Measurement Error of a Simplified Protocol for Quantitative Sensory Tests in Chronic Pain Patients

Monika Müller, MD,*† José Alberto Biurrun Manresa, PhD,‡ Andreas Limacher, PhD,†§ Konrad Streitberger, MD,* Peter Jüni, MD,** Ole Kasealer Andersen, PhD,‡ and Michele Curatolo, MD, PhD‡**

Background and Objectives: Large-scale application of Quantitative Sensory Tests (QST) is impaired by lacking standardized testing protocols. One unclear methodological aspect is the number of records needed to minimize measurement error. Traditionally, measurements are repeated 3 to 5 times, and their mean value is considered. When transferring QST to a clinical setting, reducing the number of records would be desirable to meet the time constraints encountered in a routine clinical environment and to reduce the testing burden to chronic pain patients. However, there might be a trade-off between measurement error and number of records. We determined the measurement error of a single versus the mean of 3 records of pressure pain detection threshold (PPDT), electrical pain detection threshold (EPDT), and nociceptive withdrawal reflex threshold (NWRT) in 429 chronic pain patients recruited in a routine clinical setting.

Methods: We calculated intraclass correlation coefficients and performed a Bland-Altman analysis.

Results: Intraclass correlation coefficients were all clearly greater than 0.75, and Bland-Altman analysis showed minute systematic errors with small point estimates and narrow 95% confidence intervals. Reducing the number of records from traditionally 3 to only 1 did not lead to relevant measurement error in PPDT, EPDT, or NWRT.

Conclusions: This study contributes to a standardized QST protocol, and based on the minimal measurement error of 1 single record of PPDT, EPDT, and NWRT, we submit to reduce the testing burden. This would allow saving time, resources, and patient discomfort.

(Reg Anesth Pain Med 2017;42: 660–668)

From the *Department of Anaesthesiology and Pain Medicine, Inselspital, Bern University Hospital, Bern; †Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland; ‡Center for Sensory–Motor Interaction, Department of Health Science and Technology, Aalborg University, Aalborg, Denmark; §Clinical Trials Unit Bern, Department of Clinical Research, University of Bern, Bern, Switzerland; †[Applied Health Research Centre (AHRC) Li Ka Shing Knowledge Institute of St Michael’s Hospital, University of Toronto, Toronto, Ontario, Canada; and **Department of Anaesthesiology and Pain Medicine, University of Washington, Seattle, WA.

Accepted for publication March 20, 2017.

Address correspondence to: Michele Curatolo, MD, PhD, Department of Anaesthesiology and Pain Medicine, University of Washington, 1595 NE Pacific St, Box 356540 Seattle, WA 98195 (e-mail: curatolo@uw.edu).

M.M. and J.A.B.M. contributed equally to this work.

Author contributions: M.M., J.A.B.M., and M.C. conceived the study. M.M., J.A.B.M., and K.S. were responsible for data collection. M.M., J.A.B.M., and A.L. did the data analysis. M.M. and M.C. wrote the first draft of the paper, and all authors contributed to the final draft.

The study was funded by the Scientific Funds of the Department of Anaesthesiology and Pain Medicine, Inselspital, Bern University Hospital. The results of this study were not presented at any conference; nor was the manuscript under consideration by any other journal. The authors declare no conflict of interest.

Copyright © 2017 by American Society of Regional Anesthesia and Pain Medicine.

ISSN: 1098-7339
DOI: 10.1097/AAP.0000000000000640

Original Article

CHRONIC AND INTERVENTIONAL PAIN

Participants

In 2011, we included QST as part of a routine multimodal patient assessment at our tertiary care outpatient’s facility. The present analysis was performed in a subsample of participants in whom we estimated the prevalence of central hypersensitivity as assessed by QSTs. All patients who suffered from pain lasting more than 3 months and were referred between July 1, 2011, and June 30, 2012, to the Department of Anaesthesiology and Pain Medicine at the University Hospital of Bern were included. The exclusion criteria for QST assessment were age younger than 18 years, no informed consent or inability to understand the test protocol, and known or suspected skin conditions that might interfere with QST (e.g., keratosis pilaris, psoriasis, eczema).

QST procedures

QST were performed using a standardized protocol as previously described.9,10 The recorded data included the PPDT, EPDT, and NWRT. The PPDT was assessed using a series of graded mechanical stimuli. The EPDT was measured using electrical stimuli with a constant intensity of 2 mA and a variable stimulus duration (1–10 s). The NWRT was assessed during a continuous nociceptive stimulation with a constant heat stimulus (42°C). The clinical cutoff values used were ≥30 N for the PPDT, ≥110 ms for the EPDT, and ≥71°C for the NWRT.

Statistical analysis

The statistical analysis was performed using the statistical software package R.11 The significance level was set at 5%. A p value of less than 0.05 was considered statistically significant. The differences in the prevalence of central hypersensitivity were assessed by a chi-square test. The association between the prevalence of central hypersensitivity and the demographic and clinical characteristics was assessed using multivariate logistic regression analysis. A backward selection method was used to select the variables with p values of less than 0.5 for inclusion in the final model. The results of the logistic regression analysis are presented as odds ratios with 95% confidence intervals. Correlation coefficients were calculated using Pearson’s correlation.

