Improving QST Reliability—More Raters, Tests, or Occasions?
A Multivariate Generalizability Study

Søren O’Neill* and Lotte O’Neill

*Spine Center of Southern Denmark, Lillebælt Hospital, Middelfart, Denmark.
†Institute of Regional Health Science Research, University of Southern Denmark, Odense, Denmark.
‡Centre for Medical Education, Aarhus University, INCUBA Science Park Skejby, Arhus N, Denmark.

Abstract: The reliability of quantitative sensory testing (QST) is affected by the error attributable to both test occasion and rater (examiner) and the interactions between them. Most reliability studies account for only 1 source of error. The present study employed a fully crossed, multivariate generalizability design to account for rater and occasion variance simultaneously. Nineteen healthy volunteers were examined with a battery of 7 QST procedures 4 times on 2 occasions by 2 raters. The QST battery was composed to include a mix of different pain stimuli and response domains, including threshold, intensity, tolerance, and modulation with mechanical, thermal, and chemical stimuli. The classical test-retest and interrater reliability (.19 < intraclass correlation coefficient <.92) was in line with the literature, and generalizability analysis indicated that the universe score was generally the dominant source of variation (relative contribution = 19%, 78%). Error attributable to the interaction between study participant and occasion was also influential. Dependability coefficients indicated that a substantial increase in reliability and feasibility could be achieved by employing a composite QST battery compared to single QST procedures. Reliability was improved more by repeated testing on separate occasions than by repeated testing by different raters. Perspectives: When balancing reliability and feasibility, the current findings suggest that a carefully selected battery of QST procedures repeated on a few occasions may be optimal.

© 2015 by the American Pain Society
Key words: Pain, quantitative sensory testing, reliability, validity, generalizability.
good and excellent reliability are common. The reliability is often reported as intraclass correlation coefficients (ICCs) model 2.1 and are typically found to be in the range of .6 to .9.\textsuperscript{16} Reports of the reliability of QST with intramuscular injection of hypertonic saline are fewer, but ICCs between .62 and .78 have been reported.\textsuperscript{16}

Studies such as these are typically conducted as either an interexaminer reliability study or a test-retest (intraexaminer) reliability study: That is, only one source of error variance (rater or occasion) is accounted for, though both effects are likely to affect the variance in scores simultaneously. However, with generalizability theory (GT), an extension of classical test theory (CTT), it is possible to estimate less biased reliability (generalizability) coefficients, which take into account several sources of error variance simultaneously.

The only other published GT study of QST we are aware of did not involve multiple raters and included pain measurements with or without the administration of pain-modulating drugs.\textsuperscript{22}

It was not the purpose of this study to evaluate or recommend a particular set of quantitative sensory tests as a standard for clinical application, nor were the actual reliability coefficients the main interest in this study of healthy volunteers. The aim of this paper was to examine the generalizability of diverse pain measures, taking into account the simultaneous effect of rater and occasion in a test situation. Thus, the present study's main contribution is to disentangle different sources of error variance at work simultaneously.

The objectives were to examine the estimated variance components for individual QSTs, the generalizability of individual QSTs in alternative test situations, and the composite generalizability of a test battery in alternative test situations.

Methods

Participants, Raters, and Occasions

Participants

Pain-free adult participants were recruited for the study. The participants (7 women, 12 men) were fifth-year university students who were recruited during their clinical internship at the Spine Center of Southern Denmark, Lillebælt Hospital, Denmark.

Raters

The nurses, medical doctors, and chiropractors of the department were invited to volunteer as raters for the study. Eight clinicians responded to the invitation, and of these, 2 raters were chosen randomly.

Both raters, 1 female and 1 male, were experienced clinicians (10 and 19 years, respectively). Both were familiar with performing physical and complex technical examination procedures such as diagnostic ultrasound, albeit not with QST.

Prior to data collection, the raters were instructed in the involved procedures on 2 separate occasions and allowed time to practice on each other. The principal author (S.O.) supervised the QST procedures during the practice sessions and repeatedly throughout the data collection to ensure correct procedure.

