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

Roland R Reezigt [1,2] , **Geranda E.C. Slager** [2] , **Michel W. Coppieters** [1,3,4] , **Gwendolyne G.M. Scholten-Peeters** [Corresp. 1]

[1] Department of Human Movement Sciences, Faculty of Behavioural and Movement Sciences, Amsterdam Movement Sciences - Program Musculoskeletal Health, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

[2] Academy of Health, Department of Physiotherapy, Hanze University of Applied Scienses, Groningen, Netherlands

[3] Griffith University, Menzies Health Institute Queensland, Brisbane and Gold Coast, Australia

[4] School of Health Sciences and Social Work, Griffith University, Brisbane and Gold Coast, Australia

Corresponding Author: Gwendolyne G.M. Scholten-Peeters
Email address: g.g.m.scholten-peeters@vu.nl

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 (9 raters; 90 participants). Prior to the experiment, the novice assessors practiced PPTs for 3 hours 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 kPa 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 three hours of algometer practice, there were slight differences between assessors, but reproducibility in determining PPTs was overall good.

2 Novice assessors demonstrate good intra-rater agreement and reliability when determining

3 pressure pain thresholds; a cross-sectional study

6 **Authors**
7 Roland R. Reezigt [1,2]; Geranda E.C. Slager [2]; Michel W. Coppieters [1,3,4]; Gwendolyne G.M.
8 Scholten-Peeters[1]

10 **Affiliations**
11 1: Department of Human Movement Sciences, Faculty of Behavioural and Movement Sciences,
12   Vrije Universiteit Amsterdam, Amsterdam Movement Sciences - Program Musculoskeletal
13   Health, The Netherlands
14 2: Academy of Health, Department of Physiotherapy, Hanze University of Applied Sciences,
15   Groningen, The Netherlands
16 3: Menzies Health Institute Queensland, Griffith University, Brisbane and Gold Coast, Australia
17 4: School of Health Sciences and Social Work, Griffith University, Brisbane and Gold Coast,
18   Australia

20 **ORCID**
21 Reezigt: 0000-0002-9388-0283
22 Slager: 0000-0002-9437-3101
23 Coppieters: 0000-0002-3958-4408
24 Scholten-Peeters: 0000-0002-4409-9554

26 **Corresponding author**
27 G.G.M. Scholten-Peeters
28 Faculty of Behavioural and Movement Sciences
29 Vrije Universiteit Amsterdam
30 Van der Boechorststraat 9
31 1081 BT Amsterdam
32 The Netherlands
33 g.g.m.scholten-peeters@vu.nl

35 **Category**
36 Original Research

38 **Funding sources**
39 This study was conducted with a research grant for teachers of the Dutch Organisation of
40 Scientific Research (NWO) under project number 023.011.069.
41

42 **Conflicts of interest**

43 None

44

45 **Abstract**

46

47 **Background.** Experienced assessors show good intra-rater reproducibility (within-session and

48 between-session agreement and reliability) when using an algometer to determine pressure pain

49 thresholds (PPT). However, it is unknown whether novice assessors perform equally well. This

50 study aimed to determine within and between-session agreement and reliability of PPT

51 measurements performed by novice assessors and explored whether these parameters differed

52 per assessor and algometer type.

53 **Methods.** Ten novice assessors measured PPTs over four test locations (tibialis anterior muscle,

54 rectus femoris muscle, extensor carpi radialis brevis muscle and paraspinal muscles C5-C6) in

55 178 healthy participants, using either a Somedic Type II digital algometer (10 raters; 88

56 participants) or a Wagner Force Ten FDX 25 digital algometer (9 raters; 90 participants). Prior to

57 the experiment, the novice assessors practiced PPTs for 3 hours per algometer. Each assessor

58 measured a different subsample of ~9 participants. For both the individual assessor and for all

59 assessors combined (i.e, the group representing novice assessors), the standard error of

60 measurement (SEM) and coefficient of variation (CV) were calculated to reflect within and

61 between-session agreement. Reliability was assessed using intraclass correlation coefficients

62 $(ICC_{1,1})$.

63 **Results.** Within-session agreement expressed as SEM ranged from 42 to 74 kPa, depending on

64 the test location and device. Between-session agreement, expressed as SEM, ranged from 36 to

65 76 kPa and the CV ranged from 9-16% per body location. Individual assessors differed from the

66 mean group results, ranging from -55 kPa to +32 kPa or from -9.5 to +6.6 percentage points.

67   Reliability was good to excellent (ICC$_{1,1}$: 0.87 to 0.95). Results were similar for both types of

68   algometers.

69   **Conclusions.** Following three hours of algometer practice, there were slight differences between

70   assessors, but reproducibility in determining PPTs was overall good.

71

72   **Keywords**

73   Quantitative Sensory Testing, Reproducibility, Reliability, Mechanical hyperalgesia, Pain

74   measurement, Central sensitisation, Pressure pain threshold

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). PPT is defined as the minimal applied pressure that induces a painful sensation (Bruton et al., 2000). 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

98 measurement error), which is important for evaluation purposes. Reliability, on the other hand,

99 accounts for between-subject variability within the tested sample. It shows how well the test can

100 distinguish people from each other, which is essential for diagnostic purposes (de Vet et al.,

101 2006).

102   Generally, PPTs measured with a digital algometer show good to excellent intra-rater

103 agreement and reliability for most body locations when an experienced rater performs the

104 measurements. For example, for agreement, a standard error of measurement (SEM) of 63 to 88

105 kPa on larger muscles associated with a higher PPTs, such as the tibialis anterior muscle, and 41

106 to 53 kPa on smaller muscles wither lower PPTs, such as the paraspinal muscles in the neck

107 region, and good coefficients of variations (CVs) (e.g., 8 to 19% for tibialis anterior muscle, and

108 14 to 18% for the neck region) have been reported. Furthermore, the reported reliability is high

109 (e.g., intraclass correlation coefficients (ICCs) of 0.79 to 0.94 for the tibialis anterior muscle

110 (Sterling et al., 2002; Jørgensen et al., 2014; Srimurugan Pratheep, Madeleine & Arendt-Nielsen,

111 2018), and 0.82 to 0.91 for the neck region (Sterling et al., 2002; Ylinen et al., 2007; Knapstad et

112 al., 2018)).

113   Agreement and reliability are generally reported as moderate to good in both larger and

114 smaller muscles. However, the reported parameters vary between studies. This variability can

115 partly be attributed to rater characteristics as PPT measurements seem to be operator dependent.

116 Determining PPTs requires practical and psychomotor skills of the individual rater. For example,

117 variability in application rate of the pressure increase negatively impacts on reliability and

118 agreement (Linde et al., 2017). Acquiring competence in such skills can be achieved through

119 guided practice, repetition and reinforcement, going through the cognitive, associative and

120 autonomous phases of motor learning (Schmidt, Richard A. Lee, 2005; Tynjälä & Gijbels, 2012;

121   Oermann, Muckler & Morgan, 2016; Reilly, Beran-Shepler & Paschal, 2020). Novice raters may

122   still be developing their psychomotor skills, so it could be assumed that they have not yet

123   achieved comparable skills to experienced raters (Sattelmayer et al., 2016). In studies examining

124   reliability of PPTs, only one study included novice raters (Walton et al., 2011). This study

125   showed lower reliability values (ICC 0.79) compared to studies with experienced raters (ICC

126   0.92) (Waller et al., 2016). As individual psychomotor skills differ between people (Schmidt,

127   Richard A. Lee, 2005; van Duijn, Swanick & Donald, 2014), this may lead to different

128   reproducibility parameters per individual. As such, it remains unclear whether novice raters are

129   able to determine PPT measurements with an acceptable reproducibility, following limited hours

130   of training. Insight into reproducibility is important as QST measures are often performed by

131   novice raters in student projects and higher degree research projects (Geber et al., 2009). Simply

132   extrapolating reproducibility parameters from experienced raters is debatable.

133   To summarise, research about agreement and reliability parameters in novice raters when

134   performing PPTs is scarce despite the fact that novice raters often perform these measurements.

135   Therefore, this study aimed to determine intra-rater reproducibility parameters (within-session

136   and between-session agreement and reliability) of PPTs measured by novice (i.e., inexperienced)

137   raters. Additionally, this study explored the extent of individual differences in intra-rater

138   reproducibility parameters.

139

140   **METHODS**

141

142   *DESIGN*

143   Using a cross-sectional, observational test-retest study design, ten novice raters

144   determined PPTs using two types of digital algometers to assess intra-rater reproducibility

145   parameters. For agreement, the SEM was used as absolute value for the measurement error and

146   the CV was used as relative value for the measurement error compared to the mean PPT of the

147   muscle (Mokkink et al., 2010). For reliability, the $ICC_{1,1}$ was used as ratio between the

148   measurement error (Mokkink et al., 2010). All intra-rater reproducibility parameters were

149   measured per algometer.

