chi^2_{1-\alpha, dfe}}} \right]$ , where SSE is the sum of squares error (from the ANOVA),  $\chi^2_{\alpha, dfe}$  is the chi-square value for significance level  $\alpha$  (0.05) and the corresponding degrees of freedom (df) of the SSE term (Stratford & Goldsmith, 1997; Armitage & Berry, 2001).

Second, as each sample of ratings (per location, per rater) may differ from the average PPT (*e.g.*, PPT ratings on the leg are higher than those on the arm), the CV, also called the relative measurement error (*i.e.*, SEM%), is used as an accompanying agreement parameter to increase comparability between locations and raters. The CV is calculated as the SEM divided by the sample average threshold and is reported as a percentage:

$$\text{Within-session CV} = \frac{SEM_{\text{Within-session}}}{\text{AVG}(3 \text{ ratings})} * 100, \text{ where}$$

$$SEM_{\text{Within-session}} = \sqrt{EMS_{\text{Within-session}}}$$

A lower SEM and lower CV reflect a better within-session agreement and as such a higher precision (Atkinson & Nevill, 1998).

### **Between-session agreement**

Between-session agreement reflects how close the differences were between the test and retest average ratings (Fig. 1C). For the test and retest sessions, the average of the three measurements was used to calculate the PPT on each location. Subsequently, the

between-session agreement could be calculated as the absolute SEM and relative CV. Additionally, the minimal detectable change of 90% ( $MDC_{90}$ ) based on the SEM (Weir, 2005) was calculated for clinical purposes (Donoghue et al., 2009) by:

$$MDC_{90} = SEM * \sqrt{2} * 1.65$$

### **Dose-dependent differences**

When dose-dependent differences are present, a relative parameter (e.g., CV) reflects the measurement error better than an absolute parameter (e.g., SEM) (Atkinson & Nevill, 2000). As such, the data were analysed to see whether the differences between the test and retest increased proportionally to the average threshold increment (i.e., heteroscedasticity of the data; e.g., when more force is needed to determine the PPT, the difference between the test and retest may be larger compared to low force PPT thresholds due to less accuracy when using more strength). First, Bland–Altman plots were created to visualise and analyse the potential dose-dependent differences (Bland & Altman, 1999; Carkeet, 2015; Gerke, 2020).

Second, linear regression lines, based on the absolute value of the differences (i.e., the distance from zero) between the test and retest PPT values, were added to the Bland–Altman plots. The regression analyses were used to estimate the dose-dependent difference more accurately than visual inspection by using the regression coefficient and adjusted  $R^2$  (variance explained as an indicator of the strength of the relationship between increasing thresholds and increasing differences).

In this study, presence of dose-dependent differences was defined as: (a) the slope of the regression line was  $>0.1$  and (b) the adjusted explained variance ( $R^2$ ) by the regression model was  $>0.1$ . These regression analyses were partly based on suggestions of Atkinson & Nevill (2000) and the method of Ho (2018), where we adjusted dose-dependent systematic bias to dose-dependent differences (Atkinson & Nevill, 2000; Ho, 2018). For these linear regression analyses, per algometer and per location, outlier or influential case detection and removal were based on standardised residuals from the regression greater than three (Parrinello et al., 2016).

### **Reliability**

Reliability reflects the measurement error relative to the subject variability. Reliability estimates were calculated using ANOVA-based ICCs (Koo & Li, 2016). The intra-rater reliability represents novice raters in general, and as such each participant was rated by a different rater who was randomly chosen from the set of ten raters. Therefore, ICC estimates and their 95% CI were calculated based on a single rater, absolute agreement, one-way random-effects model ( $ICC_{1,1}$ ) (Koo & Li, 2016).

Interpretation of ICCs is based on the criteria of Koo & Li (2016): values less than 0.5 indicate poor reliability, 0.50 till 0.75 indicate moderate reliability, 0.75 till 0.90 indicate good reliability and values greater than 0.90 indicate excellent reliability.



**Table 1** Characteristics of the participants.

	Somedic algometer group	Wagner algometer group	P-value
<b>Demographics</b>	<i>n</i> = 88	<i>n</i> = 90	
Sex, male (%)	44 (50%)	49 (54%)	<i>p</i> = 0.65 <sup>b</sup>
Age, years	22 (20–25) <sup>a</sup>	22 (20–23) <sup>a</sup>	<i>p</i> = 0.50 <sup>c</sup>
<b>Anthropometrics</b>			
Height, cm	179 (11)	178 (9)	<i>p</i> = 0.35
Weight, kg	74 (12)	75 (13)	<i>p</i> = 0.58
BMI, kg/m <sup>2</sup>	22.9 (2.7)	23.6 (3.2)	<i>p</i> = 0.12
<b>Lifestyle</b>			
Sports, hours/week	5 (2.5–6.5) <sup>a</sup>	5 (3.0–7.5) <sup>a</sup>	<i>p</i> = 0.69 <sup>c</sup>
Alcohol use, AUDIT (total score)	8.0 (4.5)	8.2 (4.4)	<i>p</i> = 0.93
Smoking, <i>n</i> (%)	6 (6.8%)	16 (17.8%)	<i>p</i> = 0.04 <sup>b*</sup>
Sleep, PSQI (total score)	4.3 (1.9)	4.7 (2.4)	<i>p</i> = 0.28
<b>Psychosocial</b>			
Anxiety, GAD-7 (total score)	2 (1–3) <sup>a</sup>	2 (0–3) <sup>a</sup>	<i>p</i> = 0.68 <sup>c</sup>
Depression, CES-D (total score)	3 (1–6) <sup>a</sup>	4 (2–6) <sup>a</sup>	<i>p</i> = 0.38 <sup>c</sup>

**Note:**

Data are presented as mean (SD) unless otherwise specified, <sup>a</sup>Median (IQR), <sup>b</sup>Chi Square Test, <sup>c</sup>Mann–Whitney U Test. \*Significant difference between both groups:  $\chi^2_{(1, n = 178)} = 4.93, p = 0.04$ . BMI, Body mass index; AUDIT, Alcohol Use Disorders Identification Test; PSQI, Pittsburgh Sleep Quality Index; GAD-7, Generalised anxiety disorder 7 items; CES-D, Center for Epidemiologic Studies Depression Scale.