Occasions and QST Sessions

From a period of 1 month (between September 23 and October 22, 2013), 11 logistically opportune dates were chosen. Of these, 2 examination occasions (separate days) were chosen randomly. The 2 random occasions were 7 days apart.

QST was performed twice on each occasion, for a total of 4 QST sessions for each participant. On each occasion, the 2 QST sessions were spaced by approximately 2 hours for each participant.

To avoid or minimize a spillover effect from the first to the second QST session (particularly in relation to intramuscular saline injection), a QST examination schedule was prepared in advance to ensure that test order (first and second QST sessions) and test side (left and right) were alternated between raters and between occasions. For example, on occasion 1, participant X was tested by rater A on the right-hand side in the first QST session and by rater B on the left-hand side in the second QST session. On occasion 2, participant X was tested by rater B on the right-hand side in the first QST session and by rater A on the left-hand side in the second QST session. QST session (first/second) and test side (left/right) were also evenly distributed between the 2 raters. Test side and QST session order were thus evenly distributed across raters and occasions to the greatest possible extent.

Thus, 19 healthy, pain-free participants ($n_p = 19$) were tested by the same 2 raters ($n_r = 2$) who administered the 7 pain sensitivity tests ($n_t = 7$) on 2 different occasions ($n_o = 2$).

Participants, raters, and occasions could be considered as random samples of the universes of admissible observations as described in GT.\textsuperscript{3}

QST

A battery of 7 quantitative sensory pain tests were performed in the following order:

- Mechanical pressure pain detection threshold (PPDT)
- Sustained mechanical pressure on the thumb (ie, pressure pain intensity [PPI])
- Cold pressor test (CPT) (tolerance), duration of pain (CPTD)
- CPT, pain summated (CPTS)
- Conditioned pain modulation (CPM)
- Saline-induced pain duration (SPD)
- Saline-induced pain summated (SPS)

The QST battery was not constructed with any particular clinical application in mind or as being indicative of any particular aberration in pain sensitivity. Rather, the battery was composed with the aim of covering a reasonably broad and diverse spectrum of stimulation modalities and pain domains. The QST pain tests are described in detail below.

Participants were instructed in the use of visual analog scales (VASs), which were 100-mm scales on paper,
marked “no pain” at one end and “worst pain imaginable” at the other, or similar scales displayed on a computer screen (continuous sampling at 1 Hz). The participants were positioned comfortably (seated) for the QST examination (Appendix).

Mechanical PPDT

PPDTs at the tibialis anterior muscle were measured using a pressure algometer (with a 1-cm² probe, model 2; Somedic, Horby, Sweden). Pressure was applied manually with a near-constant velocity of approximately 50 kPa/s until the participant indicated that the pressure was becoming painful by pressing an indicator button connected to the algometer. PPT measurements were repeated 3 times, with approximately 10-second rest intervals. The head of the algometer was placed such that repeat applications overlapped partly. If no pain had been elicited by 1,000 kPa, this was recorded as the PPDT. If the first and second measurements were 1,000 kPa, a third was not performed.

Mechanical PPI

A simple spring-operated tool clamp (100 mm, with two 14 × 13 mm pressure pads, product no. 72644; Millarco, Lystrup, Denmark) was used to apply sustained mechanical pressure on the thumbnail for 10 seconds.17 The clamp exerted a pressure of approximately 5 kg at a 7-mm opening. The spring clamp was placed in such a way that the upper pressure pad was placed as far proximal on the nail as possible, without overlapping the eponychium. The lower pad was placed proximal enough to prevent the clamp from sliding forward.

CPTD and CPTS

A 25-L water tub (Mobicool C40; Dometic WAECO, Dubai, United Arab Emirates) was kept refrigerated at 0 to 2°C, and the temperature was monitored throughout with a mercury thermometer.