150   Data were acquired between February and June 2019. The study was approved by the

151   Medical Ethical Committee of the University Medical Center in Groningen, The Netherlands

152   (METc 2016.613; M17.207169). All participants and raters signed an informed consent prior to

153   the measurements. Reported reproducibility parameters are based upon the COSMIN group

154   recommendations for studies on psychometrics (Gagnier et al., 2021; Mokkink et al., 2010; de

155   Vet et al., 2016). The Guideline for Reporting Reliability and Agreement Studies (GRRAS) was

156   used (Kottner et al., 2011; Gerke et al., 2018).

157

158   ***RATERS***

159   Novice raters were students in the last year of their undergraduate physiotherapy

160   program. Ten raters (six females, four males) were recruited from the Hanze University of

161   Applied Sciences in Groningen, The Netherlands. Physiotherapy students without experience in

162   determining PPTs were eligible to participate. Exclusion criteria were being physically unable to

163   perform PPTs on the selected muscles (in terms of positioning and strength) or not being able to

164   communicate in the Dutch language. Their mean (SD) age, weight and height were 23.0 (3.0)

165   years, 75 (11) kg and 179 (6) cm. For this project, all raters attended two four-hour training

166   session to perform PPT measurements with a Somedic and a Wagner digital algometer. The

167   training included an explanation of the study procedures, standardised participant instructions,

168   palpating and marking of the body locations (approximately two hours) and the use of the two

169   digital algometers (for approximately three hours per algometer). Practising assessment of PPTs

170   was performed on healthy participants who did not participate in the main study.

171

172   *PARTICIPANTS*

173   Pain-free, healthy participants were recruited from the staff and student populations of the

174   Hanze University of Applied Sciences, Academy of Health Care, in Groningen, The Netherlands,

175   and the local community. The selection of healthy participants was based on the aim to

176   determine rater-dependence in PPT measurements and as such to eliminate potential

177   confounding variables of dysfunction or disease. Convenience sampling took place via an

178   announcement on the university's intranet and an email was sent to all staff and students. All

179   participants underwent a short screening to determine eligibility (Gierthmühlen et al., 2015).

180   Inclusion criteria were being pain-free, between 18 and 65 years old and an adequate

181   understanding of the Dutch language. Participants were not permitted to use alcohol, caffeine or

182   nicotine-containing products two hours before testing, or using analgesics 24 hours prior to

183   testing (Girdler et al., 2005; Baratloo et al., 2016; Bagot et al., 2017). Exclusion criteria were the

184   presence of acute or chronic pain, neurological, orthopedic or cognitive disorders, pregnancy or

185   if participants had undergone surgery to the legs, arms or neck.

186

187   *SAMPLE SIZE*

188   Sample size estimation was performed for both agreement and reliability parameters for

189   the group of raters. For agreement, using an upper bound of the 95%CI of +15% of the SEM, a

190   minimum of 81 participants was required (per algometer) (Stratford & Goldsmith, 1997). For

191 reliability, based on α=0.05, β=0.8, an expected $ICC_{1,1}$ of >0.7 with the lower bound of the

192 95%CI of 0.5, a minimum of 63 participants was required to calculate reliability per algometer

193 (Walter, Eliasziw & Donner, 1998).

194

195 ***PPT*** *MEASUREMENTS*

196   After collecting the demographic, anthropometric, lifestyle and psychosocial data,

197 participants were randomly allocated to a rater and went into one of four rooms where they

198 received standardised instructions. Then, every participant was familiarised with the

199 measurement on the dorsal aspect of the hand by the allocated rater. Figure 1 shows the

200 procedure and methods used in the study. Four rooms were used to simultaneously measure all

201 participants. The rooms were comparable regarding size, temperature and noise levels.

202   PPTs were measured using either a Somedic digital algometer (Type II, Somedic AB,

203 Stockholm, Sweden) or Wagner digital algometer (Force Ten FDX 25, Wagner, Greenwich,

204 USA). Both digital algometers were used with a $1cm^2$ rubber tip with an application rate of ~50

205 kPa/s. The Somedic algometer provides visual feedback regarding the application rate and has a

206 hand-held switch for the participant to press when the feeling of pressure changes into painful

207 pressure. The Wagner device does not provide feedback regarding the application rate.

208 Consequently, the raters had to estimate the rate themselves. Because visual feedback about the

209 slope might assist raters in obtaining more stable PPTs, the two different algometers were used to

210 assess whether the intra-rater reproducibility parameters are influenced by the algometer

211 properties. Furthermore, the Wagner algometer has no patient switch, so participants verbally

212 indicated when their pain threshold was reached.

213 After marking the locations with a skin pencil, raters measured the PPTs three times in

214 each location, with an inter-stimulus interval of 20 seconds to avoid sensitisation of pain

215 (Reezigt et al., 2021). The measurements were taken on the dominant side, in a fixed order in

216 two circuits. In the first circuit, with the participant in sitting, PPTs were determined on the

217 tibialis anterior muscle (~5 cm distal to the tibial tuberosity) and rectus femoris muscle (~15 cm

218 proximal to the patella base) on the dominant side, alternating between the two locations (Fig.

219 1B). In the second circuit, with the participant in prone and the dominant arm above the head,

220 PPTs were determined between extensor carpi radialis brevis muscle (~5cm distal to the lateral

221 epicondyle of the humerus) and the paraspinal muscles at the C5-C6 level of the neck, again in

222 an alternating manner (Fig. 1B). The test locations were selected based on the different levels of

223 complexity of the measurements and height of the PPT. Furthermore, these muscles are often

224 selected in PPT studies and their reproducibility parameters are available for experienced raters

225 (Sterling et al., 2002; Ylinen et al., 2007; Jones, Kilgour & Comtois, 2007; Jørgensen et al.,

226 2014; Waller et al., 2016; Jakorinne, Haanpää & Arokoski, 2018; Srimurugan Pratheep,

227 Madeleine & Arendt-Nielsen, 2018; Knapstad et al., 2018; Middlebrook et al., 2020).

228 After removal of the skin markings and a 5-minute wash-out period, the rater re-marked

229 the locations and retested the PPTs.

230

231 *RANDOMISATION AND BLINDING*

232 Participants were randomly allocated to either being measured (1) with the Somedic or

233 Wagner algometer, and (2) by one of the ten raters. Computer-generated randomisation and

234 opaque envelopes were used. To blind the raters for the measurement results, they handed the

235 algometer to a research assistant after each measurement. The research assistant read the value

236   on the display, recorded the result, reset the algometer and handed the device back to the rater.

237   Participants were also blinded for their test values by shielding the display of the algometer

238   during the measurements.

239

240   *STATISTICAL ANALYSIS*

241   ***Descriptive statistics***

242   Descriptive analyses were used to report participant characteristics containing

243   demographic, anthropometric, lifestyle and psychosocial data. The normality assumption was

244   visually inspected and tested using the Shapiro-Wilk test. Normally distributed data were

245   described as mean (SD). All statistical analyses were performed in SPSS version 28 (IBM,

246   Armonk, NY, USA). Participant's characteristics measured in the Somedic algometer group and

247   Wagner algometer group were analysed to investigate comparability of the two samples and to

248   be able to indirectly compare the reproducibility parameters between the two algometers.

249   Continuous data were tested using either an independent t-test (in normal distribution) or non-

250   parametric Mann-Whitney U Test. The Chi-Square Test was used for categorical data. All PPT

251   data of the Somedic algometer were expressed in kPa. All PPT data of the Wagner algometer

252   were converted from Newton/cm^2 to kPa.

253

254   ***Reproducibility parameters***

255   All reproducibility parameters were intra-rater parameters and calculated per algometer.

256   Group reproducibility parameters reflect novice raters as a group, whereas individual

257   reproducibility parameters reflect each novice rater individually. Intra-rater absolute (i.e., SEM)

258   and relative (i.e., CV) parameters were calculated for both within-session and between-session

259 agreement. The $ICC_{1,1}$ was calculated for reliability. For all analyses, per algometer and per

260 location, outlier or influential case detection and removal were based on average PPT and

261 difference scores greater than three SDs (reducing an overrepresentation of extreme cases in a

262 normal distribution) (Parrinello et al., 2016).

263 ***Within-session agreement***

264   Within-session agreement reflects the precision within a measurement session; i.e., how

265 close the three ratings of the first test session are (Fig. 1C). First, for both the group and each

266 individual rater, the within-session agreement was calculated as an absolute value in the form of

267 the SEM. To eliminate the bias of subject variability, the calculation was based on the error mean

268 square term (EMS) of the Analysis of Variance (ANOVA) used for the $ICC_{1,1}$ calculations,

269 according to the formula: $SEM = \sqrt{EMS}$ (Weir, 2005). Confidence intervals were calculated based

270 on the sample error variance; $\left[\sqrt{\dfrac{SSE}{\hat{e}_{-}"_{\hat{\imath}\pm,dfe}^{2}}};\sqrt{\dfrac{SSE}{\hat{e}_{-}"_{1-\hat{\imath}\pm,dfe}^{2}}}\right]$, where SSE is the sum of squares error

271 (from the ANOVA), $\hat{e}_{-}"_{\hat{\imath}\pm,dfe}^{2}$ is the chi-square value for significance level α (0.05) and the

272 corresponding degrees of freedom (df) of the SSE term (Stratford & Goldsmith, 1997; Armitage

273 & Berry, 2001).