## RESULTS

Ten raters measured a total of 178 participants; 88 participants were measured with the Somedic algometer and 90 with the Wagner algometer. Both groups were similar in participant characteristics and baseline PPTs, except for smoking status, which was significantly higher in the Wagner algometer group (Table 1). One rater's data (rater 10) using the Wagner algometer, were excluded as only two participants were measured with this device due to illness of this rater. As a result, the exploration of individual data using the Wagner algometer contains nine raters.

### Group reproducibility

All within-session and between-session agreement and reliability values for the Somedic and Wagner algometer are presented in Table 2. Within-session agreement, as SEM, ranged from 42 to 72 kPa, and as CV from 10.8% to 14.5% for the Somedic algometer. For the Wagner algometer, it ranged from 42 to 74 kPa (SEM), and from 11.5% to 14.8% (CV). Measurements at the rectus femoris muscle showed the best within-session agreement, whereas the measurements at the extensor carpi radialis brevis muscle showed the lowest within-session agreement using either algometer (Fig. 2).

Between-session agreement ranged from 36 to 71 kPa (SEM) and from 10.4% to 16.1% (CV) for the Somedic algometer, and from 47 to 76 kPa (SEM) and from 9.4% to 14.6% (CV) for the Wagner algometer. The best between-session agreement was found at the

**Table 2** Reproducibility parameters on group level.

Somedic algometer	Within-session agreement			Between-session agreement			Between-session reliability
	Mean PPT (kPa)	Standard error of measurement (kPa, 95% CI)	Coefficient of variation (95% CI)	Standard error of measurement (kPa, 95% CI)	Minimal detectable change 90% (kPa, 95% CI)	Coefficient of variation (95% CI)	Intraclass correlation coefficient $_{1,1}$ (95% CI)
Tibialis anterior	578	61 [55–70]	11.3% [10.0–12.9%]	71 [64–82]	166 [149–191]	12.9% [11.5–4.8%]	0.91 [0.87–0.94]
Rectus femoris	600	61 [55–70]	10.8% [9.6–12.4%]	60 [53–68]	140 [124–159]	10.4% [9.2–11.9%]	0.93 [0.89–0.95]
Extensor carpi radialis brevis	402	56 [50–64]	14.5% [12.9–16.7%]	63 [56–72]	147 [131–168]	16.1% [14.3–18.4%]	0.89 [0.84–0.93]
Paraspinal muscles C5-C6	323	42 [37–48]	13.7% [12.1–15.6%]	36 [32–41]	84 [75–96]	11.3% [10.1–13.0%]	0.93 [0.89–0.95]
Wagner algometer							
Tibialis anterior	579	74 [66–85]	13.0% [11.5–14.9%]	76 [67–87]	177 [156–203]	13.1% [11.6–15.0%]	0.88 [0.83–0.92]
Rectus femoris	655	74 [66–85]	11.5% [10.2–13.2%]	62 [55–71]	145 [128–166]	9.4% [8.4–10.8%]	0.95 [0.92–0.97]
Extensor carpi radialis brevis	364	52 [46–60]	14.8% [13.2–17.0%]	47 [42–54]	110 [98–126]	12.9% [11.5–14.8%]	0.90 [0.85–0.93]
Paraspinal muscles C5-C6	325	42 [38–49]	13.4% [11.9–15.4%]	47 [42–54]	110 [98–126]	14.6% [13.0–16.7%]	0.87 [0.81–0.92]

**Note:**  
kPa, kilo Pascal.

rectus femoris muscle using either algometer (Fig. 2). Measurements at the extensor carpi radialis brevis muscle showed the lowest agreement for the Somedic algometer (CV: 16.1% (95% CI [14.3–18.4%]) on an average PPT of 393 kPa), whereas measurements at the paraspinal muscles at C5-C6 were the lowest for the Wagner algometer (CV: 14.6% (95% CI [13.0–16.7%]) on an average PPT of 325 kPa). The values of the  $MDC_{90}$  are presented in Table 2.

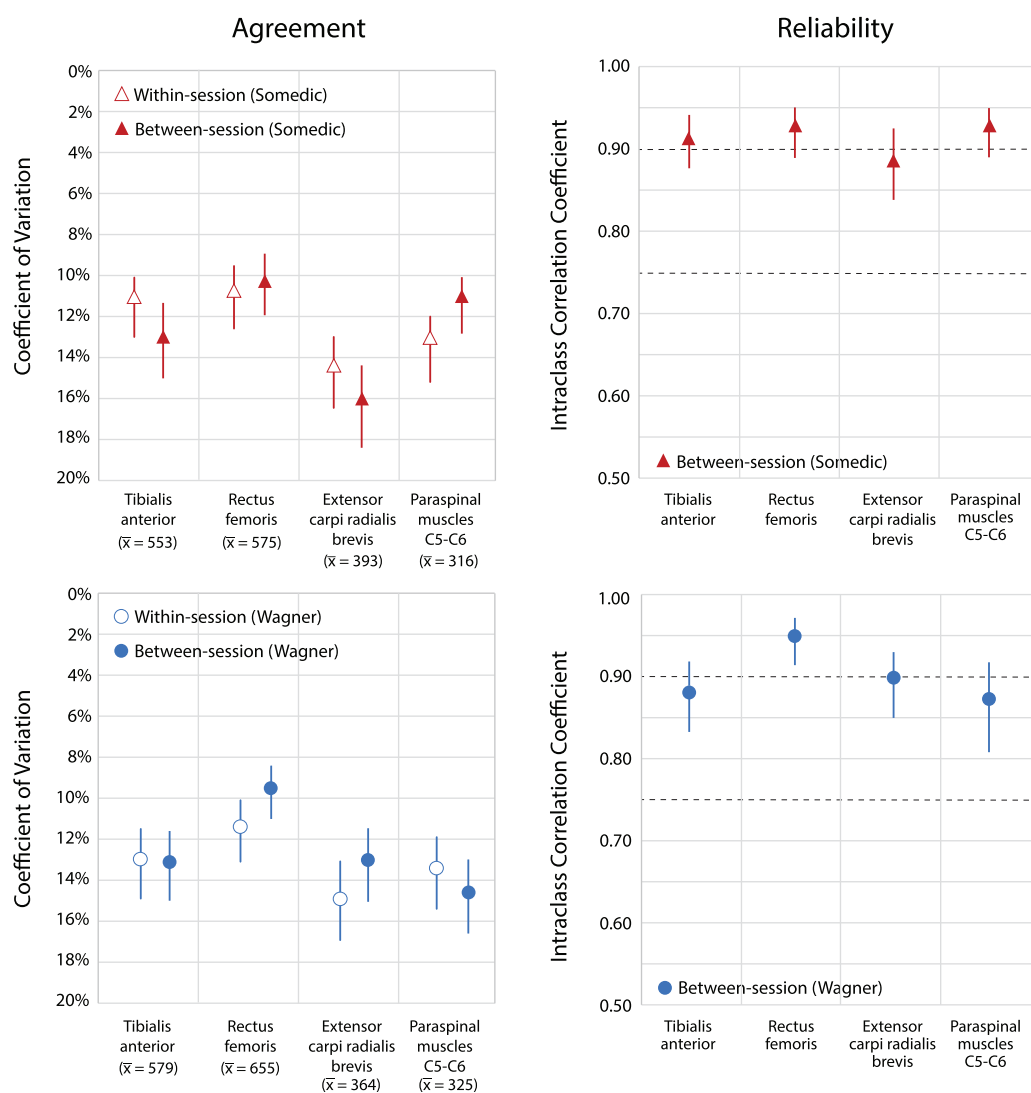
Dose-dependent differences (higher average thresholds associated with larger differences) were found at the tibialis anterior muscle (slope of 0.12x,  $R^2 = 0.301$ ) and the extensor carpi radialis brevis muscle (slope of 0.15x,  $R^2 = 0.253$ ) using the Somedic algometer. Using the Wagner algometer, dose-dependent differences were present at the paraspinal muscles of C5-C6 (slope of 0.17x,  $R^2 = 0.225$ ). Measurements at the rectus femoris muscle showed the lowest dose-dependent differences (Figs. 3 and 4).