Participants were instructed to submerge their nonclenched hand up to the wrist in the circulating water (submersible pump, 10 L/min, .5 bar; Reich, Arnhem, The Netherlands) and keep it there for 1 minute or until the pain became unbearable. Participants rated the cold pain on a computerized VAS (“no pain” = 0 to “worst pain imaginable” = 100) sampled at a frequency of 1 Hz. Cold pain was summarized as duration and pain summation (area under the curve) and recorded as SPD and SPS, respectively.

Injection of Hypertonic Saline (SPD/SPS)

After appropriate skin disinfection, .5 mL of sterile, hypertonic saline (at room temperature) was injected into the tibialis anterior muscle. The saline was injected gradually at an even rate over approximately 1 second. Participants indicated the pain intensity on a computerized VAS (“no pain” = 0 to “worst pain imaginable” = 100, sampled at 1 Hz).

Saline pain was summarized as duration and pain summation (area under the curve) and recorded as SPD and SPS, respectively.

Analysis

The test situation described above corresponds to the multivariate generalizability design \( p^* \times r^* \times o^* \) (or patients crossed with raters crossed with occasions) as described in GT by Brennan.3(p207) If \( \phi \) is the variance, the total variance in scores for each of the pain measures (t) can best be described with the linear model in equation 1.

\[
X(\text{pro}) = \mu + \phi(p) + \phi(r) + \phi(o) + \phi(pr) + \phi(po) \\
+ \phi(ro) + \phi(pro)
\]

(Equation 1)

Fig 1 is a visual representation of the variance components that were disentangled for each of the 7 pain measures in this study, and Table 1 gives a verbal explanation of these.

Based on GT, a dependability (\( \phi \)) coefficient for each pain measure, for a test situation with 1 rater on 1 occasion, was calculated using the general equation (equation 2):

\[
\phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_r^2 + \sigma_o^2 + \sigma_{pr}^2 + \sigma_{po}^2 + \sigma_{ro}^2 + \sigma_{pro,e}^2}
\]

(Equation 2)

where \( \sigma^2 \) is the variance and \( p, r, o, \) and \( pro,e \) are the involved effects giving rise to variance in scores. The involved effects giving rise to variance in scores.

Figure 1. Venn diagram of the variance components estimated for each of the subjects. The intersections between \( p, r, \) and \( o \) effects represent their interactions, and \( pro,e \) represents the interaction of all 3 components and error.
study, with equation 4 derived from equation 2 above:

\[ R = \frac{\sigma_r^2}{\sigma_p^2 + \sigma_r^2} \quad \text{(Equation 3)} \]

where \( \tau \) refers to true score and \( \varepsilon \) to error. Commonly used reliability measures such as ICC and kappa coefficients also derive from this basic form (equation 3). GT is merely an extension of CTT, which allows for the error variance \( \sigma_e^2 \) to be decomposed into all relevant systematic sources of error actually at work—simultaneously—in the test situation. Therefore, GT most often allows for more valid descriptions of the majority of real-life test situations. In CTT, either an interrater reliability coefficient or a test-retest reliability coefficient is typically estimated. However, as both facets (rater and occasion) actually affect the real test situation simultaneously, reliability coefficients that include only one or the other effect (like ICCs and kappas) will by design be biased measures of the actual reliability of the test situation. In fact, from a GT perspective, ICCs and biased measures of the actual reliability of the test other effect (like ICCs and kappas) will by design be reliability coefficients that include only one or the other effect (like ICCs and kappas) will by design be biased measures of the actual reliability of the test situation. Different participants may respond differently to pain on different occasions. Other effect (like ICCs and kappas) will by design be reliability coefficients that include only one or the other effect (like ICCs and kappas) will by design be biased measures of the actual reliability of the test situation. Different participants may respond differently to pain on different occasions.

\[ \sigma_{ro}^2 \]

The variance in scores attributable to the interaction between the rater and the occasion (day) the measurement took place. Different raters may perform the tests differently on different days.

\[ \sigma_{po}^2 \]

The residual, which includes interaction between all effects (\( p, r, \) and o) as well as random error (\( e \)).

