Scandinavian Journal of Pain 16 (2017) 93–98

Contents lists available at ScienceDirect

Scandinavian Journal of Pain

journal homepage: www.ScandinavianJournalPain.com

Clinical pain research

Multifactorial assessment of measurement errors affecting intraoral quantitative sensory testing reliability☆

Estephan J. Moana-Filho[a,*], Aurelio A. Alonso[b], Flavia P. Kapos[h,i], Vladimir Leon-Salazar[c], Scott H. Durand[d], James S. Hodges[e], Donald R. Nixdorf[a,f,g]

[a] Division of TMD and Orofacial Pain, School of Dentistry, University of Minnesota, 6-320d Moos Tower, 515 Delaware St. SE, Minneapolis, MN 55455, United States
[b] Center for Translational Pain Medicine, Department of Anesthesiology, Duke University School of Medicine, United States
[c] Division of Pediatric Dentistry, School of Dentistry, University of Minnesota, United States
[d] Private Dental Practice, 115 East Main Street, Wabasha, MN, 55981, United States
[e] Division of Biostatistics, School of Public Health, University of Minnesota, United States
[f] Department of Neurology, Medical School, University of Minnesota, United States
[g] HealthPartners Institute for Education and Research, United States
[h] Department of Epidemiology, School of Public Health, University of Washington, United States
[i] Department of Oral Health Sciences, School of Dentistry, University of Washington, United States

HIGHLIGHTS

- A comprehensive approach to assess multiple sources of intraoral QST variation is proposed.
- Most variability come from differences between participants and visits-within-participant.
- Comprehensive reliability appraisal aids in clinical decision-making and resources allocation.

ARTICLE INFO

Article history:
Received 3 November 2016
Received in revised form 11 March 2017
Accepted 29 March 2017
Available online 1 May 2017

Keywords:
Multisensory perception
Nervous system
Neuroscience/neurobiology
Oral diagnosis
Pain

ABSTRACT

Background and purpose (aims): Measurement error of intraoral quantitative sensory testing (QST) has been assessed using traditional methods for reliability, such as intraclass correlation coefficients (ICCs). Most studies reporting QST reliability focused on assessing one source of measurement error at a time, e.g., inter- or intra-examiner (test–retest) reliabilities and employed two examiners to test inter-examiner reliability. The present study used a complex design with multiple examiners with the aim of assessing the reliability of intraoral QST taking account of multiple sources of error simultaneously.
Methods: Four examiners of varied experience assessed 12 healthy participants in two visits separated by 48 h. Seven QST procedures to determine sensory thresholds were used: cold detection (CDT), warmth detection (WDT), cold pain (CPT), heat pain (HPT), mechanical detection (MDT), mechanical pain (MPT) and pressure pain (PPT). Mixed linear models were used to estimate variance components for reliability assessment; dependability coefficients were used to simulate alternative test scenarios.
Results: Most intraoral QST variability arose from differences between participants (8.8–30.5%), differences between visits within participant (4.6–52.8%), and error (13.3–28.3%). For QST procedures other than CDT and MDT, increasing the number of visits with a single examiner performing the procedures would lead to improved dependability (dependability coefficient ranges: single visit, four examiners = 0.12–0.54; four visits, single examiner = 0.27–0.68). A wide range of reliabilities for QST procedures, as measured by ICCs, was noted for inter- (0.39–0.80) and intra-examiner (0.10–0.62) variation.

DOI of refers to article: http://dx.doi.org/10.1016/j.sjpain.2017.04.066.
☆ *Disclosures*: Research reported in this publication was supported by the National Institutes of Health grants UL1-TR000114 and K12-RR23247. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
* Corresponding author.
E-mail addresses: moana004@umn.edu (E.J. Moana-Filho), aurelio.alonso@duke.edu (A.A. Alonso), kapos001@umn.edu (F.P. Kapos), leons002@umn.edu (V. Leon-Salazar), drdurand@wabashadentistry.com (S.H. Durand), hodge003@umn.edu (J.S. Hodges), nixdorf@umn.edu (D.R. Nixdorf).

http://dx.doi.org/10.1016/j.sjpain.2017.03.007

Conclusion: Reliability of sensory testing can be better assessed by measuring multiple sources of error simultaneously instead of focusing on one source at a time. In experimental settings, large numbers of participants are needed to obtain accurate estimates of treatment effects based on QST measurements. This is different from clinical use, where variation between persons (the person main effect) is not a concern because clinical measurements are done on a single person.
Implications: Future studies assessing sensory testing reliability in both clinical and experimental settings would benefit from routinely measuring multiple sources of error. The methods and results of this study can be used by clinical researchers to improve assessment of measurement error related to intraoral sensory testing. This should lead to improved resource allocation when designing studies that use intraoral quantitative sensory testing in clinical and experimental settings.