The reliability was good to excellent at all locations with both algometers. The  $ICC_{s_{1,1}}$  for the Somedic algometer ranged from 0.89 to 0.93, and for the Wagner algometer from 0.87 to 0.95 (Fig. 2).

### Exploration of individuals' reproducibility

Individual within-session agreement expressed as SEM ranged from 24 to 120 kPa, depending on the rater, location and algometer. Consequently, the maximal individual offset compared to the group findings was  $-26$  and  $+46$  kPa (both on the rectus femoris muscle). The CV ranged from 5.1% to 21.3%, with maximal differences compared to the group findings of  $-6.0$  to  $+9.3$  percentage points (both on the rectus femoris muscle).

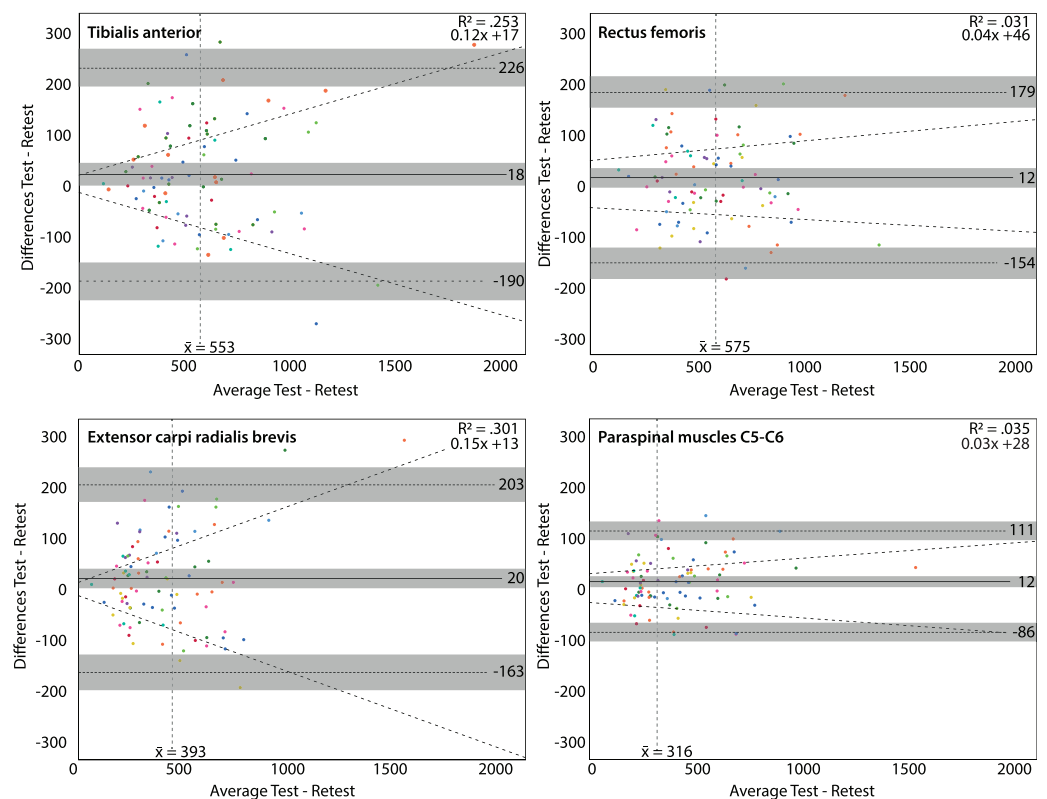
Individual between-session agreement expressed as SEM, ranged from 15 to 115 kPa, depending on the rater, location and algometer. Compared to the group findings, the



**Figure 2** Within-session and between-session agreement and reliability for both types of digital algometers. Upper parts show agreement and reliability of the Somic algometer; lower parts show agreement and reliability of the Wagner algometer. Within-session and between-session agreement expressed as coefficient of variation (CV), including 95% Confidence Intervals. A lower percentage in the coefficient of variation indicates a better agreement. Average PPT's are given per location in kilo Pascal. Reliability is expressed as intraclass correlations coefficients,  $ICC_{1,1}$ , including 95% Confidence Intervals. Dotted horizontal lines represent classification of *Koo & Li (2016)*: values  $\geq 0.75$  indicate good reliability and  $\geq 0.9$  excellent reliability.