### Analytical Software

mGENOVA for PC (Robert L. Brennan, Iowa Testing Programs, University of Iowa, Iowa City, IA) was used to estimate the variance-covariance matrices involved in this multivariate generalizability design. Based on the estimated variance-covariance matrices, a series of D studies was subsequently performed with mGENOVA, in which dependability coefficients for test situations with alternative numbers of raters and occasions were calculated for each pain measure with equation 4.

mGENOVA was also used to calculate composite generalizability coefficients for alternative “test batteries.” This feature allows the 7 pain measures to be combined and weighted differently, so that the researcher may examine optimal composite generalizability while also considering the feasibility of different potential test situations.

For comparison with results of previously published reliability studies, ICC (model 2.1) coefficients were calculated for interrater and test-retest reliability based on the same data using the statistical package R (version 2.15.2; R Foundation for Statistical Computing, Vienna, Austria).

### Ethical Approval

The study was approved by the Regional Committees on Health Research Ethics for Southern Denmark (ID S-20130088), and informed consent was obtained from all participants.

### Results

### QST

QST test side and order were alternated between raters and occasions as described in the methodology to ensure that any effect thereof was evenly distributed. Formal testing (Mann-Whitney U test) of group differences (left vs right and first vs second) for each of the 7 QST variables revealed no statistically significant differences (.43 < \( P < .85 \)).

The mean QST results for all participants were as follows: PPDT = 662.36 kPa (95% confidence interval [CI] = 613.17, 711.54), PPI = 15.13 mm (95% CI = 12.18, 18.08), CPTD = 57.51 seconds (95% CI = 55.94, 59.08),

<table>
<thead>
<tr>
<th>VARIANCE COMPONENT</th>
<th>EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_p^2 )</td>
<td>The variance in scores attributable to real differences in participants’ pain. This is known as the “universe score” variance in GT. The equivalent in CTT is the “true score” variance ( \sigma_t^2 ).</td>
</tr>
<tr>
<td>( \sigma_r^2 )</td>
<td>The variance in scores attributable to rater differences in administering the test, eg, differences in knowledge, skills, and attitudes.</td>
</tr>
<tr>
<td>( \sigma_o^2 )</td>
<td>The variance in scores attributable to the choice of the occasion (day) the measurement took place.</td>
</tr>
<tr>
<td>( \sigma_{po}^2 )</td>
<td>The variance in scores attributable to the interaction or “chemistry” between rater and participant.</td>
</tr>
<tr>
<td>( \sigma_{ro}^2 )</td>
<td>The variance in scores attributable to the interaction between the participant and the occasion (day) the measurement took place.</td>
</tr>
<tr>
<td>( \sigma_{po}^2 )</td>
<td>The variance in scores attributable to the interaction between the rater and the occasion (day) the measurement took place. Different raters may perform the tests differently on different days.</td>
</tr>
<tr>
<td>( \sigma_{po}^2 )</td>
<td>The variance in scores attributable to the interaction between all effects (( p, r, ) and o) as well as random error (( e )).</td>
</tr>
</tbody>
</table>

As seen in equation 4, all error variances can be minimized by a factor corresponding to the \( n_r \) and \( n_o \) used in the test situation. D studies give the researcher the power to plan and control the reliability of an important test situation while also taking into account the feasibility.
Table 2. ICCs (Model 2.1) for Test-Retest Reliability (Days Apart) and Interexaminer Reliability (Hours Apart)

<table>
<thead>
<tr>
<th>QST</th>
<th>RATER 1</th>
<th>RATER 2</th>
<th>OCCASION 1</th>
<th>OCCASION 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI</td>
<td>.57</td>
<td>.50</td>
<td>.44</td>
<td>.81</td>
</tr>
<tr>
<td>CPM</td>
<td>.19</td>
<td>.38</td>
<td>.33</td>
<td>.49</td>
</tr>
<tr>
<td>PPDT</td>
<td>.70</td>
<td>.57</td>
<td>.79</td>
<td>.62</td>
</tr>
<tr>
<td>SPD</td>
<td>.48</td>
<td>.44</td>
<td>.34</td>
<td>.59</td>
</tr>
<tr>
<td>SPS</td>
<td>.65</td>
<td>.88</td>
<td>.72</td>
<td>.84</td>
</tr>
<tr>
<td>CPTD</td>
<td>.48</td>
<td>.56</td>
<td>.31</td>
<td>.86</td>
</tr>
<tr>
<td>CPTS</td>
<td>.76</td>
<td>.86</td>
<td>.87</td>
<td>.92</td>
</tr>
</tbody>
</table>