1. Introduction

Somatosensory system assessment is part of a clinical examination of a patient presenting with pain, including tests to assess various sensory functions [1]. Such evaluation of the orofacial region includes traditional procedures such as thermal/electrical pulp tests, tooth percussion, palpation, and anaesthetic blocks [2] as well as various thermal, mechanical, and chemical stimuli [3]. In clinical settings, these tests are qualitative and lack standardization regarding stimulus application and assessment of evoked sensations [1,3]. When performed in a systematic manner using strictly defined stimulus properties, these tests are called quantitative sensory testing (QST) [4,5]. A standard QST protocol has been developed by the German Research Network on Neuropathic pain (DFNS) [6], and further developed for intraoral use [7].

Measurement error assessment is important for sensory testing given the multiple sources of variation: variation in stimulus (delivery methods), between examiners (experience, dexterity), between participants (sensitivity, attention, previous experiences), or between multiple visits. Intraoral QST measurement error has been investigated in healthy participants [7] and patients with persistent intraoral pain [8]. These studies focused on two measures of reliability, intra- and inter-examiner, as previously assessed in other studies of QST reliability [9,10]. This approach only accounts for one source of variation at a time – examiner or visit – and thus does not identify or measure other factors, e.g., related to participants, interactions between factors, or random error. Recently studies have investigated multiple sources of variation for sensory testing [11,12]. Such a comprehensive approach can identify factors that, once addressed, can reduce variation and guide resource allocation for studies employing sensory testing and also help evaluate these tests' applicability in clinical practice [13,14].

Our aim was to assess multiple sources of variation in a battery of intraoral QST procedures to: (i) determine their main source(s) of variation; and (ii) evaluate the influence of the number of examiners and participant visits for QST measurements' dependability.

2. Methods

2.1. Participants

Healthy participants were recruited from the UMN community. Eligibility criteria were absence of bodily pains in the previous six months and no visible oral disease. Telephone or in-person screening was initially done, then a clinical evaluation determined participant eligibility.

2.2. Examiners

Four examiners with varied clinical experience performed the intraoral QST procedures: one faculty member, one post-doctoral fellow, one dental resident, and one dental student. The faculty examiner underwent a 2-day training session in the intraoral QST protocol at the University of Washington. He then conducted a 1-day training session for the other three examiners, after which all four examiners practiced the procedures together on two further occasions.

2.3. Study design

The intraoral QST protocol was based on the DFNS adapted for intraoral use [7], retaining seven of the 13 original procedures due to time constraints and limited available resources. It included procedures measuring thresholds for thermal (cold detection [CDT], warmth detection [WDT], cold pain [CPT], and heat pain [HPT]) and mechanical (mechanical detection [MDT], mechanical pain [MPT], pressure pain [PPT]) sensory functions. Sensory testing was performed in four intraoral sites, one over the buccal premolar gingival mucosa in each quadrant. Thermal tests were performed in 2 quadrants, which were selected randomly in each participant for each thermal test done by each examiner; mechanical tests were performed in all quadrants.

Each participant was measured on two visits separated by 48 h, with each visit lasting a half-day. Before each session, all examiners convened to review the protocol. Separate dental operatory stations were used for these procedures: (1) PPT, (2) MDT, (3) MPT, and (4) thermal. Each participant remained seated in a given station and received that station's procedure(s) from each examiner, then moved to the next station to be examined by each examiner with that station's assigned procedure(s), until all seven procedures were performed on each participant by all four examiners.