Full-size DOI: [10.7717/peerj.14565/fig-2](https://doi.org/10.7717/peerj.14565/fig-2)

maximal offset was  $-55$  and  $+39$  kPa (both on the tibialis anterior muscle). The CV ranged from 3.6% to 26.5%, and differences ranged from  $-9.5$  to  $+6.6$  percentage points compared to the group findings. For example, on the tibialis anterior muscle using the Wagner algometer, rater 3 had a CV of 3.6% (95% CI [2.4–7.5%], on an average PPT of 595 kPa) and the CV of rater 6 was 17.8% (95% CI [13.3–27.6%], on an average PPT of 645 kPa). Overall, most raters showed acceptable agreement scores, only rater 9 scored a CV of 26.5% (95% CI [18.3–50.8%], on an average PPT of 304 kPa) on the extensor carpi radialis brevis muscle. In general, the extensor carpi radialis brevis muscle showed the most



**Figure 3** Bland & Altman plots for the Somedic algometer. Bland & Altman plots with the average PPT in kilo Pascal of the test and retest on the x-axis and the differences between the test and retest on the y-axis. Grey areas are the 95% confidence intervals of the upper limit, lower limit and systematic bias. The vertical dashed line is the mean threshold for that location. The diagonal dashed line is the linear regression line indicating potential dose-dependent differences. Dose-dependent differences were observed for the tibialis anterior muscle (upper left panel) and extensor carpi radialis brevis muscle (lower left panel). [Full-size !\[\]\(5f471a71b78d7676bc356df190b88ab4\_img.jpg\) DOI: 10.7717/peerj.14565/fig-3](https://doi.org/10.7717/peerj.14565/fig-3)

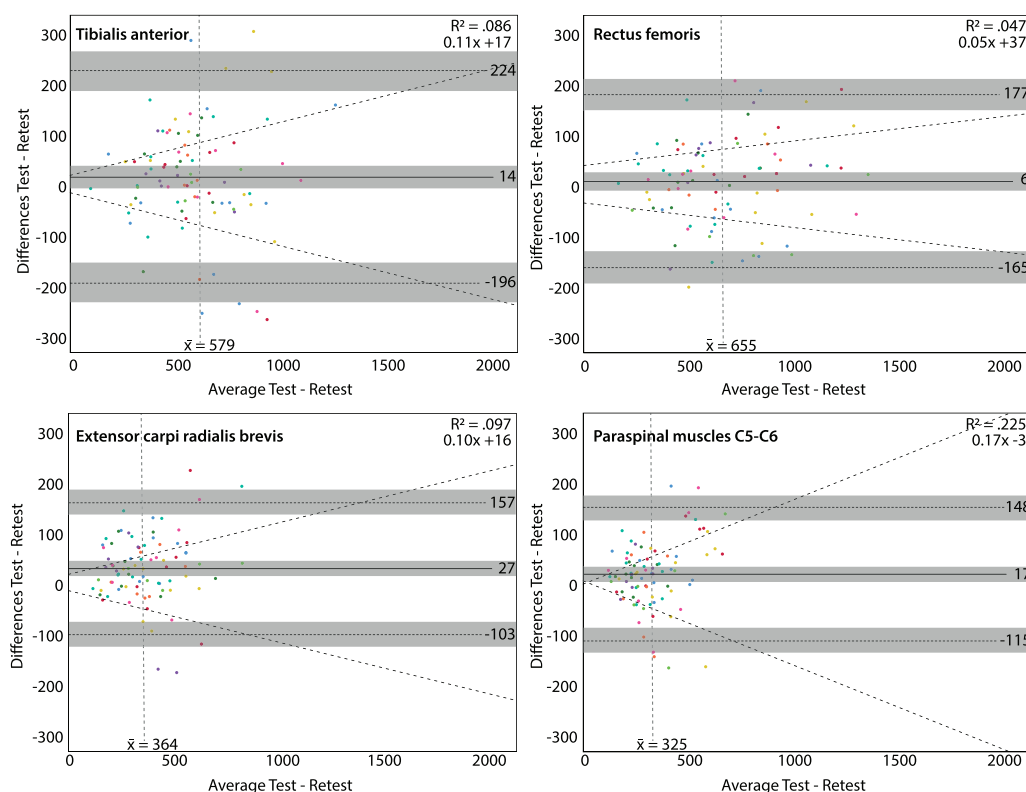
unfavourable agreement scores, probably due to the smaller size increasing the complexity of the measurement (*e.g.*, more difficult to localise and more difficult to keep the algometer on the muscle during the measurement). All individual rater reproducibility parameters are shown in [Fig. 5](#) and are presented in [Appendices A](#) and [B](#).

Between-session agreement and reliability were comparable for males and females except for the tibialis anterior muscle using the Wagner algometer ([Appendix C](#)).

## DISCUSSION

This study showed good intra-rater reproducibility parameters for novice raters after 8 h of training, including 6 h of performing PPTs, with a Somedic and Wagner digital algometer (*i.e.*, 3 h per algometer). Evaluation of individual intra-rater, between- and within-session agreement showed small differences per rater, but within an acceptable range.

The reproducibility parameters found for novice raters were comparable or even slightly better than those reported for experienced raters ([Sterling et al., 2002](#); [Ylinen et al., 2007](#); [Jones, Kilgour & Comtois, 2007](#); [Jørgensen et al., 2014](#); [Waller et al., 2016](#); [Jakorinne, Haanpää & Arokoski, 2018](#); [Srimurugan Pratheep, Madeleine & Arendt-Nielsen, 2018](#);