RESULTS

The results of this study are presented in Table 1. The prevalence of central hypersensitivity was 17.4% (95% confidence interval: 13.8%–21.1%). The demographic and clinical characteristics of the patients with and without central hypersensitivity are presented in Table 2.

The prevalence of central hypersensitivity was significantly associated with age, gender, educational level, and employment status. The odds ratios for the variables with a p value of less than 0.5 are presented in Table 3. The prevalence of central hypersensitivity was also significantly associated with the presence of comorbid conditions, such as depression, anxiety, and sleep disturbances. The odds ratios for the variables with a p value of less than 0.5 are presented in Table 4.

The results of the correlation analysis are presented in Table 5. The correlation coefficients were statistically significant for all the variables included in the analysis. The results of the correlation analysis are presented in Table 6. The correlation coefficients were statistically significant for all the variables included in the analysis.

CONCLUSIONS

The prevalence of central hypersensitivity was 17.4% in this population of chronic pain patients. The prevalence of central hypersensitivity was significantly associated with age, gender, educational level, employment status, and the presence of comorbid conditions such as depression, anxiety, and sleep disturbances. These results are consistent with previous studies that have shown a higher prevalence of central hypersensitivity in older age groups and women.1,2,12,13 The results of this study also suggest that the presence of comorbid conditions is associated with a higher prevalence of central hypersensitivity.

The results of this study have important implications for the management of chronic pain. The identification of patients with central hypersensitivity can help guide the selection of appropriate interventions. For example, patients with central hypersensitivity may benefit from interventions that target the central nervous system, such as cognitive-behavioral therapy and neurostimulation. These results also have implications for the design of clinical trials. The prevalence of central hypersensitivity can be used to identify the appropriate sample size for clinical trials.

In conclusion, the prevalence of central hypersensitivity was 17.4% in this population of chronic pain patients. The prevalence of central hypersensitivity was significantly associated with age, gender, educational level, employment status, and the presence of comorbid conditions such as depression, anxiety, and sleep disturbances. These results are consistent with previous studies and have important implications for the management of chronic pain and the design of clinical trials. The prevalence of central hypersensitivity can be used to identify the appropriate sample size for clinical trials.
Pain Medicine of the University Hospital of Bern in Switzerland were eligible for the study. We excluded patients with neurological comorbidities potentially affecting the neurological function of the lower extremity to be tested (palsy, paresthesia, polyneuropathy), patients with rheumatic inflammatory disease, and patients with chronic pain as result of evident peripheral lesions (oncological pain in the region of infiltration by a primary tumor, metastasis, or peripheral vascular disorder). Other reasons of exclusion were psychiatric comorbidity except unipolar depressive disorder, pregnancy, and language problems. We informed all patients in written form on the background of the QSTs and on the use of their data for scientific purposes. All patients gave informed consent before the tests, which were performed according to a prospective protocol approved by the local research ethics committee and in accordance with the Declaration of the World Medical Association.9

Sociodemographic, Psychological, and Clinical Characteristics

We recorded sociodemographic, psychological, and clinical characteristics for descriptive purposes. Sociodemographic characteristics were sex, age, and working status. Psychological characteristics were depression and catastrophizing assessed with the Coping Strategies Questionnaire (CSQ).11 Clinical characteristics were body mass index and pain-related variables such as history of trauma or surgery related to pain, pain duration, pain intensity, pain-related life interference, type of pain, and current pain medication. We measured pain intensity and pain-related sleep interference with a numerical rating scales (NRS) ranging from 0 (no pain or interference) to 10 (worst pain or interference imaginable). We classified the type of pain as musculoskeletal, neuropathic, orofacial, or visceral and summarized pain syndromes rarely encountered in our clinic such as noncervicogenic headache, complex regional pain syndrome, and phantom limb pain in the separate class of “rare pain syndromes.” We based the definition of the type of pain on our previous publication of the prevalence of central hypersensitivity as assessed by QSTs.8 We defined daily intake of pain medication as intake of at least 1 pain medication of the following classes: opioids, nonsteroidal anti-inflammatory drugs, acetaminophen, metamizole, or analgesic comedication such as antidepressants and/or anticonvulsants.