2.4. QST procedures

2.4.1. Thermal testing

PATHWAY Pain & Sensory Evaluation System (Medoc, Israel) with an intraoral thermode having a round active contact surface (diameter = 6 mm) was used for all thermal tests, which were performed in the sequence: CDT-WDT-CPT-HPT. For each test, the intraoral thermode was held in place by the examiner, with a baseline temperature of 32 °C, and temperature change rate of 1 °C/s for CDT and WDT; for CPT and HPT, the rate of temperature change from baseline was 1.5 °C/s; the rate of return to baseline was 8 °C/s. Cut-off temperatures for thermal tests were 0 °C and 54 °C. Participants were instructed to hold a response unit and press its button once a particular sensation (coolness, warmth, cold pain, heat pain) was first perceived, ending the trial. Detection thresholds were calculated as the temperature difference from baseline; pain thresholds were determined from the absolute temperature reached. Each test included three measurements; the average of the three measurements was used as threshold.

2.4.2. Mechanical testing

2.4.2.1. Mechanical detection threshold. Thirteen modified von Frey monofilaments (OptiHair$_2$, MARSTOCK nervtest, Germany) were used to determine the MDT. The force range was 0.125–512 mN, with each successive filament increasing force by a multiple of 2. Threshold measurement used an adaptation of the method of limits described previously [15], obtaining three infra- and three supra-thresholds. The geometric mean of these six values was used as the MDT.

2.4.2.2. Mechanical pain threshold. MPT was measured using a custom-made weighted set of eight calibrated pinprick instruments with a flat contact surface of 0.2 mm diameter (4–512 mN force range, factor 2 progression). Each instrument was applied perpendicular to the intraoral sites, with a contact time of approximately 2 s. The method of limits was used to determine six values, with their geometric mean used as the MPT.

2.4.2.3. Pressure pain threshold. PPT was measured using a digital pressure algometer (SOMEDIC, Sweden) fitted with a probe (surface area: 0.18 cm^2, diameter: 4.8 mm). Participants held a switch connected to the algometer and were instructed to press it at the first painful sensation. After placing the probe tip over the gingiva, pressure was increased at a rate of 50 kPa/s until the participant interrupted the stimulus. The average of three trials was used as the PPT.

2.5. Data analysis

Analyses used mixed linear models with the following variance components:

Main effects:

1. *Bona fide* differences between participants in threshold ("true scores") (σ^2_p);
2. Differences between a participant's quadrants ($\sigma^2_{q[p]}$);
3. Examiner differences in administering a test (σ^2_e), e.g., knowledge, skill, experience, and biases;
4. Differences between visits (σ^2_v);

Interactions:

5. "Examiner-by-participant" ($\sigma^2_{e,p}$), variation between examiners in their measured differences between participants;
6. "Visit-by-participant" ($\sigma^2_{v,p}$), variation between visits in a participant's measurements;
7. "Examiner-by-visit" ($\sigma^2_{e,v}$), variation between visits in an examiner's measurements;
8. "Examiner-by-quadrant within participant" ($\sigma^2_{e,q[p]}$), variation between examiners in the differences between quadrants that they measure;
9. "Visit-by-quadrant within participant" ($\sigma^2_{v,q[p]}$); variation between visits in differences between quadrants;
10. "Examiner-by-visit-by-participant" ($\sigma^2_{e,v,p}$); possible influencing factors similar to items 5, 6 and 7 taken in conjunction;

And finally:

11. Residual error (σ^2_{Resid}), which includes both the interaction "examiner-by-visit-by-quadrant within participant" and random error.

For these mixed linear models, the dependent variables were the test measurements on their raw scales except for MDT and MPT, for which the common logarithms (log to base 10) were used as dependent variables. Dependability coefficients for alternative settings (different number of examiners or visits) were also computed for each test (see Supplementary file). These coefficients simulate alternative test settings beyond those used in our study, and allow us to estimate how adding more visits would impact the variability of test results. Simple summaries for all seven measures (average, 2.5 and 97.5 percentiles) are presented for descriptive purposes, ignoring the complex study design (i.e., multiple examiners, visits, etc.).