274   Second, as each sample of ratings (per location, per rater) may differ from the average

275 PPT (e.g., PPT ratings on the leg are higher than those on the arm), the CV, also called the

276 relative measurement error (i.e., SEM%), is used as an accompanying agreement parameter to

277 increase comparability between locations and raters. The CV is calculated as the SEM divided by

278 the sample average threshold and is reported as a percentage:

279 $Within-session\ CV = \dfrac{SEM_{Within-session}}{AVG\ (3\ ratings)}*100$, where

280 $SEM_{Within-session} = \sqrt{EMS_{Within-session}}$.

281   A lower SEM and lower CV reflect a better within-session agreement and as such a higher

282   precision (Atkinson & Nevill, 1998).

283

284   ***Between-session agreement***

285   Between-session agreement reflects how close the differences were between the test and

286   retest average ratings (Fig. 1C). For the test and retest sessions, the average of the three

287   measurements was used to calculate the PPT on each location. Subsequently, the between-

288   session agreement could be calculated as the absolute SEM and relative CV. Additionally, the

289   minimal detectable change of 90% (MDC90) based on the SEM (Weir, 2006) was calculated for

290   clinical purposes (Donoghue et al, 2009) by:

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

292

293   ***Dose-dependent differences***

294   When dose-dependent differences are present, a relative parameter (e.g., CV) reflects the

295   measurement error better than an absolute parameter (e.g., SEM) (Atkinson & Nevill, 2000). As

296   such, the data were analysed to see whether the differences between the test and retest increased

297   proportionally to the average threshold increment (i.e., heteroscedasticity of the data; e.g., when

298   more force is needed to determine the PPT, the difference between the test and retest may be

299   larger compared to low force PPT thresholds due to less accuracy when using more strength).

300   First, Bland-Altman plots were created to visualise and analyse the potential dose-dependent

301   differences (Bland & Altman, 1999; Carkeet, 2015; Gerke, 2020).

302   Second, linear regression lines, based on the absolute value of the differences (i.e., the

303   distance from zero) between the test and retest PPT values, were added to the Bland-Altman

304   plots. The regression analyses were used to estimate the dose-dependent difference more

305   accurately than visual inspection by using the regression coefficient and adjusted R^2 (variance

306   explained as an indicator of the strength of the relationship between increasing thresholds and

307   increasing differences).

308   In this study, presence of dose-dependent differences was defined as: a) the slope of the

309   regression line was >0.1 and b) the adjusted explained variance (R^2) by the regression model was

310   >0.1. These regression analyses were partly based on suggestions of Atkinson et al. (2000) and

311   the method of Ho et al. (2018), where we adjusted dose-dependent systematic bias to dose-

312   dependent differences (Atkinson & Nevill, 2000; Ho, 2018). For these linear regression analyses,

313   per algometer and per location, outlier or influential case detection and removal were based on

314   standardised residuals from the regression greater than three (Parrinello et al., 2016).

315

316   ***Reliability***

317   Reliability reflects the measurement error relative to the subject variability. Reliability

318   estimates were calculated using ANOVA-based ICCs (Koo & Li, 2016). The intra-rater

319   reliability represents novice raters in general, and as such each participant was rated by a

320   different rater who was randomly chosen from the set of ten raters. Therefore, ICC estimates and

321   their 95%CI were calculated based on a single rater, absolute agreement, one-way random-

322   effects model ($ICC_{1,1}$) (Koo & Li, 2016).

323   Interpretation of ICCs is based on the criteria of Koo et al. (Koo & Li, 2016): values less

324   than 0.5 indicate poor reliability, 0.50 till 0.75 indicate moderate reliability, 0.75 till 0.90

325   indicate good reliability and values greater than 0.90 indicate excellent reliability.

326

327   **RESULTS**

328

329   Ten raters measured a total of 178 participants; 88 participants were measured with the

330   Somedic algometer and 90 with the Wagner algometer. Both groups were similar in participant

331   characteristics and baseline PPTs, except for smoking status, which was significantly higher in

332   the Wagner algometer group (Table 1). One rater's data (rater 10) using the Wagner algometer,

333   were excluded as only two participants were measured with this device due to illness of this

334   rater. As a result, the exploration of individual data using the Wagner algometer contains nine

335   raters.

336

337   *GROUP REPRODUCIBILITY*

338   All within-session and between-session agreement and reliability values for the Somedic

339   and Wagner algometer are presented in Table 2. Within-session agreement, as SEM, ranged from

340   42 to 72 kPa, and as CV from 10.8% to 14.5% for the Somedic algometer. For the Wagner

341   algometer, it ranged from 42 to 74 kPa (SEM), and from 11.5% to 14.8% (CV). Measurements at

342   the rectus femoris muscle showed the best within-session agreement, whereas the measurements

343   at the extensor carpi radialis brevis muscle showed the lowest within-session agreement using

344   either algometer (Fig. 2).

345   Between-session agreement ranged from 36 to 71 kPa (SEM) and from 10.4% to 16.1%

346   (CV) for the Somedic algometer, and from 47 to 76 kPa (SEM) and from 9.4% to 14.6% (CV)

347   for the Wagner algometer. The best between-session agreement was found at the rectus femoris

348   muscle using either algometer (Fig. 2). Measurements at the extensor carpi radialis brevis muscle

349   showed the lowest agreement for the Somedic algometer (CV: 16.1% (95% CI: 14.3% - 18.4%)

350   on an average PPT of 393 kPa), whereas measurements at the paraspinal muscles at C5-C6 were

351   the lowest for the Wagner algometer (CV: 14.6% (95% CI: 13.0% - 16.7%) on an average PPT

352   of 325 kPa). The values of the MDC90 are presented in Table 2.

353   Dose-dependent differences (higher average thresholds associated with larger differences)

354   were found at the tibialis anterior muscle (slope of 0.12x, R^2=0.301) and the extensor carpi

355   radialis brevis muscle (slope of 0.15x, R^2=0.253) using the Somedic algometer. Using the

356   Wagner algometer, dose-dependent differences were present at the paraspinal muscles of C5-C6

357   (slope of 0.17x, R^2=0.225). Measurements at the rectus femoris muscle showed the lowest dose-

358   dependent differences (Fig. 3 and Fig. 4).

359   The reliability was good to excellent at all locations with both algometers. The $ICCs_{1,1}$ for

360   the Somedic algometer ranged from 0.89 to 0.93, and for the Wagner algometer from 0.87 to

361   0.95 (Fig. 2).

362

363   *EXPLORATION OF INDIVIDUALS' REPRODUCIBILITY*

364   Individual within-session agreement expressed as SEM ranged from 24 kPa to 120 kPa,

365   depending on the rater, location and algometer. Consequently, the maximal individual offset

366   compared to the group findings was -26 kPa and +46 kPa (both on the rectus femoris muscle).

367   The CV ranged from 5.1 to 21.3%, with maximal differences compared to the group findings of

368   −6.0 to +9.3 percentage points (both on the rectus femoris muscle).

369   Individual between-session agreement expressed as SEM, ranged from 15 kPa to 115

370   kPa, depending on the rater, location and algometer. Compared to the group findings, the

371   maximal offset was -55 kPa and +39 kPa (both on the tibialis anterior muscle). The CV ranged

372   from 3.6% to 26.5%, and differences ranged from -9.5 to +6.6 percentage points compared to the

373   group findings. For example, on the tibialis anterior muscle using the Wagner algometer, rater 3

374   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

375   was 17.8% (95%CI: 13.3% - 27.6%, on an average PPT of 645 kPa). Overall, most raters showed

376   acceptable agreement scores, only rater 9 scored a CV of 26.5% (95%CI: 18.3% - 50.8%, on an

377   average PPT of 304 kPa) on the extensor carpi radialis brevis muscle. In general, the extensor

378   carpi radialis brevis muscle showed the most unfavourable agreement scores, probably due to the

379   smaller size increasing the complexity of the measurement (e.g., more difficult to localise and

380   more difficult to keep the algometer on the muscle during the measurement). All individual rater

381   reproducibility parameters are shown in Figure 5 and are presented in Appendices A and B.

382   Between-session agreement and reliability were comparable for males and females except

383   for the tibialis anterior muscle using the Wagner algometer (Appendix C).

384

385   **DISCUSSION**

386

387   This study showed good intra-rater reproducibility parameters for novice raters after eight

388   hours of training, including six hours of performing PPTs, with a Somedic and Wagner digital

389   algometer (i.e., three hours per algometer). Evaluation of individual intra-rater, between- and

390   within-session agreement showed small differences per rater, but within an acceptable range.

391   The reproducibility parameters found for novice raters were comparable or even slightly

392   better than those reported for experienced raters (Sterling et al., 2002; Ylinen et al., 2007; Jones,

393   Kilgour & Comtois, 2007; Jørgensen et al., 2014; Waller et al., 2016; Jakorinne, Haanpää &

394   Arokoski, 2018; Srimurugan Pratheep, Madeleine & Arendt-Nielsen, 2018; Knapstad et al.,

395   2018; Middlebrook et al., 2020). Between-session agreement was comparable, depending on the

396 location. In most locations, this study revealed equal or better CVs (e.g., ~22% at the extensor

397 carpi radialis brevis muscle compared to 16% in our study), except at the tibialis anterior muscle.