Quantitative Sensory Testing

We implemented the assessment of PPDT at the second toe, pain detection threshold after single electrical stimulation (EPDT), and the NWRT after single electrical stimulation because these assessments can be easily applied in routine clinical practice and in a previous investigation displayed good discriminative ability for hypersensitivity among 26 tests.12 Four different health care professionals routinely performed all QSTs at the extremity contralateral to the side of most pain. In case of bilateral pain, the testing extremity was randomly selected according to a computer-generated list. The assessment was standardized according to a prespecified protocol and included standardized oral explanation of the experimental setting, training session, and recording phase for all patients. The training session is considered essential to familiarize patients with the stimulation procedure.13 These sessions took on average 5 minutes and included 3 training records for pressure stimulation to assess PPDT and 3 training records for electrical stimulation to assess EPDT and NWRT. Thereafter we again performed 3 records to definitively assess PPDT, EPDT, and NWRT. We measured pain detection threshold after pressure stimulation at the center of the pulp of the second toe using an electronic pressure algometer with a 1 cm² surface probe (Somedic AB, Södala, Sweden).13 Pressure was increased at a rate of 30 kPa/s until patients perceived the stimulus as painful, what we defined as pain detection threshold. In case the stimulation was not perceived as painful, we considered the maximum stimulation intensity of 1000 kPa as threshold. We performed electrical stimulation of the sural nerve with bipolar surface Ag/AgCl electrodes placed distal to the lateral malleolus. A computer-controlled constant current stimulator (NCS System, Evidence 3102 evo; Neurosoft, Moscow, Russia) delivered a train-of-5 1-millisecond square-wave pulses of an overall duration of 25 milliseconds, which was perceived as a single stimulus by the patients. In a single increasing intensity staircase, the current intensity was increased from 1 mA in steps of 1 mA until the electrical stimulus was perceived as painful (EPDT), and until an NWR of the biceps femoris with an amplitude higher than 20 μV for at least 10 milliseconds in the 50- to 150-millisecond poststimulation interval was elicited (NWRT).14,15 We did not apply a maximum current intensity.

Statistical Analysis

Quantitative Sensory Tests were not performed or repeated 3 times for logistic reasons in 103 patients (Fig. 1). We excluded these patients from the method comparison analysis but evaluated a potential selection bias by comparing their sociodemographic, psychological, and clinical characteristics, with patients included in the analysis using χ² and Student t tests. We also compared characteristics of patients with and without successful NWRT assessment using χ² and Student t test. To evaluate measurement error when reducing the number of records to 1, we first calculated intraclass correlation coefficients (ICCs) based on mixed-effects linear regression models with a random intercept for subjects to account for clustering of repeated records within patients. The ICC was calculated as the estimated variance of the measurements between subjects divided by the sum of the estimated variances of the measurements between and within subjects. Then, we performed a method comparison analysis as suggested by Bland and Altman.16–18 We considered the value of record 1 as measurement method M2 and compared it with the mean value of records 1, 2, and 3, which was considered as measurement method M1.

In a first set of mixed-effects linear regression models, we used the 3 single records of PPDT, EPDT, and NWRT as dependent variable and the subject identifier as random effect. Intraclass correlation coefficients in these models estimate the concordance of the 3 records and show how much of the total variance can be attributed to within-patient variability and between-patient variability. We then included an indicator variable for record number to investigate a possible systematic effect of sequential recording. In a second set of mixed-effects linear regression models, we used measurement methods (M2 vs M1) of PPDT, EPDT, and NWRT as dependent variable and the subject identifier as random effect. The ICCs in these models estimate the concordance between the 2 measurement methods, thus representing the measurement error relative to the variability between patients. Intraclass correlation coefficient values greater than 0.75 suggest an excellent, between 0.4 and 0.75 a moderate, and less than 0.4 a poor correlation of the 3 records or of the 2 measurement methods.19 Then, we performed a Bland-Altman analysis.16–18 Bland and Altman suggested that the extent of agreement between 2 measurement methods could be examined by comparing the differences between the pairs of measurements with the mean of each pair.16–18 The mean difference between the 2 methods explains whether there is a systematic error in the new method (M2) as compared with the standard method (M1). The limits of agreement are defined as the systematic error (or mean difference) ± 1.96 times the standard error.
the SD of the difference. They delimit the range within which 95% of the differences between results of QSTs may be expected to lie if the number of records is limited to 1. In close relation to this definition, the coefficient of repeatability (CR) is defined as the value below which 95% of the absolute differences between thresholds may be expected to lie. If the systematic error is close to 0, the limits of agreement and the CR are expected to be similar. We generated Bland-Altman plots allowing visual inspection of the measurement error by plotting the mean of the 2 measurement methods against the difference of the 2 methods. If the measurement error is unrelated to the size of the outcome variable, a random scatter can be expected.

Finally, we conducted 2 sensitivity analyses and several stratified exploratory analyses. First, we changed the definition of measurement method M1 to account for the high arithmetic correlation between the mean value of records 1, 2, and 3 and the value of record 1. We thus compared the value of record 1 (M2) with the mean value of records 2 and 3 (M1). Second, we stratified the analysis according to assessor to evaluate if measurement error was comparable in all 4 assessors. To evaluate the effect of gender (male vs female), age (≥65 vs <65 years), depression (BDI-FS ≥4 vs <4), catastrophizing (CSQ median value of ≥3.17 vs <3.17), pain duration (below 1 year, 1–2 years, >2 years), type of pain, pain intensity (NRS median value of ≥6 vs <6), daily

FIGURE 1. Flowchart of patients undergoing first consultation between July 1, 2011, and June 30, 2012. *Two patients with neuroborreliosis, 2 patients with bilateral paresis of unknown origin, 3 patients with bilateral sensibility disorders of unknown origin, and 2 patients with restless legs syndrome.
intake of any medication (yes vs no), and daily intake of different classes of pain medications (yes vs no) on measurement error, we performed secondary exploratory analyses stratified according to these variables (Appendix, Supplemental Digital Content 1, http://links.lww.com/AAP/A210). We performed all statistical analyses with STATA (version 12.1; StataCorp, College Station, Texas).