Intraclass correlation coefficients (ICC) describing inter- and intra-examiner (test–retest) reliability were calculated for comparison to previous reliability studies. Study design complexity was taken into account by adapting Shrout and Fleiss's "ICC (2,1)" model [16] (see Supplemental material). Inter- and intra-examiner levels of agreement were deemed poor (ICC < 0.40), fair ($0.40 \leq$ ICC ≤ 0.59), good ($0.60 \leq$ ICC ≤ 0.74), or excellent (ICC ≥ 0.75) using published guidelines [17].

All analyses were implemented using SAS (v. 9.3, SAS Institute Inc., Cary, NC) and the R system v.3.3.0 [18].

Table 1
Means for QST procedures.

Procedure	Mean	Percentiles 2.5–97.5
CDT[a] (°C)	9.7	3.1–23.1
WDT[a] (°C)	9.1	3.3–13.9
CPT[b] (°C)	11.5	0.0–23.9
HPT[b] (°C)	44.7	39.0–50.0
MDT[c] (mN)	5.5	0.2–32.0
MPT[c] (mN)	148.6	12.0–406.0
PPT (kPa)	219.0	78.0–454.0

[a] Difference from baseline temperature (32 °C).
[b] Absolute thresholds.
[c] Geometric means.

Table 2
Intraclass correlation coefficients (ICC) for QST procedures.

Procedure	Inter-examiner	Intra-examiner (test–retest)
CDT	0.39	0.62
WDT	0.66	0.10
CPT	0.80	0.34
HPT	0.63	0.42
MDT	0.39	0.52
MPT	0.60	0.27
PPT	0.53	0.37

Study design complexity accounted for by adapting the "ICC 2,1" model proposed by Shrout and Fleiss (1979) – see supplemental material for details.
Levels of agreement: poor (ICC < 0.40), fair ($0.40 \leq$ ICC ≤ 0.59), good ($0.60 \leq$ ICC ≤ 0.74), and excellent (ICC ≥ 0.75).

3. Results

Twelve participants (10 females, average age ± SD: 39.3 ± 11.9 years) participated in the first visit; 8 returned for the second visit (6 females, age: 39.8 ± 10.7), as the other participants had time constraints preventing their return for a second visit.

Table 1 lists simple summaries of the QST measures; Table 2 shows intra- and inter-examiner ICCs. Inter-examiner reliability was poor for CDT and MDT, fair for PPT, good for WDT, HPT, and MPT, and excellent for CPT. Intra-examiner (test–retest) ICCs had lower reliability, mostly poor (WDT, CPT, MPT, PPT) and fair (HPT, MDT).

Table 3 presents variance component estimates for each procedure both as absolute values and as percentages of total variance. For each procedure, most variance (>75%) arises from the participant main effect (σ^2_p), differences between visits within participant ($\sigma^2_{v,p}$), and residual error (σ^2_{Resid}). Also, CDT had large variation between examiners in the difference between quadrants ($\sigma^2_{e,q[p]}$); MDT had large variation between quadrants and between examiners ($\sigma^2_{q[p]}$, σ^2_e); and PPT had a large examiner-by-visit-by-participant interaction ($\sigma^2_{e,v,p}$). The participant main effect

Table 3
Variance component estimates for each QST procedure.

Variance component	CDT	WDT	CPT	HPT
σ^2_{p}	4.42 (20.6)	0.70 (8.8)	19.80 (30.5)	2.55 (33.7)
$\sigma^2_{q[p]}$	0.42 (2.0)	0.00 (0.0)	2.49 (3.8)	0.01 (0.1)
σ^2_{e}	0.19 (0.9)	0.00 (0.0)	0.00 (0.0)	0.00 (0.0)
σ^2_{v}	0.07 (0.3)	0.00 (0.0)	0.56 (0.8)	0.00 (0.0)
$\sigma^2_{e,p}$	2.42 (11.3)	0.13 (1.6)	0.00 (0.0)	0.26 (3.4)
$\sigma^2_{v,p}$	3.39 (15.8)	4.19 (52.8)	29.21 (45.0)	1.86 (24.5)
$\sigma^2_{e,v}$	0.00 (0.0)	0.00 (0.0)	0.00 (0.0)	0.08 (1.1)
$\sigma^2_{e,q[p]}$	5.92 (27.7)	0.00 (0.0)	0.00 (0.0)	0.33 (4.3)
$\sigma^2_{v,q[p]}$	0.00 (0.0)	0.37 (4.7)	0.00 (0.0)	0.35 (4.6)
$\sigma^2_{e,v,p}$	0.00 (0.0)	0.72 (9.1)	0.54 (0.8)	0.00 (0.0)
σ^2_{Resid}	4.58 (21.4)	1.83 (23.0)	12.36 (19.0)	2.15 (28.3)
σ^2_{Total}	21.42 (100.0)	7.93 (100.0)	64.94 (100.0)	7.58 (100.0)