398 A previous study found a slightly better CV at the tibialis anterior muscle in experienced raters

399 (8% vs 10-14%) (Sterling et al., 2002) compared to our study. In contrast, these authors found

400 slightly lower ICC values for reliability (ICC 0.65 to 0.94) compared to our study (ICC 0.87 to

401 0.95). The mean PPTs we obtained at the different locations were comparable with previously

402 reported values (Sterling et al., 2002; Ylinen et al., 2007; Jones, Kilgour & Comtois, 2007;

403 Jørgensen et al., 2014; Waller et al., 2016; Jakorinne, Haanpää & Arokoski, 2018; Srimurugan

404 Pratheep, Madeleine & Arendt-Nielsen, 2018; Knapstad et al., 2018; Middlebrook et al., 2020).

405 One other study included novice undergraduate physiotherapy students, using a Wagner

406 algometer at the tibialis anterior muscle and paraspinal muscles in a comparable population of

407 healthy, young participants (mean of 25.4 years) (Walton et al., 2011). They found slightly lower

408 reliability (ICC 0.79 versus 0.88 in our study) and within-session agreement values (CV 19%

409 versus 13% in our study). The lower reliability and within session agreement that they found

410 could probably be explained by the lower number of training hours compared to our study (1

411 hour versus 8 hours) (Walton et al., 2011). Earlier studies also attributed lower individual

412 reliability scores to systematic errors from novice raters (Chung, Um & Kim, 1992; Goulet,

413 Clark & Flack, 1993). Due to potential different learning curves (not every rater learns equally

414 fast and the number of participants measured to reach acceptable reproducibility may differ), the

415 training could have a variable effect on each rater's performance (van Duijn, Swanick & Donald,

416 2014). Additionally, it is unknown how much training is needed to perform PPTs with an

417 acceptable reproducibility.

418   In this study, an arbitrary number of ten raters was chosen, as we hypothesised that

419   individual (psychomotor) skills could influence the reproducibility parameters. Using a group of

420   multiple raters increased the generalisability of the findings and allowed exploration of

421   differences in the reproducibility parameters of individual raters. Even though novice raters show

422   good group reproducibility parameters compared to experienced raters, individual differences in

423   novice were identified. No previous studies assessed individual reproducible parameters in

424   experienced or novice raters. Multiple raters showed a CV and SEM which were over two times

425   smaller than the group's CV and SEM, as such they have excellent measurement skills

426   techniques compared to their peer raters.

427   This study has some limitations. Ideally, for exploration of rater differences, all raters

428   should have tested the same group of participants to minimise the subject variability. However,

429   different subsamples were chosen to reduce the number of tests per participant and prevent

430   potentially attention bias, salience and sensitisation effects (Villemure & Bushnell, 2002;

431   Goffaux et al., 2007; Hall & Rodríguez, 2017; Moore et al., 2019). Consequently, we could not

432   directly compare reproducibility values between the two different algometers. However, the two

433   samples were comparable based on demographics and PPT values. Another limitation is that not

434   all raters measured an equal number of participants (min = 6; max = 14), which influenced the

435   width of the confidence intervals and the amount of experience gained while acquiring the

436   reproducibility data. Remarkably, raters who measured a higher number of participants showed

437   noticeable higher reproducibility parameters. This could possibly be explained by the differences

438   in gained experience through measuring. The individual parameters should, however, be

439   interpreted with caution due to the low subsample sizes.

440   We found some individual differences in reproducibility values, which were within

441   acceptable ranges. Consequently, novice raters who attended an eight-hour training can

442   participate as raters in research projects and use PPTs in clinical practice adequately. Since

443   individual differences in reproducibility parameters may exist, researchers and clinicians should

444   be cautious when using reproducibility parameters of other raters from former studies. Future

445   studies should focus on quantifying these differences and include methods to explore whether

446   rater experience or psychomotor skills (e.g., strength or dexterity) may explain these differences.

447   Furthermore, the relationship between the duration of training and reproducibility parameters

448   should be explored further to recommend the minimal duration of training needed to perform

449   PPTs adequately.

450   In conclusion, although slight differences between individual assessors exist, this study

451   revealed that novice raters show good reproducibility parameters in determining PPTs across

452   four body locations following three hours of practice of PPT measures in addition to surface

453   palpation skills.

455 **REFERENCES**

456

457 Amiri M, Alavinia M, Singh M, Kumbhare D. 2021. Pressure Pain Threshold in Patients With
458   Chronic Pain: A Systematic Review and Meta-Analysis. *American journal of physical*
459   *medicine & rehabilitation* 100:656–674. DOI: 10.1097/PHM.0000000000001603.
460 Armitage P, Berry G. 2001. *Statistical Methods in Medical Research*. Wiley-Blackwell, 115–
461   117, 221–223.
462 Atkinson G, Nevill AM. 1998. Statistical Methods For Assessing Measurement Error
463   (Reliability) in Variables Relevant to Sports Medicine. *Sports Medicine* 26:217–238. DOI:
464   10.2165/00007256-199826040-00002.
465 Atkinson G, Nevill A. 2000. Measures of Reliability in Sports Medicine and Science. *Sports*
466   *Medicine* 30:375–381. DOI: 10.2165/00007256-200030050-00005.
467 Bagot KS, Wu R, Cavallo D, Krishnan-Sarin S. 2017. Assessment of pain in adolescents:
468   Influence of gender, smoking status and tobacco abstinence. *Addictive Behaviors* 67:79–85.
469   DOI: 10.1016/j.addbeh.2016.12.010.
470 Baratloo A, Rouhipour A, Forouzanfar MM, Safari S, Amiri M, Negida A. 2016. The Role of
471   Caffeine in Pain Management: A Brief Literature Review. *Anesthesiology and Pain*
472   *Medicine* 6. DOI: 10.5812/aapm.33193.
473 Bisset L, Evans K, Tuttle N. 2015. Reliability of 2 protocols for assessing pressure pain
474   threshold in healthy young adults. *Journal of manipulative and physiological therapeutics*
475   38:282–287. DOI: 10.1016/J.JMPT.2015.03.001.
476 Bland JM, Altman DG. 1999. Measuring agreement in method comparison studies. *Statistical*
477   *Methods in Medical Research* 8:135–160. DOI: 10.1177/096228029900800204.
478 Bruton A, Conway JH, Holgate ST. 2000. Reliability: What is it, and how is it measured?
479   *Physiotherapy* 86:94–99. DOI: 10.1016/S0031-9406(05)61211-4.
480 Carkeet A. 2015. Exact Parametric Confidence Intervals for Bland-Altman Limits of Agreement.
481   *Optometry and Vision Science* 92:e71–e80. DOI: 10.1097/OPX.0000000000000513.
482 Chung S-C, Um B-Y, Kim H-S. 1992. Evaluation of Pressure Pain Threshold in Head and Neck
483   Muscles by Electronic Algometer: Intrarater and Interrater Reliability. *CRANIO®* 10:28–34.
484   DOI: 10.1080/08869634.1992.11677888.
485 Donoghue, D., Murphy, A., Jennings, A., McAuliffe, A., O'Neil, S., Charthaigh, E.N., Griffin,
486   E., Gilhooly, L., Lyons, M., Galvin, R., Gallagher, S., Ward, S., Mhaille, E.N., Stokes, E.K.
487   (2009). How much change is true change? The minimum detectable change of the Berg
488   Balance Scale in elderly people. *J Rehabil Med* 41, 343–346 DOI: 10.2340/16501977-0337.
489 van Duijn AJ, Swanick K, Donald EK. 2014. Student Learning of Cervical Psychomotor Skills
490   Via Online Video Instruction Versus Traditional Face-to-Face Instruction. *Journal of*
491   *Physical Therapy Education* 28:94–102. DOI: 10.1097/00001416-201410000-00015.
492 Fischer AA. 1987. Pressure algometry over normal muscles. Standard values, validity and
493   reproducibility of pressure threshold. *Pain* 30:115–126. DOI: 10.1016/0304-
494   3959(87)90089-3.
495 Gagnier, J.J., Lai, J., Mokkink, L.B., Terwee, C.B. (2021). COSMIN reporting guideline for
496   studies on measurement properties of patient-reported outcome measures. Qual Life Res 30,
497   2197–2218 DOI: 10.1007/s11136-021-02822-4.