**RESULTS**

We screened 724 patients who were referred to our pain clinic for a first consultation between July 1, 2011, and June 30, 2012, and excluded 192 patients (Fig. 1). The most important reason for exclusion was concomitant neurological disorder, accounting for 54% of all exclusions. For logistic reasons, most commonly due to time constraints, we did not perform any QST in 65 patients (12%) and were unable to repeat QST 3 times in 38 patients (7%) of all included patients. These 103 patients with incomplete QSTs were significantly older ($P = 0.002$) as compared with the 429 patients analyzed, but did not differ in any other patients’ characteristics listed in Table 1. Data on PPDT and EPDT were complete. Maximum PPDTs for records 1, 2, and 3 were 982, 736, and 707 kPa, respectively. We were unable to determine NWRT in 130 patients because electrical stimulation became intolerable before a reflex could be detected. We found no differences in characteristics of patients with and without successful NWRT assessment using $\chi^2$ and Student’s $t$ test and, except for age and daily intake of any pain medication. Mean ages were 48.6 (SD, 14.8) years and 52.5 (SD, 14.5) years in patients with and without NWRT, respectively ($P = 0.01$). A smaller proportion of patients with NWRT took any pain medication (69% vs 80%, $P = 0.02$). Intraclass correlation coefficients and results of the Bland-Altman analysis were based on 429 patients for PPDT and EPDT and on a subsample of 299 patients for NWRT.

Table 1 shows sociodemographic, psychological, and clinical characteristics of all 429 patients. More than half were female (56%), suffered from depression (54%), and had pain lasting more than 2 years (58%). One hundred forty-seven patients (33%) were unemployed because of their chronic pain condition. Most frequently, patients suffered from musculoskeletal pain (287 patients [67%]). Three hundred ten patients (72%) took at least 1 pain medication on a daily basis. One hundred eleven (26%) of all patients regularly took opioids; 126 (29%), nonsteroidal anti-inflammatory drugs; 44 (10%), metamizol; 150 (35%), acetaminophen; and 128 (30%) took coanalgesics such as antidepressants or anticonvulsants. Figure 2 illustrates results of PPDT, EPDT, and NWRT for records 1, 2, and 3. The mean values and SDs were similar for all 3 records. We found evidence for a systematic effect of sequential recording for PPDT and EPDT but not for NWRT. Values of PPDT significantly increased, and values for EPDT significantly decreased for every additional record. The ICCs of the 3 records were 0.91 (95% confidence interval [CI], 0.89–0.92), 0.95 (95% CI, 0.94–0.95), and 0.90 (95% CI, 0.88–0.91) for PPDT, EPDT, and NWRT, respectively.

Table 2 displays the results of the main analysis comparing record 1 with the mean of all 3 records, as well as the sensitivity analysis comparing record 1 with the mean of records 2 and 3 for all 3 QSTs. As for the main analysis, the point estimates of the ICCs were 0.96, 0.97, and 0.95 for PPDT, EPDT, and NWRT, respectively, with lower bounds of the 95% CIs greater than 0.93. The Bland-Altman analysis showed a slight overestimation for PPDT if the number of records is limited to one, as compared with the mean of 3 records, with a systematic error of 5.98 kPa (95% CI, 2.81–9.14 kPa). For EPDT, results were slightly underestimated when reducing the number of records to 1. For NWRT, again the point estimate indicated that results were underestimated when reducing the number of records to 1; however, the 95% CI was compatible with both underestimation and overestimation (systematic error of −0.11 mA; 95% CI, −0.26–0.05 mA). The CRs were 66.7 kPa, 1.44 mA, and 2.75 mA for PPDT, EPDT, and NWRT, respectively. As for the sensitivity analysis, ICCs were robust for all 3 QSTs. Measurement errors were slightly larger than in the main analysis with systematic errors of 8.72 kPa, −0.55 mA, and −0.36 mA, as well as CRs of 100 kPa, 2.31 mA, and 4.19 mA for PPDT, EPDT, and NWRT, respectively. Figure 3 illustrates Bland-Altman plots of PPDT, EPDT, and NWRT of the mean and the sensitivity analysis for visual inspection of the maximal measurement error.

Table 3 displays the results of the second sensitivity analysis investigating measurement error after stratifying for assessor. Again, point estimates and lower bounds of 95% CI of all ICCs were clearly greater than 0.75, and ICCs per testing modality were comparable between assessors. The Bland-Altman analysis showed numerically different systematic errors between the 4 assessors for reducing the number of records to 1; however, the 95% CI was compatible with both underestimation and overestimation (systematic error of −0.11 mA; 95% CI, −0.26–0.05 mA).
all 3 tests, with PPDT being the test with the largest difference between assessor on the one hand and the largest discrepancy from the main analysis on the other hand. For EPDT and NWRT, however, the differences between assessors were small, and the point estimates per assessor were comparable to the results of the main analysis.