CPTS = 2,478.05 seconds · mm (95% CI = 2,148.30, 2,807.81), CPM = 4.91 mm (95% CI = 2.73, 7.09), SPD = 152.41 seconds (95% CI = 142.73, 162.09), and SPS = 4,133.72 seconds · mm (95% CI = 3,497.84, 4,769.60).

ICCs

The ICCs (model 2.1) for test-retest and for interexaminer reliability are presented in Table 2. The ICCs ranged from .19 to .92, with a mean of .6.

In other words, classical reliability testing indicated that 5 ICC tests had poor reliability (ICC < .4), 10 were fair (.4 ≤ ICC < .6), 4 were good (.6 ≤ ICC < .75), and 9 were excellent (ICC ≥ .75).

Generalizability

Variance Components

The absolute and relative contributions to total variance in QST scores are presented in Table 3.

Generally speaking, the largest variance component was attributable to the participant (p effect, Table 3) for all QST scores, except for CPM. For example, 50.8% of the total variance in PPDT scores was explained by differences between participants.

Table 3. Variance Components for 7 Measures of Pain: Absolute Values and Relative Contributions to the Total Variance in Scores

<table>
<thead>
<tr>
<th>VARIANCE COMPONENTS</th>
<th>PPI</th>
<th>CPM</th>
<th>PPDT</th>
<th>SPD</th>
<th>SPS</th>
<th>CPTD</th>
<th>CPTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>σ²₁ (%)</td>
<td>72.55 (41.6)</td>
<td>18.43 (19.3)</td>
<td>26,463.21</td>
<td>610.13 (30.1)</td>
<td>5,624,811.68</td>
<td>30.9 (52.4)</td>
<td>1,683,507.48</td>
</tr>
<tr>
<td>σ²₂ (%)</td>
<td>4.59 (2.6)</td>
<td>0.0 (0.0)</td>
<td>4,544.58 (8.7)</td>
<td>88.75 (4.4)</td>
<td>340,914.97</td>
<td>26.4 (4.4)</td>
<td>62,032.88</td>
</tr>
<tr>
<td>σ²₃ (%)</td>
<td>0.0 (0.0)</td>
<td>4.48 (4.7)</td>
<td>3,881.94 (7.4)</td>
<td>295.58 (14.6)</td>
<td>324,508.20</td>
<td>26.4 (4.4)</td>
<td>62,032.88</td>
</tr>
<tr>
<td>σ²₄ (%)</td>
<td>14.78 (8.5)</td>
<td>12.49 (12.9)</td>
<td>3,588.29 (6.9)</td>
<td>287.67 (14.2)</td>
<td>385,189.14</td>
<td>26.4 (4.4)</td>
<td>62,032.88</td>
</tr>
<tr>
<td>σ²₅ (%)</td>
<td>42.25 (24.2)</td>
<td>22.23 (23.3)</td>
<td>7,369.43 (14.1)</td>
<td>274.99 (13.6)</td>
<td>385,189.14</td>
<td>26.4 (4.4)</td>
<td>62,032.88</td>
</tr>
<tr>
<td>σ²₆ (%)</td>
<td>0.76 (4.4)</td>
<td>0.0 (0.0)</td>
<td>1,762.15 (3.4)</td>
<td>40.23 (2.0)</td>
<td>5,397.09</td>
<td>26.4 (4.4)</td>
<td>62,032.88</td>
</tr>
<tr>
<td>σ²₇ (%)</td>
<td>39.48 (22.6)</td>
<td>37.96 (39.8)</td>
<td>4,500.65 (8.6)</td>
<td>427.81 (21.1)</td>
<td>1,484,643.32</td>
<td>27.59 (46.7)</td>
<td>161,263.54</td>
</tr>
<tr>
<td>σ²₈ (%)</td>
<td>174.4 (100)</td>
<td>95.38 (100)</td>
<td>52,110.25</td>
<td>2,025.17 (100)</td>
<td>8,166,094.4</td>
<td>59.02 (100)</td>
<td>2,165,697.93</td>
</tr>
<tr>
<td>σ²ₑ (%)</td>
<td>108.86</td>
<td>76.96</td>
<td>25,647.04</td>
<td>1,415.04</td>
<td>2,541,282.71</td>
<td>28.11</td>
<td>482,190.45</td>
</tr>
<tr>
<td>SD (σₑ)</td>
<td>8.52</td>
<td>4.29</td>
<td>162.68</td>
<td>24.70</td>
<td>2,371.67</td>
<td>5.56</td>
<td>1,297.50</td>
</tr>
<tr>
<td>SD (σₑ)</td>
<td>10.09</td>
<td>8.77</td>
<td>160.15</td>
<td>37.62</td>
<td>1,594.14</td>
<td>5.30</td>
<td>694.40</td>
</tr>
</tbody>
</table>