Variance component	MDT[a]	MPT[a]	PPT
σ^2_{p}	0.07 (18.4)	0.03 (13.7)	2418.11 (30.0)
$\sigma^2_{q[p]}$	0.04 (11.5)	0.01 (4.0)	259.32 (3.2)
σ^2_{e}	0.08 (20.6)	0.00 (1.4)	421.13 (5.2)
σ^2_{v}	0.01 (2.4)	0.01 (2.8)	88.74 (1.1)
$\sigma^2_{e,p}$	0.02 (5.8)	0.02 (8.7)	0.00 (0.0)
$\sigma^2_{v,p}$	0.02 (4.6)	0.08 (41.0)	877.09 (10.9)
$\sigma^2_{e,v}$	0.01 (1.4)	0.00 (0.0)	0.00 (0.0)
$\sigma^2_{e,q[p]}$	0.02 (6.0)	0.00 (0.0)	126.06 (1.6)
$\sigma^2_{v,q[p]}$	0.01 (4.0)	0.00 (0.0)	670.65 (8.3)
$\sigma^2_{e,v,p}$	0.00 (0.7)	0.01 (4.5)	2117.68 (26.3)
σ^2_{Resid}	0.09 (24.7)	0.05 (23.8)	1074.26 (13.3)
σ^2_{Total}	0.37 (100.0)	0.20 (100.0)	8053.04 (100.0)

Variance (%).
[a] Log transformed.

(variation between their "true scores") accounted for 8.8–33.7% of total variance, depending on the test. As examples, the largest components of variance for CPT were the participant main effect (30.5%) and differences between visits within participant (45%), while for MDT the largest components were the residual (24.7%) and the main effects for examiners (20.6%), participants (18.4%), and quadrants (11.5%).

Table 4 shows dependability coefficients for alternative test scenarios. As expected, the worst (1 visit, 1 examiner) and best (4 visits, 4 examiners) scenarios gave a wide range of dependability coefficients, for example 0.23–0.60 for CDT. Except for CDT and MDT, having more participant visits with a single examiner improved the procedure's dependability coefficient more than having a single visit with 4 examiners. Other test settings can be simulated using the "variance components estimates calculator" available as a supplementary file.

4. Discussion

The present study used a complex design with four examiners having different training levels, examining each participant in two visits, allowing estimation of several variance components for each of seven QST procedures. This allowed detailed examination of multiple error sources that contribute to the procedures' performance. This approach departs from the usual practice of focusing on one source of error such as variability between examiners (inter-examiner) or between visits by the same examiner (intra-examiner), which others have previously noted [11].

Average intraoral QST measures reported here (Table 1) are slightly lower but comparable to previous studies [7,19], especially considering the 2.5th and 97.5th percentiles. We calculated ICCs to compare our results to those studies, though as noted our calculations needed elaboration to accommodate our study design and use the whole dataset. To the best of our knowledge, this is the first time ICCs have incorporated such design complexity; by doing so we avoided reporting 56 ICCs (4 examiners × 2 visits × 7 procedures) based on subsets of the data, and also obtained more stable estimates.

Despite the aforementioned complexity, all reliability calculations using ICCs follow the same basic formula [20]:

$$\text{reliability} = \frac{\text{between subjects variability (signal)}}{\text{between subjects variability (signal)} + \text{error (noise)}}$$

From this, it follows that an ICC ranges between 0 (no reliability) to 1 (perfect reliability), and its value depends on the magnitude of "signal" relative to "signal + noise". The mathematical model rationalizing these ICCs in the present study included in "signal" the variance components for the main effects for "participant" and "quadrant"; in addition to these, the inter-examiner ICC included in "signal" the interactions "visit-by-participant" and "visit-by-quadrant within participant", while the intra-examiner (test–retest) ICC included "examiner-by-participant" and "examiner-by-quadrant" (see supplemental material). This aids understanding of Table 2's ICC values. Consider CPT for example, with inter- and intra-examiner ICCs of 0.80 and 0.34 respectively. In Table 3's variance component estimates for CPT, "visit-by-participant" contributes 45% of the total variance, which means that almost half of the variability for CPT arises from variation between participants in the change in their thresholds between the first and second visits. Because this component is considered "signal" for inter-examiner ICC but "noise" for intra-examiner ICC, it becomes clear why the former ICC is much larger than the latter.