RESULTS

In this large-scale-method comparison analysis of 429 chronic pain patients recruited in a routine clinical setting, we found that reducing the number of records from traditional 3,6,7 to only 1 did not lead to relevant measurement error in PPDT, EPDT, and NWRT. Point estimates and lower bounds of 95% CIs of ICCs were clearly greater than 0.75 for all QSTs and thus showed excellent correlation of the 3 sequential records. When directly comparing the value of record 1 with the mean value of records 1, 2, and 3, ICCs of the 2 measurement methods were 0.95 or higher, again suggesting excellent correlation of both measurement methods.

Results of the Bland-Altman analysis showed minute systematic errors with small point estimates and narrow 95% CIs for all 3 QSTs. The CRs were 66.7 kPa, 1.44 mA, and 2.75 mA for PPDT, EPDT, and NWRT, respectively. The low measurement error is evident when observing the narrow lower and upper limits of agreement (Fig. 3), as compared with the corresponding wide ranges of measurement reported in Figure 2. For EPDT, most of the differences between a single measurement and the average 3 will be smaller than 1.7 mA, which corresponds to a variability of less than 2 current steps (usually of 1 mA) of the electrical stimulator. For PPDT, most of the differences will be smaller than 70 kPa, which corresponds to a variability of approximately 2 seconds on the algometer pressure test (at a rate of 30 kPa/s). Results of the sensitivity analysis showed slightly larger measurement error when comparing the value of record 1 with the mean value of records 2 and 3. The exploratory secondary analyses suggest that there are no relevant effects of gender, age, depression, catastrophizing, pain duration, type of pain, pain intensity, and daily intake of medication on measurement error. However, the results have to be interpreted with caution because we performed these analyses post hoc, and stratification partly resulted in small subgroups.

To our knowledge, only 1 study assessed measurement error when reducing the number of records to 1; the investigation was limited to NWRT. 15 The authors found high correlations between NWRT of 1 record and mean NWRT of 2 records and concluded that performing only 1 record is acceptable. The findings of our study confirm these results and showed that this was also the case for PPDT and EPDT. Although there are several limitations when using correlation coefficients to estimate measurement error, the authors refrained to perform a formal method comparison analysis in accordance with the method suggested by Bland and Altman. 16,17 This is the most recommended and commonly used statistical method to estimate measurement error. 18,19 We therefore interpret our results in the context of previous test-retest reliability studies investigating whether QSTs reliably yield similar results if repeated at different time points. 20,22 This includes 2 studies that determined measurement error over time for PPDT, 21,22 2 studies for EPDT, 20,22 and 1 study for NWRT. 20 Only our previous studies 20,22 used Bland-Altman analysis to estimate measurement error. In our first study, we compared QSTs of 3 sessions with a mean of 7.7 days between the sessions. We found mean CRs of the 3

FIGURE 2. Box plots of each record of (A) PPDT, (B) EPDT, and (C) NWRT after single electrical stimulation. Values are medians with interquartile range (IQR), means with SDs, and β coefficients (Coef) with corresponding 95% CIs and P values based on mixed-effects linear regressions.

DISCUSSION

In this large-scale-method comparison analysis of 429 chronic pain patients recruited in a routine clinical setting, we found that reducing the number of records from traditional 3,6,7 to only 1 did not lead to relevant measurement error in PPDT, EPDT, and NWRT. Point estimates and lower bounds of 95% CIs of ICCs were clearly greater than 0.75 for all QSTs and thus showed excellent correlation of the 3 sequential records. When directly comparing the value of record 1 with the mean value of records 1, 2, and 3, ICCs of the 2 measurement methods were 0.95 or higher, again suggesting excellent correlation of both measurement methods. Results of the Bland-Altman analysis showed minute systematic errors with small point estimates and narrow 95% CIs for all 3 QSTs. The CRs were 66.7 kPa, 1.44 mA, and 2.75 mA for PPDT, EPDT, and NWRT, respectively. The low measurement error is evident when observing the narrow lower and upper limits of agreement (Fig. 3), as compared with the corresponding wide ranges of measurement reported in Figure 2. For EPDT, most of the differences between a single measurement and the average 3 will be smaller than 1.7 mA, which corresponds to a variability of less than 2 current steps (usually of 1 mA) of the electrical stimulator. For PPDT, most of the differences will be smaller than 70 kPa, which corresponds to a variability of approximately 2 seconds on the algometer pressure test (at a rate of 30 kPa/s). Results of the sensitivity analysis showed slightly larger measurement error when comparing the value of record 1 with the mean value of records 2 and 3. The exploratory secondary analyses suggest that there are no relevant effects of gender, age, depression, catastrophizing, pain duration, type of pain, pain intensity, and daily intake of different medications on measurement error. However, the results have to be interpreted with caution because we performed these analyses post hoc, and stratification partly resulted in small subgroups.