498   Geber C, Scherens A, Pfau D, Nestler N, Zenz M, Tölle T, Baron R, Treede RD, Maier C. 2009.
499   [Procedure for certification of QST laboratories]. *Schmerz (Berlin, Germany)* 23:65–69.
500   DOI: 10.1007/S00482-008-0771-4.
501   Gerke O. 2020. Reporting standards for a bland-altman agreement analysis: A review of
502   methodological reviews. *Diagnostics* 10:1–17. DOI: 10.3390/diagnostics10050334.
503   Gerke O, Möller S, Debrabant B, Halekoh U, Odense Agreement Working Group. 2018.
504   Experience Applying the Guidelines for Reporting Reliability and Agreement Studies
505   (GRRAS) Indicated Five Questions Should Be Addressed in the Planning Phase from a
506   Statistical Point of View. *Diagnostics (Basel, Switzerland)* 8:69. DOI:
507   10.3390/diagnostics8040069.
508   Gierthmühlen J, Enax-Krumova EK, Attal N, Bouhassira D, Cruccu G, Finnerup NB, Haanpää
509   M, Hansson P, Jensen TS, Freynhagen R, Kennedy JD, Mainka T, Rice ASC, Segerdahl M,
510   Sindrup SH, Serra J, Tölle T, Treede RD, Baron R, Maier C. 2015. Who is healthy? Aspects
511   to consider when including healthy volunteers in QST-based studies - A consensus
512   statement by the EUROPAIN and NEUROPAIN consortia. *Pain* 156. DOI:
513   10.1097/j.pain.0000000000000227.
514   Girdler SS, Maixner W, Naftel HA, Stewart PW, Moretz RL, Light KC. 2005. Cigarette
515   smoking, stress-induced analgesia and pain perception in men and women. *Pain* 114:372–
516   385. DOI: 10.1016/j.pain.2004.12.035.
517   Goffaux P, Redmond WJ, Rainville P, Marchand S. 2007. Descending analgesia - When the
518   spine echoes what the brain expects. *Pain* 130:137–143. DOI: 10.1016/j.pain.2006.11.011.
519   Goulet J -P, Clark GT, Flack VF. 1993. Reproducibility of examiner performance for muscle and
520   joint palpation in the temporomandibular system following training and calibration.
521   *Community Dentistry and Oral Epidemiology* 21:72–77. DOI: 10.1111/j.1600-
522   0528.1993.tb00724.x.
523   Hall G, Rodríguez G. 2017. Habituation and conditioning: Salience change in associative
524   learning. *Journal of Experimental Psychology: Animal Learning and Cognition* 43:48–61.
525   DOI: 10.1037/xan0000129.
526   Ho KM. 2018. Using linear regression to assess dose-dependent bias on a Bland-Altman plot.
527   *Journal of Emergency and Critical Care Medicine* 2:68–68. DOI:
528   10.21037/JECCM.2018.08.02.
529   Jakorinne P, Haanpää M, Arokoski J. 2018. Reliability of pressure pain, vibration detection, and
530   tactile detection threshold measurements in lower extremities in subjects with knee
531   osteoarthritis and healthy controls. *Scandinavian Journal of Rheumatology* 47:491–500.
532   DOI: 10.1080/03009742.2018.1433233.
533   Jones DH, Kilgour RD, Comtois AS. 2007. Test-retest reliability of pressure pain threshold
534   measurements of the upper limb and torso in young healthy women. *The journal of pain*
535   8:650–656. DOI: 10.1016/J.JPAIN.2007.04.003.
536   Jørgensen R, Ris I, Falla D, Juul-Kristensen B. 2014. Reliability, construct and discriminative
537   validity of clinical testing in subjects with and without chronic neck pain. *BMC*
538   *musculoskeletal disorders* 15. DOI: 10.1186/1471-2474-15-408.
539   Keele KD. 1954. PAIN-SENSITIVITY TESTS. *The Lancet* 263:636–639. DOI: 10.1016/S0140-
540   6736(54)92347-8.
541   Knapstad MK, Nordahl SHG, Naterstad IF, Ask T, Skouen JS, Goplen FK. 2018. Measuring
542   pressure pain threshold in the cervical region of dizzy patients-The reliability of a pressure

543   algometer. *Physiotherapy research international : the journal for researchers and clinicians*
544   *in physical therapy* 23. DOI: 10.1002/PRI.1736.
545   Koo TK, Li MY. 2016. A Guideline of Selecting and Reporting Intraclass Correlation
546   Coefficients for Reliability Research. *Journal of chiropractic medicine* 15:155–63. DOI:
547   10.1016/j.jcm.2016.02.012.
548   Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M,
549   Streiner DL. 2011. Guidelines for Reporting Reliability and Agreement Studies (GRRAS)
550   were proposed. *International Journal of Nursing Studies* 48:661–671. DOI:
551   10.1016/j.ijnurstu.2011.01.016.
552   Linde LD, Dinesh ; Kumbhare A, Joshi M, Srbely JZ. 2017. The Relationship between Rate of
553   Algometer Application and Pain Pressure Threshold in the Assessment of Myofascial
554   Trigger Point Sensitivity. DOI: 10.1111/papr.12597.
555   Middlebrook N, Heneghan NR, Evans DW, Rushton A, Falla D. 2020. Reliability of temporal
556   summation, thermal and pressure pain thresholds in a healthy cohort and musculoskeletal
557   trauma population. *PLoS ONE* 15. DOI: 10.1371/JOURNAL.PONE.0233521.
558   Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet
559   HCW. 2010. The COSMIN study reached international consensus on taxonomy,
560   terminology, and definitions of measurement properties for health-related patient-reported
561   outcomes. *Journal of Clinical Epidemiology* 63:737–745. DOI:
562   10.1016/j.jclinepi.2010.02.006.
563   Moore DJ, Meints SM, Lazaridou A, Johnson D, Franceschelli O, Cornelius M, Schreiber K,
564   Edwards RR. 2019. The Effect of Induced and Chronic Pain on Attention. *The Journal of*
565   *Pain* 20:1353–1361. DOI: 10.1016/J.JPAIN.2019.05.004.
566   Mücke M, Cuhls H, Radbruch L, Baron R, Maier C, Tölle T, Treede RD, Rolke R. 2016.
567   Quantitative sensorische Testung (QST). *Schmerz*:1–8. DOI: 10.1007/s00482-015-0093-2.
568   Nijs J, Lahousse A, Kapreli E, Bilika P, Saraçoğlu İ, Malfliet A, Coppieters I, de Baets L,
569   Leysen L, Roose E, Clark J, Voogt L, Huysmans E. 2021. Nociplastic Pain Criteria or
570   Recognition of Central Sensitization? Pain Phenotyping in the Past, Present and Future.
571   *Journal of Clinical Medicine* 10:3203. DOI: 10.3390/jcm10153203.
572   Nussbaum EL, Downes L. 1998. Reliability of clinical pressure-pain algometric measurements
573   obtained on consecutive days. *Physical Therapy* 78:160–169. DOI: 10.1093/ptj/78.2.160.
574   Oermann MH, Muckler VC, Morgan B. 2016. Framework for teaching psychomotor and
575   procedural skills in nursing. *Journal of Continuing Education in Nursing* 47:278–282. DOI:
576   10.3928/00220124-20160518-10.
577   Ohrbach R, Gale EN. 1989. Pressure pain thresholds, clinical assessment, and differential
578   diagnosis: reliability and validity in patients with myogenic pain. *Pain* 39:157–169. DOI:
579   10.1016/0304-3959(89)90003-1.
580   Parrinello CM, Grams ME, Sang Y, Couper D, Wruck LM, Li D, Eckfeldt JH, Selvin E, Coresh
581   J. 2016. Iterative Outlier Removal: A Method for Identifying Outliers in Laboratory
582   Recalibration Studies. *Clinical Chemistry* 62:966–972. DOI:
583   10.1373/clinchem.2016.255216.
584   Reezigt RR, Kielstra SC, Coppieters MW, Scholten-Peeters GGM. 2021. No relevant differences
585   in conditioned pain modulation effects between parallel and sequential test design. A cross-
586   sectional observational study. *PeerJ* 9. DOI: 10.7717/PEERJ.12330.