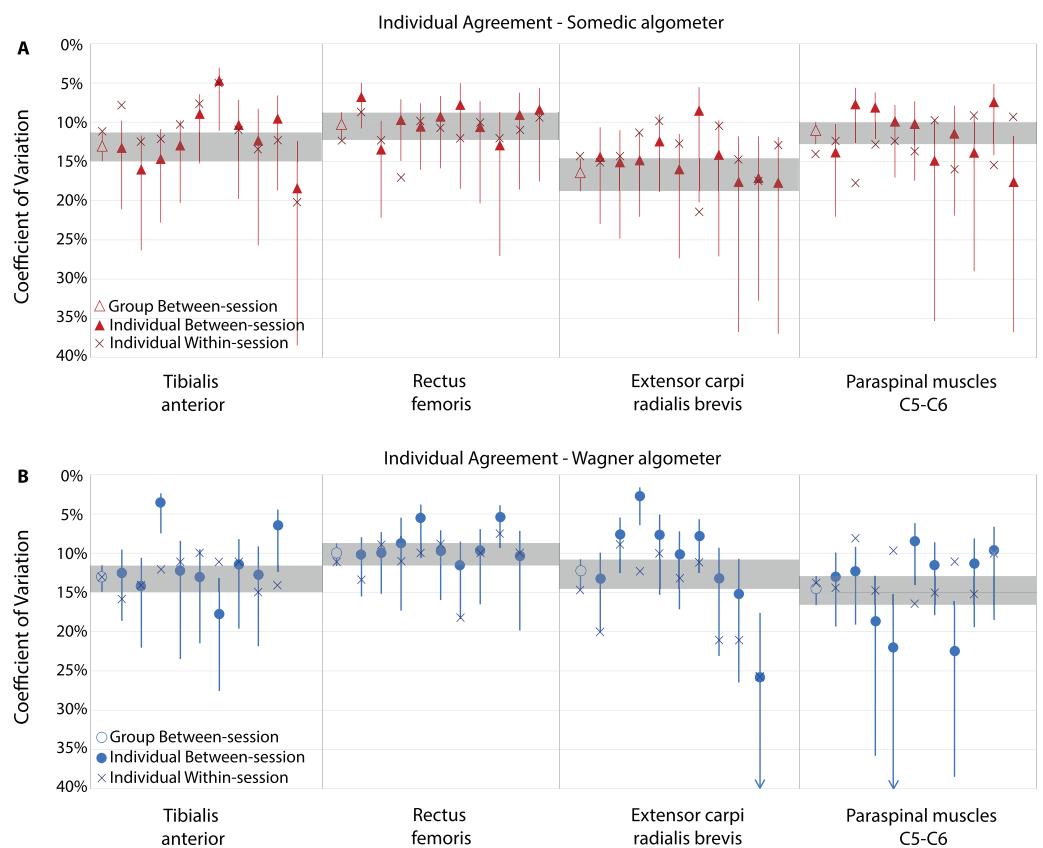


**Figure 4** Bland & Altman plots for the Wagner algometer. Bland & Altman plots with the average PPT in kilo Pascal of the test and retest on the x-axis and the differences between the test and retest on the y-axis. Grey areas are the 95% confidence intervals of the upper limit, lower limit and systematic bias. The vertical dashed line is the mean threshold for that location. The diagonal dashed line is the linear regression line indicating potential dose-dependent differences. Dose-dependent differences were observed for the paraspinal muscles (C5-C6) (lower right panel).

Full-size DOI: [10.7717/peerj.14565/fig-4](https://doi.org/10.7717/peerj.14565/fig-4)

*Knapstad et al., 2018; Middlebrook et al., 2020*). Between-session agreement was comparable, depending on the location. In most locations, this study revealed equal or better CVs (e.g., ~22% at the extensor carpi radialis brevis muscle compared to 16% in our study), except at the tibialis anterior muscle. A previous study found a slightly better CV at the tibialis anterior muscle in experienced raters (8% vs. 10–14%) (*Sterling et al., 2002*) compared to our study. In contrast, these authors found slightly lower ICC values for reliability (ICC 0.65 to 0.94) compared to our study (ICC 0.87 to 0.95). The mean PPTs we obtained at the different locations were comparable with previously reported values (*Sterling et al., 2002; Ylinen et al., 2007; Jones, Kilgour & Comtois, 2007; Jørgensen et al., 2014; Waller et al., 2016; Jakorinne, Haanpää & Arokoski, 2018; Srimurugan Pratheep, Madeleine & Arendt-Nielsen, 2018; Knapstad et al., 2018; Middlebrook et al., 2020*).

One other study included novice undergraduate physiotherapy students, using a Wagner algometer at the tibialis anterior muscle and paraspinal muscles in a comparable population of healthy, young participants (mean of 25.4 years) (*Walton et al., 2011*). They found slightly lower reliability (ICC 0.79 vs. 0.88 in our study) and within-session agreement values (CV 19% vs. 13% in our study). The lower reliability and within session



**Figure 5** Individual exploration of reproducibility parameters of novice raters. Agreement expressed as coefficient of variation (CV), including 95% Confidence Intervals, of the Somic algometer (A) and Wagner algometer (B). Shaded areas show the overall group between-session agreement confidence interval as reference. Each filled mark represents an individual rater.

Full-size  DOI: [10.7717/peerj.14565/fig-5](https://doi.org/10.7717/peerj.14565/fig-5)

agreement that they found could probably be explained by the lower number of training hours compared to our study (1 vs. 8 h) (Walton et al., 2011). Earlier studies also attributed lower individual reliability scores to systematic errors from novice raters (Chung, Um & Kim, 1992; Goulet, Clark & Flack, 1993). Due to potential different learning curves (not every rater learns equally fast and the number of participants measured to reach acceptable reproducibility may differ), the training could have a variable effect on each rater's performance (van Duijn, Swanick & Donald, 2014). Additionally, it is unknown how much training is needed to perform PPTs with an acceptable reproducibility.

In this study, an arbitrary number of ten raters was chosen, as we hypothesised that individual (psychomotor) skills could influence the reproducibility parameters. Using a group of multiple raters increased the generalisability of the findings and allowed exploration of differences in the reproducibility parameters of individual raters. Even though novice raters show good group reproducibility parameters compared to experienced raters, individual differences in novice were identified. No previous studies assessed individual reproducible parameters in experienced or novice raters. Multiple raters showed a CV and SEM which were over two times smaller than the group's CV and

SEM, as such they have excellent measurement skills techniques compared to their peer raters.

This study has some limitations. Ideally, for exploration of rater differences, all raters should have tested the same group of participants to minimise the subject variability. However, different subsamples were chosen to reduce the number of tests per participant and prevent potentially attention bias, salience and sensitisation effects (*Villemure & Bushnell, 2002; Goffaux et al., 2007; Hall & Rodríguez, 2017; Moore et al., 2019*). Consequently, we could not directly compare reproducibility values between the two different algometers. However, the two samples were comparable based on demographics and PPT values. Another limitation is that not all raters measured an equal number of participants (min = 6; max = 14), which influenced the width of the confidence intervals and the amount of experience gained while acquiring the reproducibility data. Remarkably, raters who measured a higher number of participants showed noticeable higher reproducibility parameters. This could possibly be explained by the differences in gained experience through measuring. The individual parameters should, however, be interpreted with caution due to the low subsample sizes.