To our knowledge, only 1 study assessed measurement error when reducing the number of records to 1; the investigation was limited to NWRT. 15 The authors found high correlations between NWRT of 1 record and mean NWRT of 2 records and concluded that performing only 1 record is acceptable. The findings of our study confirm these results and showed that this was also the case for PPDT and EPDT. Although there are several limitations when using correlation coefficients to estimate measurement error, the authors refrained to perform a formal method comparison analysis in accordance with the method suggested by Bland and Altman. 16,17 This is the most recommended and commonly used statistical method to estimate measurement error. 18,19 We therefore interpret our results in the context of previous test-retest reliability studies investigating whether QSTs reliably yield similar results if repeated at different time points. 20,22 This includes 2 studies that determined measurement error over time for PPDT, 21,22 2 studies for EPDT, 20,22 and 1 study for NWRT. 20 Only our previous studies 20,22 used Bland-Altman analysis to estimate measurement error. In our first study, we compared QSTs of 3 sessions with a mean of 7.7 days between the sessions. We found mean CRs of the 3
sessions of 2.3 and 5 mA for EPDT and NWRT, respectively. In our second study, we recorded QSTs 3 times within 1 session and calculated the CR for all 3 records to assess within-session reliability. We found a mean CR for all 3 records of 1.3 mA for EPDT and mean CR of 90 kPa for PPDT, respectively. Taking ICCs greater than 0.75 into account, the reliability and thus measurement error were concluded acceptable in both studies. We could confirm our previous findings in this study. We thus submit measurement error to be small when reducing the number of records to 1. Although we found evidence for a systematic effect of sequential recording, the overall correlation of the 3 records was still excellent, with ICCs for the 3 records equal or greater than 0.90. This systematic effect is likely the result of sensitization or habituation, although the quantitative changes were minimal. We believe that by reducing the number of records from 3 to 1 we would also be able to avoid such systematic effects as result of sequential recording. Of note, we performed a short training session to familiarize patients with the stimulation procedure before registering the 3 records. This is considered essential before formal testing is started and thus is common practice of QST testing protocols as it was in the test-retest reliability studies. The ideal clinical QST protocol would require very little to no familiarization to further save time, resources, and patient discomfort. Future research should therefore aim at randomizing patients to different familiarization procedures to evaluate the effect of familiarization on measurement error.

Up-to-date recommendations about the number of records necessary to yield a valid test result were not based on evidence but on expert opinion. This is the first large-scale method comparison analysis in chronic pain patients to formally assess the trade-off between measurement error and number of records needed to yield a valid QST result. Strengths of our study include the large sample size with high associated statistical precision and the study setting. We recruited all patients in a routine clinical environment, which suggests generalizability of our results. The measurement error is likely to be even smaller in a highly controlled experimental research setting. In 20% of all eligible patients, we were unable to perform a complete QST assessment because of logistical reasons, most commonly time constraints in the clinical setting. To evaluate selection bias, we compared these patients with the included patients and found no differences in patients’ characteristics except for age. Therefore, results are unlikely to be invalidated by selection bias. The results are contingent on our selection of QST and not necessarily applicable to other QST modalities. However, the results were consistent across 2 different stimulation modalities (pressure and electrical) and 2 different response modalities (pain and reflex thresholds). Our reliability analysis was done within the same session (ie, addressed “internal consistency”). Internal consistency does not necessarily imply “stability over time” (ie, reliability over weeks or months), which may be the target of future research. One limitation was that patients were not randomly allocated to the 4 assessors performing the QSTs. To address this limitation, we performed a sensitivity analysis stratified for assessor. We found different measurement errors for the 4 assessors with largest discrepancies for PPDT and small differences for EPDT and NWRT. As the allocation to the assessor was not randomized, no firm conclusions can be drawn from this finding. Pressure algometry relies on manual testing, which may introduce uncontrolled variability, as opposed to electrical stimulation. The findings suggest that trial sessions

---

**TABLE 2. Results of the Method Comparison Analysis for 3 Quantitative Sensory Tests Showing Between- and Within-Subject SDs and ICCs Based on Mixed-Effects Linear Regression Models and Systematic Error (SE) and Limits of Agreement (LoA) Based on Bland-Altman Analysis**

### Main Analysis: Value of Record 1 (M2) vs Mean Value of Records 1, 2, and 3 (M1)

<table>
<thead>
<tr>
<th></th>
<th>SD Between (95% CI)</th>
<th>SD Within (95% CI)</th>
<th>ICC (95% CI)</th>
<th>SE* (95% CI)</th>
<th>Upper LoA† (95% CI)</th>
<th>Lower LoA† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDT‡ kPa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>118 (110–127)</td>
<td>24 (22–26)</td>
<td>0.96 (0.95–0.97)</td>
<td>5.98 (2.81–9.14)</td>
<td>72.66 (67.15–78.16)</td>
<td>–60.70 (–66.21 to –55.20)</td>
<td></td>
</tr>
<tr>
<td>EPDT§ mA</td>
<td>3.28 (3.06–3.51)</td>
<td>0.59 (0.55–0.63)</td>
<td>0.97 (0.96–0.97)</td>
<td>–0.24 (–0.31 to –0.17)</td>
<td>1.21 (1.09–1.33)</td>
<td>–1.68 (–1.80 to –1.56)</td>
</tr>
<tr>
<td>NWRT</td>
<td></td>
<td>mA</td>
<td>4.49 (4.14–4.88)</td>
<td>0.98 (0.91–1.07)</td>
<td>0.95 (0.94–0.96)</td>
<td>–0.11 (–0.26 to 0.05)</td>
</tr>
</tbody>
</table>