Another feature of ICCs is that low values can arise either from large measurement error (high noise) from any of several sources (methods/instruments, examiners, visits) or from reduced between-subject variability (low signal), i.e., subjects who "look alike" when measured. This does not mean that a measurement method with low ICC has no use, but only that is has limited ability to discriminate the specific collection of subjects that was assessed. Such a method could be valuable for assessing a given patient over time if its measurement error is small; in this case, between-subject variability has no influence on the method's performance

Table 4
Dependability coefficients for QST procedures.

N_{visits}	$N_{examiners}$							
	CDT				WDT			
	1	2	3	4	1	2	3	4
1	0.23	0.33	0.38	0.42	0.09	0.11	0.11	0.12
2	0.28	0.40	0.48	0.52	0.16	0.19	0.20	0.21
3	0.30	0.44	0.52	0.57	0.22	0.26	0.27	0.28
4	0.31	0.46	0.54	0.60	0.27	0.31	0.33	0.34

N_{visits}	$N_{examiners}$							
	CPT				HPT			
	1	2	3	4	1	2	3	4
1	0.34	0.38	0.40	0.40	0.34	0.41	0.45	0.47
2	0.51	0.55	0.57	0.57	0.48	0.57	0.61	0.63
3	0.61	0.65	0.66	0.67	0.55	0.65	0.68	0.71
4	0.68	0.71	0.72	0.73	0.60	0.70	0.73	0.75

N_{visits}	$N_{examiners}$							
	MDT[a]				MPT[a]			
	1	2	3	4	1	2	3	4
1	0.30	0.42	0.49	0.54	0.18	0.22	0.24	0.25
2	0.37	0.51	0.59	0.64	0.28	0.34	0.37	0.39
3	0.40	0.55	0.63	0.68	0.34	0.42	0.45	0.47
4	0.42	0.57	0.65	0.71	0.38	0.47	0.51	0.54

N_{visits}	$N_{examiners}$ PPT			
	1	2	3	4
1	0.33	0.43	0.48	0.51
2	0.47	0.59	0.64	0.66
3	0.55	0.66	0.71	0.74
4	0.60	0.71	0.76	0.78

Dependability coefficients range: 0 (poor) – 1 (high).
[a] Log transformed.

[21]. The above formula also implies that ICC as a measure of reliability reflects a measurement method's performance for a given population sample with a given heterogeneity (context specific); when comparing a method's ICCs as measured in studies that used different samples, one needs to assess whether the samples are similarly heterogeneous, for example by evaluating the between- and within-subject standard deviations [22].

Reliability calculations focusing on one error source at a time, such as ICCs, are considered part of a framework known as "classic test theory". This has been recognized as an important limitation; an extension of this framework, called generalizability theory (GT) [23,24], was proposed so that multiple sources of measurement error could be recognized and estimated. The methods used here can be considered part of this trend, providing several benefits. First, variance components estimation allows not only calculation of reliability measures such as ICCs (how well can participants be distinguished from each other, despite measurement error), but also other aspects of measurement error such as "agreement" (how close are two measurements on the same participant). Agreement is expressed in the same units as the measurement itself and unlike ICC does not depend on the sample's heterogeneity, thus quantifying measurement error only [20–22]. Second, no consensus exists on how reproducibility of QST results should be defined or assessed [9,25,26]. Finally, in a carefully designed study (e.g., number of participants, examiners, and visits), variance components estimation is relatively straightforward and would also allow calculation of dependability coefficients, which can be used to simulate alternative test scenarios to improve allocation of resources in future studies using QST procedures for sensory testing to obtain accurate estimates of treatment effects.