587 Reilly M, Beran-Shepler K, Paschal KA. 2020. Pedagogy for Effective Learning of Clinical
588   Skills: An Integrated Laboratory Model. *Journal of Physical Therapy Education* 34:234–
589   241. DOI: 10.1097/jte.0000000000000145.
590 Rolke R, Baron R, Maier C, Tölle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F,
591   Bötefür IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W, Maihöfner
592   C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. 2006. Quantitative
593   sensory testing in the German Research Network on Neuropathic Pain (DFNS):
594   Standardized protocol and reference values. *Pain* 123:231–243. DOI:
595   10.1016/j.pain.2006.01.041.
596 Sattelmayer M, Elsig S, Hilfiker R, Baer G. 2016. A systematic review and meta-analysis of
597   selected motor learning principles in physiotherapy and medical education. DOI:
598   10.1186/s12909-016-0538-z.
599 Schmidt, Richard A. Lee TD. 2005. *Motor control and learning: A behavioral emphasis, 4th ed.*
600   *- PsycNET*.
601 Srimurugan Pratheep N, Madeleine P, Arendt-Nielsen L. 2018. Relative and absolute test-retest
602   reliabilities of pressure pain threshold in patients with knee osteoarthritis. *Scandinavian*
603   *journal of pain* 18:229–236. DOI: 10.1515/SJPAIN-2018-0017.
604 Sterling M, Jull G, Carlsson Y, Crommert L. 2002. Are cervical physical outcome measures
605   influenced by the presence of symptomatology? *Physiotherapy research international : the*
606   *journal for researchers and clinicians in physical therapy* 7:113–121. DOI:
607   10.1002/PRI.248.
608 Stratford PW, Goldsmith CH. 1997. Use of the standard error as a reliability index of interest:
609   An applied example using elbow flexor strength data. *Physical Therapy* 77:745–750. DOI:
610   10.1093/ptj/77.7.745.
611 Tynjälä P, Gijbels D. 2012. Changing World: Changing Pedagogy. In: *Transitions and*
612   *Transformations in Learning and Education*. Dordrecht: Springer Netherlands, 205–222.
613   DOI: 10.1007/978-94-007-2312-2_13.
614 de Vet HCW, Terwee CB, Knol DL, Bouter LM. 2006. When to use agreement versus reliability
615   measures. *Journal of Clinical Epidemiology* 59:1033–1039. DOI:
616   10.1016/j.jclinepi.2005.10.015.
617 de Vet, H.C.W., Terwee, C.B., Mokkink, L.B., Knol, D.L. (2016). Measurement in Medicine
618   (Cambridge University Press).
619 Villemure C, Bushnell MC. 2002. Cognitive modulation of pain: How do attention and emotion
620   influence pain processing? *Pain* 95:195–199. DOI: 10.1016/S0304-3959(02)00007-6.
621 Waller R, Smith AJ, O'Sullivan PB, Slater H, Sterling M, Alexandra McVeigh J, Straker LM.
622   2016. Pressure and cold pain threshold reference values in a large, young adult, pain-free
623   population. *Scandinavian Journal of Pain* 13:114–122. DOI: 10.1016/j.sjpain.2016.08.003.
624 Walter SD, Eliasziw M, Donner A. 1998. Sample size and optimal designs for reliability studies.
625   *Statistics in Medicine* 17:101–110. DOI: 10.1002/(SICI)1097-
626   0258(19980115)17:1<101::AID-SIM727>3.0.CO;2-E.
627 Walton D, MacDermid J, Nielson W, Teasell R, Chiasson M, Brown L. 2011. Reliability,
628   Standard Error, and Minimum Detectable Change of Clinical Pressure Pain Threshold
629   Testing in People With and Without Acute Neck Pain. *Journal of Orthopaedic & Sports*
630   *Physical Therapy* 41:644–650. DOI: 10.2519/jospt.2011.3666.

631 Weir JP. 2005. Quantifying Test-Retest Reliability Using the Intraclass Correlation Coefficient
632   and the SEM. *The Journal of Strength and Conditioning Research* 19:231. DOI:
633   10.1519/15184.1.
634 Yarnitsky D, Bouhassira D, Drewes AM, Fillingim RB, Granot M, Hansson P, Landau R,
635   Marchand S, Matre D, Nilsen KB, Stubhaug A, Treede RD, Wilder-Smith OHG. 2015.
636   Recommendations on practice of conditioned pain modulation (CPM) testing. *European*
637   *Journal of Pain (United Kingdom)* 19:805–806. DOI: 10.1002/ejp.605.
638 Ylinen J, Nykänen M, Kautiainen H, Häkkinen A. 2007. Evaluation of repeatability of pressure
639   algometry on the neck muscles for clinical use. *Manual Therapy* 12:192–197. DOI:
640   10.1016/j.math.2006.06.010.
641

Figure 1

Visual representation of the methods

Panel A represents the randomisation procedure for algometer type and rater allocation; Panel B shows the measurement procedure; Panel 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.

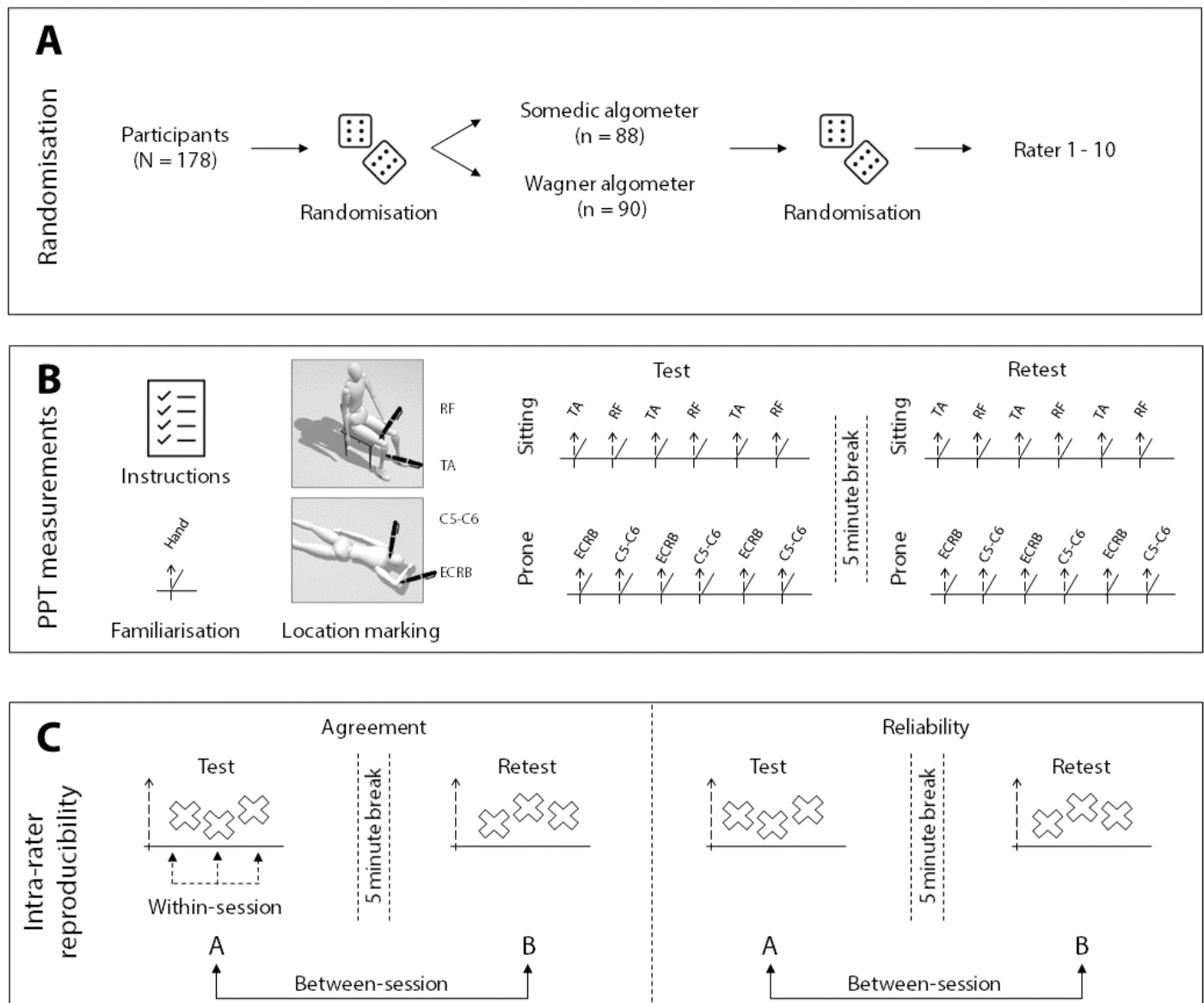



Figure 2

Within-session and between-session agreement and reliability for both types of digital algometers

Upper panels show agreement and reliability of the Somedic algometer; lower panels 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 et al (2016): values ≥0.75 indicate good reliability and ≥0.9 excellent reliability.