### Sensitivity Analysis: Value of Record 1 (M2) vs Mean Value of Records 2 and 3 (M1)

<table>
<thead>
<tr>
<th></th>
<th>SD Between (95% CI)</th>
<th>SD Within (95% CI)</th>
<th>ICC (95% CI)</th>
<th>SE* (95% CI)</th>
<th>Upper LoA† (95% CI)</th>
<th>Lower LoA† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDT‡ kPa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>114 (106–122)</td>
<td>36 (34–38)</td>
<td>0.91 (0.89–0.93)</td>
<td>8.72 (3.98–13.47)</td>
<td>108.72 (100.46–116.97)</td>
<td>–91.28 (–99.53 to –83.02)</td>
<td></td>
</tr>
<tr>
<td>EPDT§ mA</td>
<td>3.28 (3.07–3.51)</td>
<td>0.54 (0.50–0.57)</td>
<td>0.97 (0.97–0.98)</td>
<td>–0.55 (–0.66 to –0.44)</td>
<td>1.76 (1.57–1.95)</td>
<td>–2.86 (–3.05 to –2.67)</td>
</tr>
<tr>
<td>NWRT</td>
<td></td>
<td>mA</td>
<td>4.49 (4.14–4.88)</td>
<td>0.98 (0.91–1.07)</td>
<td>0.95 (0.94–0.96)</td>
<td>–0.36 (–0.60 to –0.12)</td>
</tr>
</tbody>
</table>

All values are presented with corresponding 95% CIs.
*SE that corresponds to mean difference between M1 and M2.
†Lower and upper limits of agreement that correspond to SE ±1.96 SDs of the SE.
‡PPDT second toe: PPDT at the second toe (n = 429).
§EPDT after single electrical stimulation (n = 429).
||NWRT after single electrical stimulation (n = 299).
within the assessor team may be considered to improve consistency of measurements. Another limitation was that in 30% of all patients pain of electrical stimulation became intolerable before a reflex was evoked, and we performed a complete case analysis in those patients with NWRT. To address this limitation, we compared patients' characteristics of those patients with and without NWRT and found no differences except for age and daily intake of any pain medicaments. We again argue that the magnitude of measurement error is unlikely to be influenced by missing data. A major strength of the presented study is the comprehensive statistical analysis performed to evaluate measurement error. To account for high arithmetic correlation between the mean value of records 1, 2, and 3 and the value of record 1, a sensitivity analysis was added to compare the value of record 1 with the mean value of records 2 and 3. Results were much the same as in the main analysis, and thus an underestimation of measurement error due to high arithmetic correlation is unlikely.
While QSTs are well established in research, their use for clinical practice is not widespread. One reason is the limited data on the ability of QST to support decision making. On the other hand, recent research has shown potential prognostic value and ability to predict efficacy of medications, although the results are not consistent across studies.23–25 Importantly, detecting central sensitization with QST may help patients better understand their pain condition. Simplifying the testing procedure is expected to lead to broader use in clinical practice and hopefully to more sensitization with QST may help patients better understand their pain condition. Simplifying the testing procedure is expected to lead to broader use in clinical practice and hopefully to more widespread use in clinical practice, an evidence-based approach toward standardization of QST assessment procedures is needed. This study contributes to a standardized QST testing protocol. One single record of PPDT, EPDT, and NWRT is associated with minimal measurement error, compared with the mean of 3 sequential records. Based on this result, it seems acceptable to limit the number of records to 1. We expect to save time, resources, and patients’ discomfort when reducing the number of records to 1.

### REFERENCES


### TABLE 3. Results of the Stratified Sensitivity Method Comparison Analysis Showing Between- and Within-Subject SD and Systematic Error (SE) and Limits of Agreement (LoA) Based on Bland-Altman Analysis