Table 3 shows that most variability in QST procedures arises from three variance components – differences between participants ("true scores"), differences between visits within participant, and error – but there were exceptions. For CDT, 27.7% of variation was variation between examiners in their measured differences between intraoral quadrants. One explanation is that CDT was the first thermal procedure done in all subjects, and placement of the intraoral thermal probe could elicit discomfort, thus impacting overall CDT variability. About a third of MDT's variation was variation between participants in differences between quadrants and differences between examiners, which could be related, respectively, to difficulties in positioning the Von Frey filaments and in examiners' experience in using the filaments. Other authors have found poor reliability for MDT, attributing it to the filaments' design and the method of limits used for threshold determination [7]. The interaction "examiner-by-visit-by-participant" represented 26.3% of PPT total variance, meaning that within-participant differences between visits also varied considerably between participants and examiners. This could be a consequence of the examiners' varying ability in assessing PPT across visits. Based on these findings and after addressing concerns specific to CDT, MDT, and PPT, one potential way to improve these procedures' reliability is to perform them on multiple visits. By averaging measurements from all visits, the error attributed to within-participant variability between visits can be reduced [11], allowing true differences between participants to explain a greater portion of the remaining variability of each procedure, i.e., the measurement would become more reliable.

Studies using the GT framework would provide information needed to serve the different needs of clinical decision making and clinical research: when examining a patient, the between-patient component of variation does not contribute to error in determining that patient's status (diseased vs. non-diseased), but when comparing treatments given to distinct groups of study participants, that component of variation will largely determine the statistical power for a given sample size, as the present study suggests. By knowing the several sources of error in each test's performance, improved allocation of limited resources (clinician's chair time, costs associated with procedures, study participant sample size) will ultimately lead to improved diagnosis in clinical settings and more accurate estimates of treatment effects in experimental settings. Dependability coefficients calculation for alternative test scenarios simulation could help researchers to determine which design features would improve the tests' reliability.

This study has strengths and limitations. Our sample size was relatively small; the mathematical models allowed calculations with missing data but a larger sample would give more stable estimates. A shortened version of the intraoral QST protocol was used [7], with seven out of the 13 procedures originally described and MDT and MPT modified to reduce threshold measurements from 10 to six. These modifications were needed to save time four examiners were included instead of two, as in most studies [7–11]. An additional strength was accounting for multiple sources of error at the same time including testing in different intraoral quadrants, which allows better understanding of factors affecting intraoral QST reliability.

5. Conclusions

To the best of our knowledge, this study was the first to account for multiple sources of measurement error in intraoral QST simultaneously. This allows an improved understanding of the moderate to poor reliability for these tests measured by ICC for intra- and inter-rater scenarios. Using the GT framework to determine the reliability of intraoral sensory tests can help elucidate sources of error to inform clinical decision-making and resources allocation in experimental settings.

Supplemental material

1. Rationale for ICC and dependability coefficients calculation and Dependability Coefficients.
2. Variance components estimates and Dependability Coefficients calculator.

Ethical issues

The University of Minnesota (UMN) Institutional Review Board approved the study's protocol (approval #1004M80212) and all participants gave oral and written informed consent before entering the study. Data were collected in May 2010.

Conflict of interest

The authors declare that there is no conflict of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.sjpain.2017.03.007.