Abbreviation: SD, standard deviation.

NOTE. The variance (σ²) in this table is for a test situation with 1 rater and 1 occasion. σ₁ is the absolute error variance, that is, σ₁ = σ²₁ + σ²₂ + σ²₃ + σ²₄ + σ²₅ + σ²₆ + σ²₇ + σ²₈ + σₑ.

Decisions Studies for Single QST Procedures

Generally speaking, the variation attributable to the interaction between the participant and occasion (po) and the residual (pro,e) was also considerable, whereas the variation between the rater (r), the occasion (o), the participant-rater interaction (pr), and the rater-occasion interaction (ro) was comparatively low (0.0–14.6% of the total variance).

Alternative Test Situations for a QST Battery

Table 5 lists examples of composite dependability coefficients for alternative QST batteries. The examples given in Table 5 are just a few of many possible alternative test situations.

Using a carefully selected composite test battery can increase dependability considerably relative to individual QST procedures. However, the choice and combination of QST procedures is important (compare test situations 1 and 2 in Table 5).

A test battery of all 7 tests was relatively insensitive to differential weighting of the individual tests (compare test situations 3 and 4 in Table 5).

Employing the full, equally weighted test battery on a single occasion with a single rater yielded an excellent composite dependability coefficient of .78 (test situation 3 in Table 5). Increasing the number of test occasions to 2 raised the dependability coefficient to .85 (test situation 5 in Table 5). Increasing the test situation to 2 raters on 2
occasions (i.e., similar to the present study design) yielded a dependability coefficient of .90 (see Fig 2).

Generally speaking, increasing the number of occasions resulted in a greater increase in dependability than did increasing the number of raters (see Fig 2).

Discussion

The current study demonstrated the following:

- The variance attributable to differences between study participants (the universe score) was the greatest of the observed variance components in 6 of the 7 QST procedures.
- The relative contribution of the universe score ranged from as much as 77.7% to as little as 19.3% of the total variance for individual QST procedures.
- The dependability and feasibility of QST was improved substantially by applying a test battery.
- The combination of QST procedures in the battery was important.
- Differentiated weighting of QST procedures in this battery of 7 tests added little to the dependability coefficients.
- Increasing the number of occasions resulted in greater improvements in dependability than did increasing the number of raters.

Three components in combination (p, po, and pro,e) explained most of the variation (between 65 and 99%) for individual QST procedures. In other words, the 3 greatest sources of variance were differences between participants (p, the universe score), within-participant variation over time (po), and error (e).