<table>
<thead>
<tr>
<th>Assessor 1</th>
<th>n</th>
<th>SD Between (95% CI)</th>
<th>SD Within (95% CI)</th>
<th>ICC (95% CI)</th>
<th>SE* (95% CI)</th>
<th>Upper LoA† (95% CI)</th>
<th>Lower LoA† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDT‡ (kPa)</td>
<td>101</td>
<td>137 (119, 158)</td>
<td>34 (30, 39)</td>
<td>0.94 (0.92, 0.96)</td>
<td>13.57 (4.45, 22.69)</td>
<td>105.97 (90.25, 121.70)</td>
<td>−78.83 (−94.55, −63.12)</td>
</tr>
<tr>
<td>EPDT§ (mA)</td>
<td>101</td>
<td>3.78 (3.28, 4.34)</td>
<td>0.58 (0.51, 0.67)</td>
<td>0.98 (0.97, 0.98)</td>
<td>−0.21 (−0.35, −0.06)</td>
<td>1.27 (1.02, 1.52)</td>
<td>−1.68 (−1.93, −1.43)</td>
</tr>
<tr>
<td>NWRT</td>
<td></td>
<td>63</td>
<td>3.09 (2.57, 3.71)</td>
<td>0.94 (0.79, 1.12)</td>
<td>0.92 (0.87, 0.95)</td>
<td>−0.47 (−0.77, −0.16)</td>
<td>1.95 (1.43, 2.47)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessor 2</th>
<th>n</th>
<th>SD Between (95% CI)</th>
<th>SD Within (95% CI)</th>
<th>ICC (95% CI)</th>
<th>SE* (95% CI)</th>
<th>Upper LoA† (95% CI)</th>
<th>Lower LoA† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDT‡ (kPa)</td>
<td>88</td>
<td>108 (93, 125)</td>
<td>16 (14, 19)</td>
<td>0.98 (0.97, 0.99)</td>
<td>−0.29 (−5.19, 4.61)</td>
<td>45.97 (37.53, 54.40)</td>
<td>−46.55 (−54.98, −38.12)</td>
</tr>
<tr>
<td>EPDT§ (mA)</td>
<td>88</td>
<td>3.06 (2.63, 3.56)</td>
<td>0.64 (0.55, 0.74)</td>
<td>0.96 (0.93, 0.97)</td>
<td>−0.49 (−0.64, −0.34)</td>
<td>0.92 (0.67, 1.18)</td>
<td>−1.90 (−2.16, −1.64)</td>
</tr>
<tr>
<td>NWRT</td>
<td></td>
<td>65</td>
<td>4.08 (3.42, 4.86)</td>
<td>0.77 (0.65, 0.91)</td>
<td>0.97 (0.94, 0.98)</td>
<td>0.07 (−0.19, 0.33)</td>
<td>2.19 (1.74, 2.64)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessor 3</th>
<th>n</th>
<th>SD Between (95% CI)</th>
<th>SD Within (95% CI)</th>
<th>ICC (95% CI)</th>
<th>SE* (95% CI)</th>
<th>Upper LoA† (95% CI)</th>
<th>Lower LoA† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDT‡ (kPa)</td>
<td>155</td>
<td>110 (99, 124)</td>
<td>17 (15, 19)</td>
<td>0.98 (0.97, 0.97)</td>
<td>−0.91 (−4.84, 3.03)</td>
<td>48.68 (41.87, 55.49)</td>
<td>−50.50 (−57.31, −43.69)</td>
</tr>
<tr>
<td>EPDT§ (mA)</td>
<td>155</td>
<td>3.11 (2.78, 3.48)</td>
<td>0.50 (0.46, 0.56)</td>
<td>0.97 (0.96, 0.98)</td>
<td>−0.05 (−0.14, 0.05)</td>
<td>1.17 (1.00, 1.34)</td>
<td>−1.26 (−1.43, −1.10)</td>
</tr>
<tr>
<td>NWRT</td>
<td></td>
<td>114</td>
<td>3.94 (3.45, 4.50)</td>
<td>0.90 (0.79, 1.02)</td>
<td>0.95 (0.93, 0.97)</td>
<td>0.08 (−0.16, 0.32)</td>
<td>2.65 (2.24, 3.06)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assessor 4</th>
<th>n</th>
<th>SD Between (95% CI)</th>
<th>SD Within (95% CI)</th>
<th>ICC (95% CI)</th>
<th>SE* (95% CI)</th>
<th>Upper LoA† (95% CI)</th>
<th>Lower LoA† (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDT‡ (kPa)</td>
<td>85</td>
<td>104 (89, 122)</td>
<td>26 (23, 31)</td>
<td>0.94 (0.91, 0.96)</td>
<td>16.01 (8.72, 23.30)</td>
<td>83.58 (71.04, 96.11)</td>
<td>−51.56 (−64.10, −39.03)</td>
</tr>
<tr>
<td>EPDT§ (mA)</td>
<td>85</td>
<td>2.91 (2.50, 3.40)</td>
<td>0.69 (0.59, 0.80)</td>
<td>0.95 (0.92, 0.97)</td>
<td>−0.36 (−0.53, −0.18)</td>
<td>1.28 (0.98, 1.59)</td>
<td>−2.00 (−2.30, −1.69)</td>
</tr>
<tr>
<td>NWRT</td>
<td></td>
<td>57</td>
<td>6.40 (5.30, 7.72)</td>
<td>1.35 (1.12, 1.62)</td>
<td>0.96 (0.93, 0.97)</td>
<td>−0.30 (−0.80, 0.20)</td>
<td>3.47 (2.61, 4.32)</td>
</tr>
</tbody>
</table>

Comparison of the value of record 1 (M2) with the mean value of records 1, 2, and 3 (M1) stratified per assessor. All values are presented with corresponding 95% CIs. *SE that corresponds to mean difference between M1 and M2. †Lower and upper limits of agreement that correspond to SE ± 1.96 SDs of the SE. ‡PPDT at the second toe. §EPDT after single electrical stimulation. ||NWRT after single electrical stimulation.