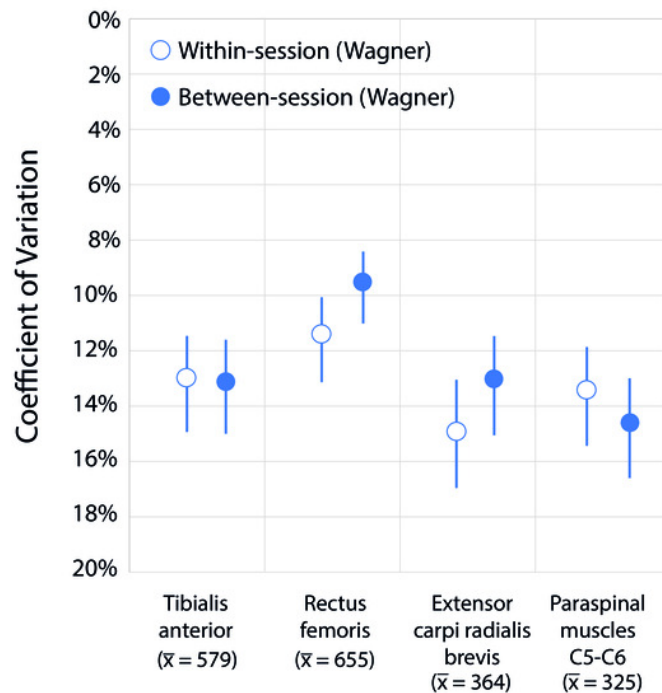
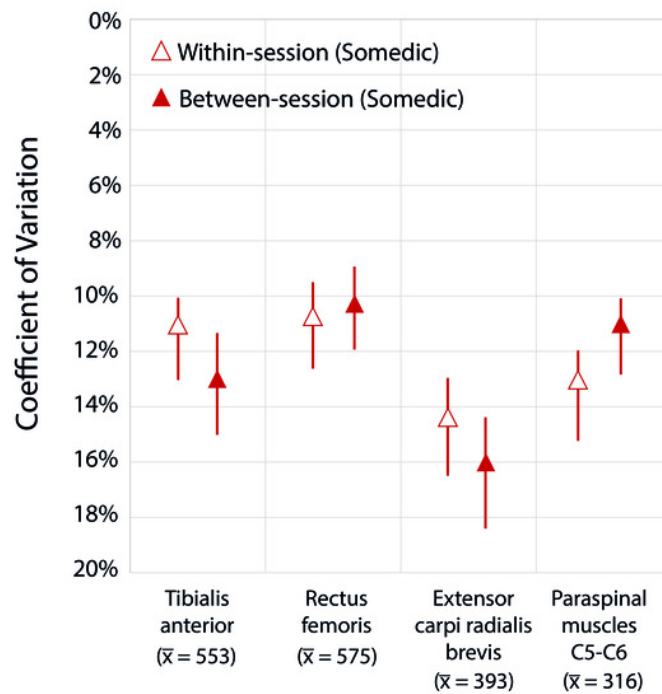
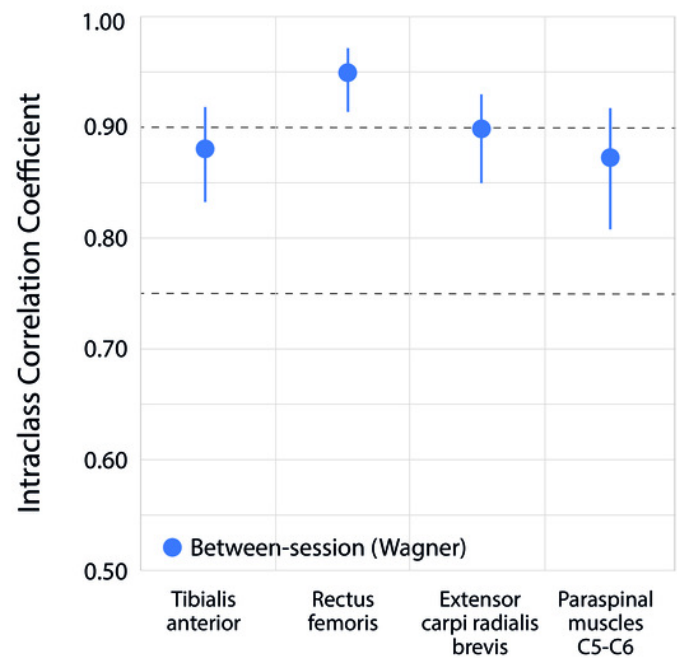
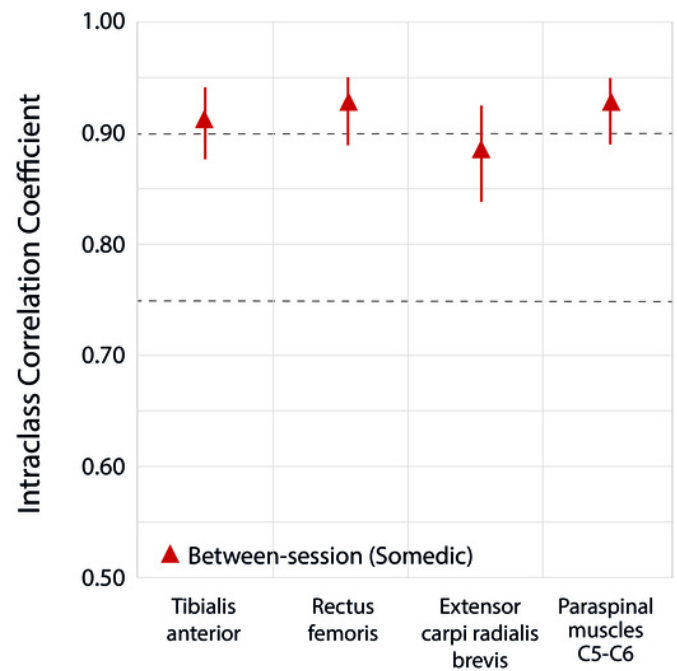



Figure 3

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).

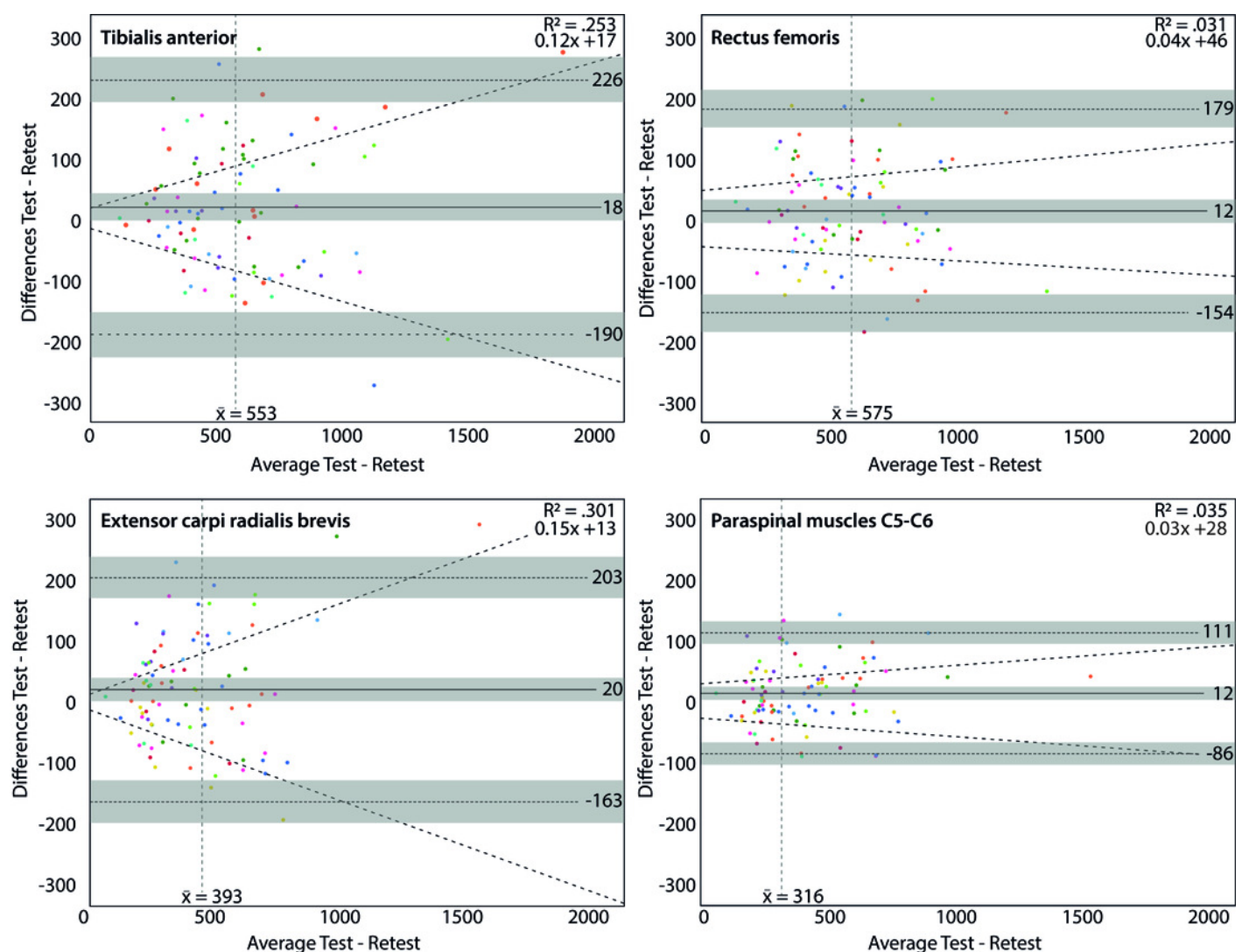



Figure 4

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).