References

[1] Hansson P, Backonja M, Bouhassira D. Usefulness and limitations of quantitative sensory testing: clinical and research application in neuropathic pain states. Pain 2007;129:256–9.
[2] Hargreaves KM, Berman LH. Cohen's Pathways of the Pulp Expert Consult. 11th ed. Mosby; 2015.
[3] Svensson P, Baad-Hansen L, Pigg M, List T, Eliav E, Ettlin D, Michelotti A, Tsukiyama Y, Matsuka Y, Jaaskelainen SK, Essick G, Greenspan JD, Drangsholt M. Guidelines and recommendations for assessment of somatosensory function in oro-facial pain conditions – a taskforce report. J Oral Rehabil 2011;38:366–94.
[4] Cruz-Almeida Y, Fillingim RB. Can quantitative sensory testing move us closer to mechanism-based pain management? Pain Med 2014;15:61–72.
[5] Gruener G, Dyck PJ. Quantitative sensory testing: methodology, applications, and future directions. J Clin Neurophysiol 1994;11:568–83.
[6] Rolke R, Baron R, Maier C, Tolle TR, Treede RD, Beyer A, Binder A, Birbaumer N, Birklein F, Botefur IC, Braune S, Flor H, Huge V, Klug R, Landwehrmeyer GB, Magerl W, Maihofner C, Rolko C, Schaub C, Scherens A, Sprenger T, Valet M, Wasserka B. Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): standardized protocol and reference values. Pain 2006;123:231–43.
[7] Pigg M, Baad-Hansen L, Svensson P, Drangsholt M, List T. Reliability of intraoral quantitative sensory testing (QST). Pain 2010;148:220–6.
[8] Baad-Hansen L, Pigg M, Yang G, List T, Svensson P, Drangsholt M. Reliability of intra-oral quantitative sensory testing (QST) in patients with atypical odontalgia and healthy controls – a multicentre study. J Oral Rehabil 2015;42:127–35.
[9] Geber C, Klein T, Azad S, Birklein F, Gierthmuhlen J, Huge V, Lauchart M, Nitzsche D, Stengel M, Valet M, Baron R, Maier C, Tolle T, Treede RD. Test–retest and inter-observer reliability of quantitative sensory testing according to the protocol of the German Research Network on Neuropathic Pain (DFNS): a multi-centre study. Pain 2011;152:548–56.
[10] Moloney NA, Hall TM, O'Sullivan TC, Doody CM. Reliability of thermal quantitative sensory testing of the hand in a cohort of young, healthy adults. Muscle Nerve 2011;44:547–52.
[11] O'Neill S, O'Neill L. Improving QST reliability – more raters, tests, or occasions? A multivariate generalizability study. J Pain 2015;16:454–62.
[12] Pryseley A, Ledent EY, Drewes AM, Staahl C, Olesen AE, Arendt-Nielsen L. Applying concepts of generalizability theory on data from experimental pain studies to investigate reliability. Basic Clin Pharmacol Toxicol 2009;105:105–12.
[13] Backonja MM, Attal N, Baron R, Bouhassira D, Drangholt M, Dyck PJ, Edwards RR, Freeman R, Gracely R, Haanpaa MH, Hansson P, Hatem SM, Krumova EK, Jensen TS, Maier C, Mick G, Rice AS, Rolke R, Treede RD, Serra J, Toelle T, Tugnoli V, Walk D, Walalce MS, Ware M, Yarnitsky D, Ziegler D. Value of quantitative sensory testing in neurological and pain disorders: NeuPSIG consensus. Pain 2013;154:1807–19.
[14] Birklein F, Sommer C. Pain: quantitative sensory testing – a tool for daily practice? Nat Rev Neurol 2013;9:490–2.
[15] Baumgartner U, Magerl W, Klein T, Hopf HC, Treede RD. Neurogenic hyperalgesia versus painful hypoalgesia: two distinct mechanisms of neuropathic pain. Pain 2002;96:141–51.
[16] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.
[17] Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284.
[18] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.
[19] Baad-Hansen L, Pigg M, Ivanovic SE, Faris H, List T, Drangsholt M, Svensson P. Intraoral somatosensory abnormalities in patients with atypical odontalgia – a controlled multicenter quantitative sensory testing study. Pain 2013;154:1287–94.
[20] Weir JP. Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. J Strength Cond Res 2005;19:231–40.
[21] de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006;59:1033–9.
[22] Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008;31:466–75.
[23] Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: a liberalization of reliability theory. Br J Stat Psychol 1963;16:137–63.
[24] Shavelson RJ, Webb NM, Rowley GL. Generalizability theory. Am Psychol 1989;44:922–32.
[25] Shy ME, Frohman EM, So YT, Arezzo JC, Cornblath DR, Giuliani MJ, Kincaid JC, Ochoa JL, Parry GJ, Weimer LH, Therapeutics, Technology Assessment Subcommittee of the American Academy of N. Quantitative sensory testing: report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. Neurology 2003;60:898–904.
[26] Werner MU, Petersen MA, Bischoff JM. Test–retest studies in quantitative sensory testing: a critical review. Acta Anaesthesiol Scand 2013;57:957–63.