Thus, the current findings suggest that investigators seeking to improve the reliability of individual QST procedures should invest their resources in repeated measurements on different occasions. By calculating an average from different occasions (after identifying and eliminating any outliers), the greatest source of error variance—that attributable to participant-occasion interaction—can be reduced.

The number of retests (on separate occasions) necessary to achieve an excellent reliability (<.75) varied between QST procedures. For the most reliable single QST procedure (CPTS), an excellent reliability (.78) could be achieved by a single test. In comparison, the commonly applied PPDT required 3 test occasions by 2 raters. With a single retest of the combined, equally weighted battery of 7 tests, an excellent reliability of .85 could be achieved.

In addition, using a larger test battery boosted the composite test generalizability to a degree that allowed for less sampling of raters and occasions (which means increased feasibility) than would have been achievable with even the most reliable single QST on its own. It is also noteworthy that the 2 single QST procedures with the highest relative true score variances were CPTS and SPS, which are themselves composite pain scores determined by time to pain onset, pain duration, and pain intensity over time (area under the curve).

Particular circumstances such as experience with QST and time and equipment available may dictate

Table 4. Dependability Coefficients (\( \Phi \)) for Alternative Test Situations for 7 Measures of Experimental Pain

<table>
<thead>
<tr>
<th>Test</th>
<th>( n_r )</th>
<th>( n_o )</th>
<th>( W_t )</th>
<th>( \Phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI</td>
<td>1,1</td>
<td>.25</td>
<td>.45</td>
<td></td>
</tr>
<tr>
<td>CPM</td>
<td>1,1</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPDT</td>
<td>1,1</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPD</td>
<td>1,1</td>
<td>.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPS</td>
<td>1,1</td>
<td>.33</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>CPTD</td>
<td>1,1</td>
<td>.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CPTS</td>
<td>1,1</td>
<td>.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: \( W_t \), weighting assigned to individual test in the test battery; \( \Phi \), composite dependability coefficient for the test battery.

NOTE. This table shows just 4 illustrative examples of many possible alternatives. Different \( n_r, n_o \) combinations may for instance also be assigned to each individual test when planning to optimize composite dependability to a target value (not shown here).
otherwise, but a suitably large, carefully chosen QST battery performed on multiple occasions is likely to considerably improve the generalizability of QST.

It should be noted that pain sensitivity may fluctuate systematically over time, for example, in relation to the female menstrual cycle,\textsuperscript{11} and this could affect the amount of variation attributable to occasion and occasion-interactions. However, any effect of such systematic variation over time would have been countered in the current study by the random selection of occasions within a 1-month period.

Closer examination of Table 3 reveals that the reliability of CPM was particularly poor. This may be related to the particular manner in which CPM was assessed in the current study: The test stimulus was not very intense (mean VAS = 15, standard deviation = 13), and although the conditioning stimulus (CPT at 0–2°C) was quite painful, it was tolerated for the full 1 minute in 59 of 76 tests; this may have been insufficient to induce a robust CPM response. The poor reliability of CPM in the current study does not necessarily reflect that of other methods of assessing CPM.

As described in the introduction, classical tests of reliability, such as ICCs, are commonly used in relation to QST and are often reported to indicate fair to excellent reliability. With 23 of 28 ICCs being fair to excellent, the current study compares well to those discussed in previous studies.\textsuperscript{4,5,10,18-21,25,26,31,33-35}

QST is used in the literature to examine differences in pain sensitivity between clinical groups, for example, chronic pain patients and healthy controls. In those contexts, QST is used as an indicator of pain sensitivity and, by inference, as an indicator of abnormalities in pain sensitivity. Whether QST is a valid indicator of pain sensitivity is inextricably linked to the reliability of QST procedures. Streiner and Norman\textsuperscript{27} state it quite clearly: “Reliability places an upper limit on validity, so that the higher the reliability, the higher the maximum possible validity (more formally, the maximum validity of a test is the square root of the reliability coefficient).” Accordingly, in modern validity theory, evidence of sufficient generalizability is a necessary step in construct validation.\textsuperscript{10} As a consequence, single QST procedures with poor reliability cannot validly assess a phenomenon such as hyperalgesia. Nonetheless, several published studies have reported differences in pain sensitivity between clinical groups based on single QST procedures.\textsuperscript{8,12,13,23,24,31}

The fact that many such studies report a significant group difference, in addition to relying on a single QST procedure, could be the result of publication bias, but it also might be an indication that differences in pain sensitivity between clinical subgroups is in fact a rather robust phenomenon. In other words, group differences in pain sensitivity appear to be pronounced enough to be detectable in spite of less reliable measurements.