We found some individual differences in reproducibility values, which were within acceptable ranges. Consequently, novice raters who attended an 8-h training can participate as raters in research projects and use PPTs in clinical practice adequately. Since individual differences in reproducibility parameters may exist, researchers and clinicians should be cautious when using reproducibility parameters of other raters from former studies. Future studies should focus on quantifying these differences and include methods to explore whether rater experience or psychomotor skills (*e.g.*, strength or dexterity) may explain these differences. Furthermore, the relationship between the duration of training and reproducibility parameters should be explored further to recommend the minimal duration of training needed to perform PPTs adequately.

In conclusion, although slight differences between individual assessors exist, this study revealed that novice raters show good reproducibility parameters in determining PPTs across four body locations following 3 h of practice of PPT measures in addition to surface palpation skills.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This study was conducted with a research grant for teachers of the Dutch Organisation of Scientific Research (NWO) under project number 023.011.069. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:  
Dutch Organisation of Scientific Research (NWO): 023.011.069.

## Competing Interests

The authors declare that they have no competing interests.

## Author Contributions

- Roland R. Reezigt conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Geranda E. C. Slager conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Michel W. Coppieters conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Gwendolyne G. M. Scholten-Peters conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

## Human Ethics

The following information was supplied relating to ethical approvals (*i.e.*, approving body and any reference numbers):

The study was approved by the Medical Ethical Committee of the University Medical Center in Groningen, The Netherlands (METc 2016.613; M17.207169).

## Data Availability

The following information was supplied regarding data availability:

The SPSS data of PPT measurements is available in the [Supplemental Files](#).

## Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.14565#supplemental-information>.

## REFERENCES

- Amiri M, Alavinia M, Singh M, Kumbhare D. 2021.** Pressure pain threshold in patients with chronic pain: a systematic review and meta-analysis. *American Journal of Physical Medicine & Rehabilitation* **100**(7):656–674 DOI [10.1097/PHM.0000000000001603](https://doi.org/10.1097/PHM.0000000000001603).
- Armitage P, Berry G. 2001.** *Statistical methods in medical research*. Fourth Edition. Hoboken: Wiley-Blackwell, 115–117, 221–223.
- Atkinson G, Nevill AM. 1998.** Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Medicine* **26**(4):217–238 DOI [10.2165/00007256-199826040-00002](https://doi.org/10.2165/00007256-199826040-00002).
- Atkinson G, Nevill A. 2000.** Measures of reliability in sports medicine and science. *Sports Medicine* **30**(5):375–381 DOI [10.2165/00007256-200030050-00005](https://doi.org/10.2165/00007256-200030050-00005).
- Bagot KS, Wu R, Cavallo D, Krishnan-Sarin S. 2017.** Assessment of pain in adolescents: influence of gender, smoking status and tobacco abstinence. *Addictive Behaviors* **67**:79–85 DOI [10.1016/j.addbeh.2016.12.010](https://doi.org/10.1016/j.addbeh.2016.12.010).



- Baratloo A, Rouhipour A, Forouzanfar MM, Safari S, Amiri M, Negida A. 2016.** The role of caffeine in pain management: a brief literature review. *Anesthesiology and Pain Medicine* **6(3)**:139 DOI [10.5812/aapm.33193](https://doi.org/10.5812/aapm.33193).
- Bland JM, Altman DG. 1999.** Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8(2)**:135–160 DOI [10.1177/096228029900800204](https://doi.org/10.1177/096228029900800204).
- Bruton A, Conway JH, Holgate ST. 2000.** Reliability: what is it, and how is it measured? *Physiotherapy* **86(2)**:94–99 DOI [10.1016/S0031-9406\(05\)61211-4](https://doi.org/10.1016/S0031-9406(05)61211-4).
- Carkeet A. 2015.** Exact parametric confidence intervals for bland-altman limits of agreement. *Optometry and Vision Science* **92(3)**:e71–e80 DOI [10.1097/OPX.0000000000000513](https://doi.org/10.1097/OPX.0000000000000513).
- Chung S-C, Um B-Y, Kim H-S. 1992.** Evaluation of pressure pain threshold in head and neck muscles by electronic algometer: intrarater and interrater reliability. *CRANIO®* **10(1)**:28–34 DOI [10.1080/08869634.1992.11677888](https://doi.org/10.1080/08869634.1992.11677888).
- de Vet HCW, Terwee CB, Knol DL, Bouter LM. 2006.** When to use agreement versus reliability measures. *Journal of Clinical Epidemiology* **59(10)**:1033–1039 DOI [10.1016/j.jclinepi.2005.10.015](https://doi.org/10.1016/j.jclinepi.2005.10.015).
- de Vet HCW, Terwee CB, Mokkink LB, Knol DL. 2016.** *Measurement in medicine*. Cambridge: Cambridge University Press.
- Donoghue D, Murphy A, Jennings A, McAuliffe A, O’Neil S, Charthaigh EN, Griffin E, Gilhooly L, Lyons M, Galvin R, Gallagher S, Ward S, Mhaille EN, Stokes EK. 2009.** How much change is true change? The minimum detectable change of the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine* **41(5)**:343–346 DOI [10.2340/16501977-0337](https://doi.org/10.2340/16501977-0337).
- Fischer AA. 1987.** Pressure algometry over normal muscles. Standard values, validity and reproducibility of pressure threshold. *Pain* **30(1)**:115–126 DOI [10.1016/0304-3959\(87\)90089-3](https://doi.org/10.1016/0304-3959(87)90089-3).
- Gagnier JJ, Lai J, Mokkink LB, Terwee CB. 2021.** COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Quality of Life Research* **30(8)**:2197–2218 DOI [10.1007/s11136-021-02822-4](https://doi.org/10.1007/s11136-021-02822-4).
- Geber C, Scherens A, Pfau D, Nestler N, Zenz M, Tölle T, Baron R, Treede RD, Maier C. 2009.** Procedure for certification of QST laboratories. *Schmerz (Berlin, Germany)* **23(1)**:65–69 DOI [10.1007/s00482-008-0771-4](https://doi.org/10.1007/s00482-008-0771-4).
- Gerke O. 2020.** Reporting standards for a bland-altman agreement analysis: a review of methodological reviews. *Diagnostics* **10(5)**:1–17 DOI [10.3390/diagnostics10050334](https://doi.org/10.3390/diagnostics10050334).
- Gerke O, Möller S, Debrabant B, Halekoh U, Odense Agreement Working Group. 2018.** Experience applying the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) indicated five questions should be addressed in the planning phase from a statistical point of view. *Diagnostics (Basel, Switzerland)* **8(4)**:69 DOI [10.3390/diagnostics8040069](https://doi.org/10.3390/diagnostics8040069).
- Gierthmühlen J, Enax-Krumova EK, Attal N, Bouhassira D, Cruccu G, Finnerup NB, Haanpää M, Hansson P, Jensen TS, Freynhagen R, Kennedy JD, Mainka T, Rice ASC, Segerdahl M, Sindrup SH, Serra J, Tölle T, Treede RD, Baron R, Maier C. 2015.** Who is healthy? Aspects to consider when including healthy volunteers in QST-based studies—A consensus statement by the EUROPAIN and NEURO-PAIN consortia. *Pain* **156(11)**:2203–2211 DOI [10.1097/j.pain.0000000000000227](https://doi.org/10.1097/j.pain.0000000000000227).
- Girdler SS, Maixner W, Naftel HA, Stewart PW, Moretz RL, Light KC. 2005.** Cigarette smoking, stress-induced analgesia and pain perception in men and women. *Pain* **114(3)**:372–385 DOI [10.1016/j.pain.2004.12.035](https://doi.org/10.1016/j.pain.2004.12.035).
- Goffaux P, Redmond WJ, Rainville P, Marchand S. 2007.** Descending analgesia—when the spine echoes what the brain expects. *Pain* **130(1)**:137–143 DOI [10.1016/j.pain.2006.11.011](https://doi.org/10.1016/j.pain.2006.11.011).