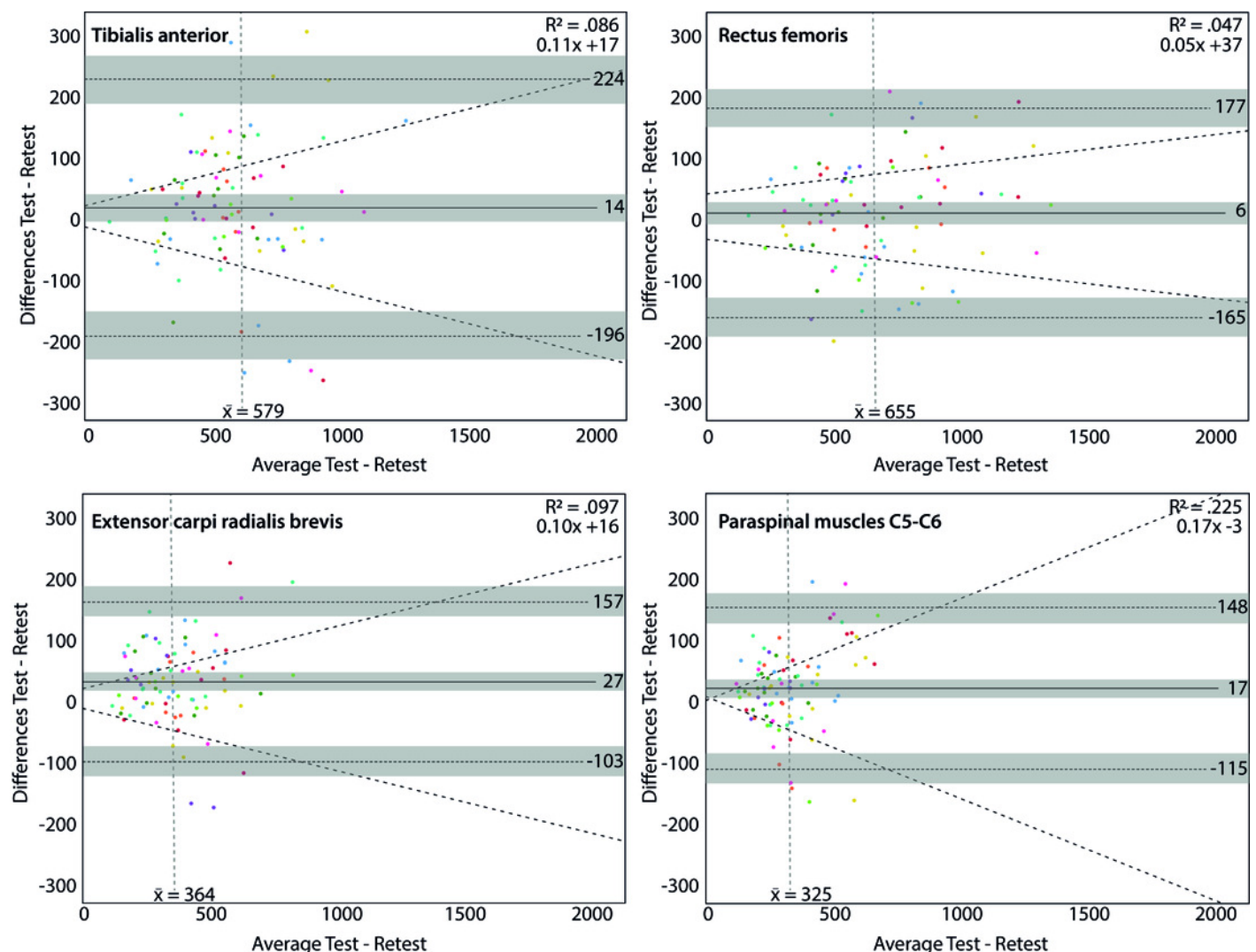
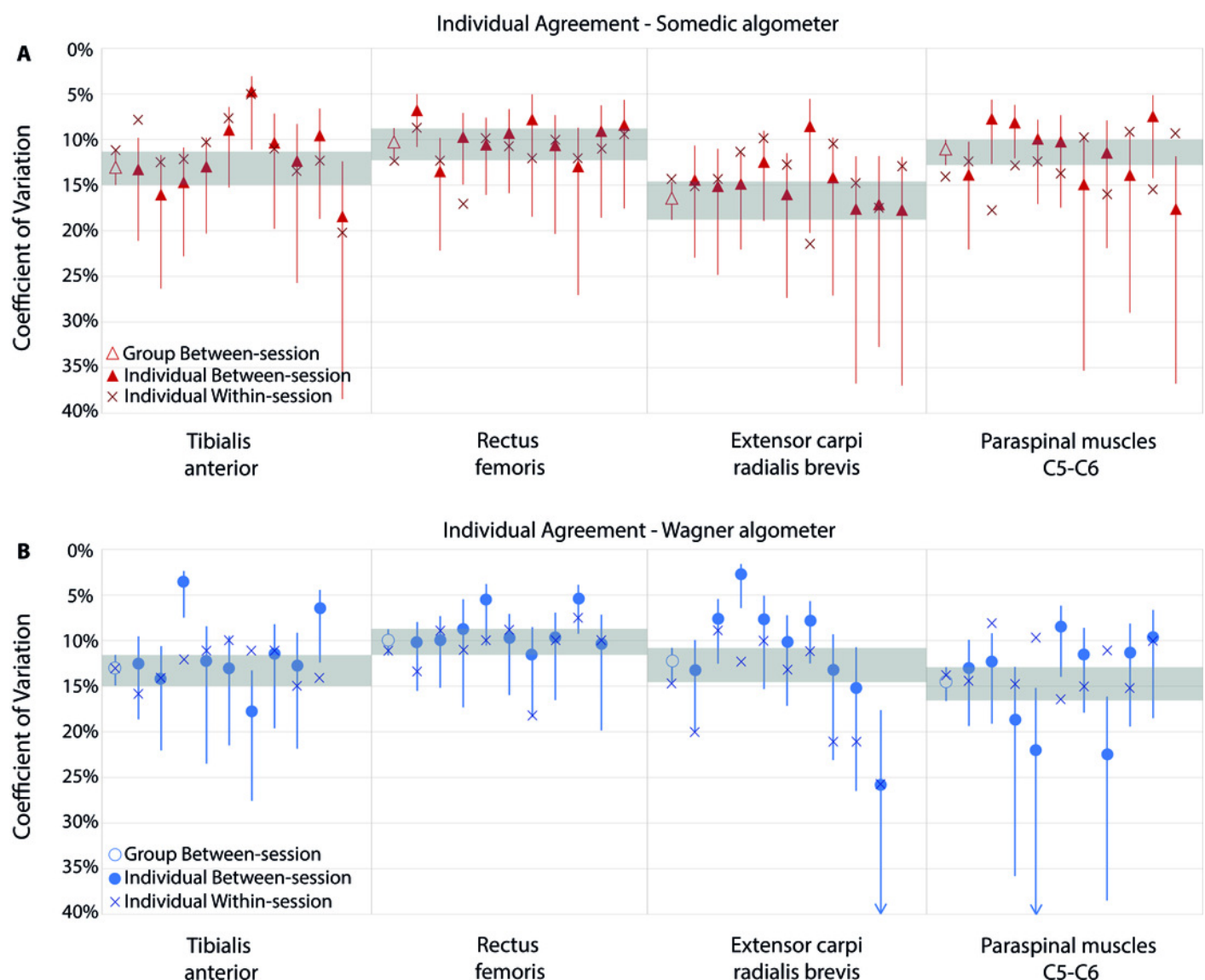



Figure 5

Individual exploration of reproducibility parameters of novice raters

Agreement expressed as coefficient of variation (CV), including 95% Confidence Intervals, of the Somedic algometer (Panel A) and Wagner algometer (Panel B). Shaded areas show the overall group between-session agreement confidence interval as reference. Each filled mark represents an individual rater

Characteristics of the participants.

Data are presented as mean (SD) unless otherwise specified, ᵃMedian (IQR), ᵇChi Square Test, ᶜMann-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

1   **Table 1** – Characteristics of the participants.

	Somedic algometer group	Wagner algometer group	P-Value
Demographics	$n = 88$	$n = 90$	
Sex, male (%)	44 (50%)	49 (54%)	$p = 0.65$[b]
Age, years	22 (20-25)[a]	22 (20-23)[a]	$p = 0.50$[c]
Anthropometrics			
Height, cm	179 (11)	178 (9)	$p = 0.35$
Weight, kg	74 (12)	75 (13)	$p = 0.58$
BMI, kg/m²	22.9 (2.7)	23.6 (3.2)	$p = 0.12$
Lifestyle			
Sports, hours/week	5 (2.5-6.5)[a]	5 (3.0-7.5)[a]	$p = 0.69$[c]
Alcohol use, AUDIT (total score)	8.0 (4.5)	8.2 (4.4)	$p = 0.93$
Smoking, n (%)	6 (6.8%)	16 (17.8%)	$p = 0.04$[b*]
Sleep, PSQI (total score)	4.3 (1.9)	4.7 (2.4)	$p = 0.28$
Psychosocial			
Anxiety, GAD-7 (total score)	2 (1-3)[a]	2 (0-3)[a]	$p = 0.68$[c]
Depression, CES-D (total score)	3 (1-6)[a]	4 (2-6)[a]	$p = 0.38$[c]

2   Data are presented as mean (SD) unless otherwise specified, [a]Median (IQR), [b]Chi Square Test,

3   [c]Mann-Whitney U Test. *Significant difference between both groups: $\chi^2_{(1,\ n=178)} = 4.93$, p = 0.04.

4   BMI - Body mass index; AUDIT - Alcohol Use Disorders Identification Test; PSQI - Pittsburgh

5   Sleep Quality Index; GAD-7 - Generalised anxiety disorder 7 items; CES-D - Center for

6   Epidemiologic Studies Depression Scale

Reproducibility parameters on group level

kPa - kilo Pascal

1 **Table 2** – Reproducibility parameters on group level

Somedic algometer	Mean PPT (kPa)	Within-session agreement		Between-session agreement			Between-session reliability
		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% - 14.8%)	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)

2 kPa, kilo Pascal