Conducting and analyzing a fully crossed generalizability study is a somewhat more involved affair than CTT studies but will in most real-life clinical situations provide less biased estimates of reliability, as most test
situations are indeed influenced by more than one source of error variance simultaneously. To our knowledge, the only other GT study on QST was that by Pryseley et al. The reliability study published by the German Research Network on Neuropathic Pain actually appears to have been a fully crossed, multivariate generalizability design similar to the present study, but the data were analyzed and presented using CTT (test-retest and interexaminer reliability).

In addition to estimating more discrete sources of error than CTT, generalizability studies empower researchers in their decisions on how to best conduct reliable and feasible experiments. It is important, however, to underline a limitation of the current study: The generalizability (coefficients) reported in the present study does not imply that the results can be generalized to other particular conditions, for example, other QST procedures or study groups such as pain patients. Rather, the current findings can help elucidate which parameters are likely to be more important in obtaining good reliability: more raters, tests, or occasions.

### Conclusions

The current study is the first generalizability study of its kind of quantitative sensory pain testing to our knowledge. The findings suggest that the reliability of some QST procedures may be overestimated by CTT reliability estimates such as ICCs. Furthermore, in order to increase reliability, it appears that researchers are best served by administering several carefully chosen individual QST procedures in a test battery and secondarily by repeating measurements on several occasions. Reliability is only marginally improved by retesting with different raters, although this is also a separate source of error variance at work simultaneously.

### References


2. Brennan RL: Computer Programs - Center for Advanced Studies in Measurement and Assessment (CASM): College of Education - The University of Iowa [Internet]. [cited 2014 Jan 1]. Available at: http://www.education.uiowa.edu/centers/casma/computer-programs#8f748e48-ff00000648cd


QST Instructions

Pressure pain threshold: “I will now apply gradually increasing pressure on your leg. When the pressure begins to feel painful, please press the button. In other words, please indicate the threshold at which the pressure changes from merely being pressure to becoming a painful pressure. […] Do you understand? […] I will repeat the pressure three times.”

Pressure pain intensity: “I will now place this spring clamp on your thumb. After 10 seconds I want you to indicate how painful the pressure is on this scale by writing an ‘X’ somewhere along the line—this end of the line represents ‘No pain’ and the other end ‘Worst imaginable pain.’ In other words, if the clamp is not painful at all, you should put your X at this end. If you can not imagine anything more painful, you should put your X at this end. If the pain is somewhere in between, your X should be somewhere in between either end of the line. Do you understand? […]”

Cold pressor test: “When I tell you, please lower your hand into the water, as far as your wrist. When your hand is in the water please do not make a fist or clench your hand and try to keep it there for 1 minute. Only withdraw your hand if the pain is unbearable. With your other hand please indicate how painful the cold water is, on the computer scale by scrolling up or down with the ball. If the pain changes over time, please indicate so on the scale. […] Same as before, this end of the scale represents ‘No pain’ and the other end ‘Worst imaginable pain.’ I will keep track of the time, and let you know when 30 and 50 seconds have passed. Do you understand? […]”

Injection of hypertonic saline: “I will now inject the saline. Please indicate the pain using the computer scale by scrolling up or down with the ball in the same manner as before. If the pain changes over time, please indicate so on the scale. […] Please try to ignore the discomfort from the needle puncture itself, and indicate the muscle pain induced by the saline. Do you understand? […]”

Appendix

462 The Journal of Pain


Improving QST Reliability