- Goulet J-P, Clark GT, Flack VF. 1993.** Reproducibility of examiner performance for muscle and joint palpation in the temporomandibular system following training and calibration. *Community Dentistry and Oral Epidemiology* **21**(2):72–77  
DOI [10.1111/j.1600-0528.1993.tb00724.x](https://doi.org/10.1111/j.1600-0528.1993.tb00724.x).
- Hall G, Rodríguez G. 2017.** Habituation and conditioning: salience change in associative learning. *Journal of Experimental Psychology: Animal Learning and Cognition* **43**(1):48–61  
DOI [10.1037/xan0000129](https://doi.org/10.1037/xan0000129).
- Ho KM. 2018.** Using linear regression to assess dose-dependent bias on a Bland-Altman plot. *Journal of Emergency and Critical Care Medicine* **2**:68 DOI [10.21037/jeccm.2018.08.02](https://doi.org/10.21037/jeccm.2018.08.02).
- Jakorinne P, Haanpää M, Arokoski J. 2018.** Reliability of pressure pain, vibration detection, and tactile detection threshold measurements in lower extremities in subjects with knee osteoarthritis and healthy controls. *Scandinavian Journal of Rheumatology* **47**(6):491–500  
DOI [10.1080/03009742.2018.1433233](https://doi.org/10.1080/03009742.2018.1433233).
- Jones DH, Kilgour RD, Comtois AS. 2007.** Test-retest reliability of pressure pain threshold measurements of the upper limb and torso in young healthy women. *The Journal of Pain* **8**(8):650–656 DOI [10.1016/j.jpain.2007.04.003](https://doi.org/10.1016/j.jpain.2007.04.003).
- Jørgensen R, Ris I, Falla D, Juul-Kristensen B. 2014.** Reliability, construct and discriminative validity of clinical testing in subjects with and without chronic neck pain. *BMC Musculoskeletal Disorders* **15**(1):253 DOI [10.1186/1471-2474-15-408](https://doi.org/10.1186/1471-2474-15-408).
- Keele KD. 1954.** Pain-sensitivity tests. *The Lancet* **263**(6813):636–639  
DOI [10.1016/S0140-6736\(54\)92347-8](https://doi.org/10.1016/S0140-6736(54)92347-8).
- Knapstad MK, Nordahl SHG, Naterstad IF, Ask T, Skouen JS, Goplen FK. 2018.** Measuring pressure pain threshold in the cervical region of dizzy patients-The reliability of a pressure algometer. *Physiotherapy Research International: The Journal for Researchers and Clinicians in Physical Therapy* **23**(4):e1736 DOI [10.1002/PRI.1736](https://doi.org/10.1002/PRI.1736).
- Koo TK, Li MY. 2016.** A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* **15**(2):155–163  
DOI [10.1016/j.jcm.2016.02.012](https://doi.org/10.1016/j.jcm.2016.02.012).
- Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL. 2011.** Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies* **48**(6):661–671  
DOI [10.1016/j.ijnurstu.2011.01.016](https://doi.org/10.1016/j.ijnurstu.2011.01.016).
- Linde LD, Kumbhare DA, Joshi M, Srbely JZ. 2017.** The relationship between rate of algometer application and pain pressure threshold in the assessment of myofascial trigger point sensitivity. *Pain Practice* **18**(2):224–229 DOI [10.1111/papr.12597](https://doi.org/10.1111/papr.12597).
- Middlebrook N, Heneghan NR, Evans DW, Rushton A, Falla D. 2020.** Reliability of temporal summation, thermal and pressure pain thresholds in a healthy cohort and musculoskeletal trauma population. *PLOS ONE* **15**(5):e0233521 DOI [10.1371/JOURNAL.PONE.0233521](https://doi.org/10.1371/JOURNAL.PONE.0233521).
- Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HCW. 2010.** The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology* **63**(7):737–745 DOI [10.1016/j.jclinepi.2010.02.006](https://doi.org/10.1016/j.jclinepi.2010.02.006).
- Moore DJ, Meints SM, Lazaridou A, Johnson D, Franceschelli O, Cornelius M, Schreiber K, Edwards RR. 2019.** The effect of induced and chronic pain on attention. *The Journal of Pain* **20**(11):1353–1361 DOI [10.1016/j.jpain.2019.05.004](https://doi.org/10.1016/j.jpain.2019.05.004).
- Mücke M, Cuhls H, Radbruch L, Baron R, Maier C, Tölle T, Treede RD, Rolke R. 2016.** Quantitative sensorische Testung (QST). *Schmerz* **35**(S3):1–8 DOI [10.1007/s00482-015-0093-2](https://doi.org/10.1007/s00482-015-0093-2).

- Nijs J, Lahousse A, Kapreli E, Bilika P, Saraçoğlu İ, Malfliet A, Coppeters I, de Baets L, Leysen L, Roose E, Clark J, Voogt L, Huysmans E. 2021. Nociceptive pain criteria or recognition of central sensitization? Pain phenotyping in the past, present and future. *Journal of Clinical Medicine* 10(15):3203 DOI 10.3390/jcm10153203.
- Nussbaum EL, Downes L. 1998. Reliability of clinical pressure-pain algometric measurements obtained on consecutive days. *Physical Therapy* 78(2):160–169 DOI 10.1093/ptj/78.2.160.
- Oermann MH, Muckler VC, Morgan B. 2016. Framework for teaching psychomotor and procedural skills in nursing. *Journal of Continuing Education in Nursing* 47(6):278–282 DOI 10.3928/00220124-20160518-10.
- Ohrbach R, Gale EN. 1989. Pressure pain thresholds, clinical assessment, and differential diagnosis: reliability and validity in patients with myogenic pain. *Pain* 39(2):157–169 DOI 10.1016/0304-3959(89)90003-1.
- Parrinello CM, Grams ME, Sang Y, Couper D, Wruck LM, Li D, Eckfeldt JH, Selvin E, Coresh J. 2016. Iterative outlier removal: a method for identifying outliers in laboratory recalibration studies. *Clinical Chemistry* 62(7):966–972 DOI 10.1373/clinchem.2016.255216.
- Reezigt RR, Kielstra SC, Coppeters MW, Scholten-Peeters GGM. 2021. No relevant differences in conditioned pain modulation effects between parallel and sequential test design. A cross-sectional observational study. *PeerJ* 9(3):e12330 DOI 10.7717/peerj.12330.
- Reilly M, Beran-Shepler K, Paschal KA. 2020. Pedagogy for effective learning of clinical skills: an integrated laboratory model. *Journal of Physical Therapy Education* 34(3):234–241 DOI 10.1097/JTE.000000000000145.
- Rolke R, Baron R, Maier C, Tölle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Bötefür IC, Braune S, Flor H, Häge V, Klug R, Landwehrmeyer GB, Magerl W, Maihöfner C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. 2006. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. *Pain* 123(3):231–243 DOI 10.1016/j.pain.2006.01.041.
- Sattelmayer M, Elsig S, Hilfiker R, Baer G. 2016. A systematic review and meta-analysis of selected motor learning principles in physiotherapy and medical education. *BMC Medical Education* 16(1):387 DOI 10.1186/s12909-016-0538-z.
- Schmidt RA, Lee TD. 2005. Motor control and learning: a behavioral emphasis. In: *Human Kinetics*. Fourth Edition. Champaign, IL, USA: Human Kinetics.
- Srimurugan Pratheep N, Madeleine P, Arendt-Nielsen L. 2018. Relative and absolute test-retest reliabilities of pressure pain threshold in patients with knee osteoarthritis. *Scandinavian Journal of Pain* 18(2):229–236 DOI 10.1515/sjpain-2018-0017.
- Sterling M, Jull G, Carlsson Y, Crommert L. 2002. Are cervical physical outcome measures influenced by the presence of symptomatology? *Physiotherapy Research International: The Journal for Researchers and Clinicians in Physical Therapy* 7:113–121 DOI 10.1002/(ISSN)1471-2865.
- Stratford PW, Goldsmith CH. 1997. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. *Physical Therapy* 77(7):745–750 DOI 10.1093/ptj/77.7.745.
- Tynjälä P, Gijbels D. 2012. Changing world: changing pedagogy. In: *Transitions and Transformations in Learning and Education*. Dordrecht: Springer Netherlands, 205–222.
- van Duijn AJ, Swanick K, Donald EK. 2014. Student learning of cervical psychomotor skills via online video instruction versus traditional face-to-face instruction. *Journal of Physical Therapy Education* 28(1):94–102 DOI 10.1097/00001416-201410000-00015.

- Villemure C, Bushnell MC. 2002.** Cognitive modulation of pain: how do attention and emotion influence pain processing? *Pain* **95**(3):195–199 DOI [10.1016/S0304-3959\(02\)00007-6](https://doi.org/10.1016/S0304-3959(02)00007-6).
- Waller R, Smith AJ, O’Sullivan PB, Slater H, Sterling M, Alexandra McVeigh J, Straker LM. 2016.** Pressure and cold pain threshold reference values in a large, young adult, pain-free population. *Scandinavian Journal of Pain* **13**(1):114–122 DOI [10.1016/j.sjpain.2016.08.003](https://doi.org/10.1016/j.sjpain.2016.08.003).
- Walter SD, Eliasziw M, Donner A. 1998.** Sample size and optimal designs for reliability studies. *Statistics in Medicine* **17**:101–110 DOI [10.1002/\(ISSN\)1097-0258](https://doi.org/10.1002/(ISSN)1097-0258).
- Walton D, MacDermid J, Nielson W, Teasell R, Chiasson M, Brown L. 2011.** Reliability, standard error, and minimum detectable change of clinical pressure pain threshold testing in people with and without acute neck pain. *Journal of Orthopaedic & Sports Physical Therapy* **41**(9):644–650 DOI [10.2519/jospt.2011.3666](https://doi.org/10.2519/jospt.2011.3666).
- Weir JP. 2005.** Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength and Conditioning Research* **19**(1):231 DOI [10.1519/15184.1](https://doi.org/10.1519/15184.1).
- 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 OHG. 2015.** Recommendations on practice of conditioned pain modulation (CPM) testing. *European Journal of Pain (United Kingdom)* **19**(6):805–806 DOI [10.1002/ejp.605](https://doi.org/10.1002/ejp.605).
- Ylinen J, Nykänen M, Kautiainen H, Häkkinen A. 2007.** Evaluation of repeatability of pressure algometry on the neck muscles for clinical use. *Manual Therapy* **12**(2):192–197 DOI [10.1016/j.math.2006.06.010](https://doi.org/10.1016/j.math.2006.06.010).