J Musculoskelet Neuronal Interact 2023; 23(1):145-164

Journal of Musculoskeletal and Neuronal Interactions JMNI

Review Article

Psychometric properties of quantitative sensory testing in healthy and patients with shoulder pain: A systematic review

Paraskevi Bilika[1], Achilleas Paliouras[1], Konstantina Savvoulidou[1], Alberto Arribas-Romano[2], Zacharias Dimitriadis[3], Evdokia Billis[4], Nikolaos Strimpakos[3,5], Eleni Kapreli[1]

[1]Clinical Exercise Physiology and Rehabilitation Research Laboratory, Department of Physiotherapy, School of Health Sciences, University of Thessaly, Lamia, Greece;
[2]Department of Physical Therapy, Occupational Therapy, Rehabilitation and Physical Medicine, Rey Juan Carlos University, Madrid, Spain;
[3]Health Assessment and Quality of Life Research Laboratory, Department of Physiotherapy, School of Health Sciences, University of Thessaly, Lamia, Greece;
[4]Physiotherapy Department, School of Health Rehabilitation Sciences, University of Patras, Greece;
[5]Division of Musculoskeletal & Dermatological Sciences School of Biological Sciences Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

Abstract

Quantitative Sensory Testing (QST) is a psychophysical battery of various tests developed to quantify the subjects' self-reported sensory experience. Although the use of QST is valuable for the clinical assessment of pain, standard evaluation protocols have not yet been established. This systematic review aimed to investigate the level of evidence for the psychometric properties of QST in healthy and patients with shoulder pain. Eight databases were searched for peer-reviewed studies published until August 2021. The methodological quality of studies was evaluated using the COSMIN checklist. Twelve studies were included for qualitative synthesis, which included three different tests (Pressure Pain Threshold (PPT), Conditioned Pain Modulation (CPM) and Temporal Summation (TS)). As the body of evidence consisted of studies of low methodological quality, the psychometric properties of PPT, CPM, and TS in healthy and patients with shoulder pain were classified as unknown. Although there is a risk that the conclusions may be 'superficial' in nature, the reliability seems to be nearly excellent for the PPT, however, the protocols' variation and the low methodological quality of the studies do not allow for clear conclusions. Further studies are required for the CPM and TS in patients with shoulder pain.

Keywords: Psychometric properties, Quantitative Sensory Testing, Shoulder, Systematic Review

Introduction

Shoulder Pain (SP) is a complex condition that is often accompanied by pain, disability, sleep disorders, loss of workdays and psychological distress[1-4]. Forty percent of the patients with SP report persistent symptoms 12 months after onset[5,6]. Various factors have been blamed from time to time for delaying recovery, but the pathophysiological mechanisms have not yet been elucidated[7-10]. Generalized hyperalgesia, widespread pain, emotional distress and comorbidities have been reported in a subgroup of patients with chronic SP[11-15]. These indications can not be explained by clinical signs of tissue damage or inflammation, but they have been associated with Central Nervous System (CNS) sensitization. Central Sensitization (CS) is an amplification of neural signaling within the CNS which leads to pain hypersensitivity[15]. Pain hypersensitivity may play a role in the transition from acute to chronic pain[16]. Patients with pain hypersensitivity seem more likely to develop CS[17,18]. The presence of CS requires a different approach to rehabilitation as its management with standard therapeutic procedures is rather ineffective[19,20]. At this point it should be mentioned that SP can be further the result of serious pathology such as in patients with gastrointestinal and hepatic diseases[21], however these cases extend beyond the objectives of

The authors have no conflict of interest.

Corresponding author: Paraskevi Bilika, Clinical Exercise Physiology and Rehabilitation Research Laboratory, Department of Physiotherapy, School of Health Sciences, University of Thessaly, 35100 Lamia, Greece
E-mail: pbilika@uth.gr

Edited by: G. Lyritis
Accepted 29 August 2022

the present study which investigates musculoskeletal shoulder pain.

At present, there is no consensus on the most appropriate tool for assessing pain sensitivity and CS[22]. Self-reported questionnaires can detect symptoms and disorders associated with predominant CS pain mechanisms, however the literature highlights Quantitative Sensory Testing (QST) as the most representative indicator for determining the presence of CS[23-25]. QST is a general term of a battery of tests assessing perception, pain tolerance, and pain threshold through different stimuli (standardized for the assessment) and designed to quantify the subjects' self-reported sensory experience[26,27]. QST modalities evaluate the sensory processing of large and small sensory fibers and can provide significant information on pain mechanisms[28,29]. Indications of peripheral or central sensitization can be assessed using QST in the affected or non-affected areas[25]. The most common stimuli used in QST tests are pressure, heat, cold, electrical and vibration, most of which require special equipment. QST involves both static and dynamic tests[27,30]. Static QST (such as Pressure Pain Threshold PPT) include the determination of a stimulus (pain detection, pain tolerance and pain threshold) and the determination of stimulus intensity[27]. Dynamic QST are central integration tests such as Temporal Summation (TS) and Conditioned Pain Modulation (CPM) that control the endogenous pain inhibition mechanisms[27]. The difference between the two categories is that static tests control the sensory response to a single stimulus at a single test site and clearly can not provide a complete view of sensory processing systems. On the other hand, dynamic tests include the contribution of at least 2 stimuli assessing the function of descending pain pathways or pain wind up[31]. Dynamic tests are predictive of the development of CS mechanisms[32]. The prognostic ability of QST has also been studied in individuals with chronic musculoskeletal pain[33]. Recent studies have highlighted the importance of QST, as it appears that pain hypersensitivity can predict acute postoperative pain[34,35], worse musculoskeletal outcomes of pain, disability and negative effects[33].

Although the use of QST is valuable for the clinical assessment of pain, different protocols have been proposed in the literature confusing healthcare practitioners. Studies have reported that standardization of instruction to subjects, technique training, instruments' calibration, and stimulus selection are all necessary for QST reliability[36]. The use of reliable and validated testing procedures in clinical practice is the cornerstone of both assessment and monitoring the effectiveness of treatment. Although there are reliability studies, there is no aggregation of all this evidence to create a complete view of the psychometric properties of QST. Therefore, the purpose of this systematic review (SR) was to identify studies that describe the psychometric properties of one or more QST tests in healthy subjects and patients with SP aimed at objectively evaluating pain sensitivity and evaluating their methodological quality. A secondary purpose was to highlight the most reliable tools and procedures from the various tests for evaluating patients with SP.

Methods

Protocol and Registration

This SR was developed in accordance with Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines[37] and COSMIN recommendations[38]. The protocol for this SR was registered in the PROSPERO database https://www.crd.york.ac.uk/prospero/ with registration number CRD42021232778.

Eligibility criteria

The inclusion and exclusion criteria have been described in detail in the published SR protocol[39]. The construct was the psychometric properties of QST in the shoulder area. The present study investigated the use of the QST in pain-free adults or patients with self-reported musculoskeletal shoulder pain of any etiology. In order to enhance the sensitivity of the search, studies using a neck-shoulder pain population without a well-defined diagnosis of neck pathology were included. Only studies written in English and Greek were included. In this SR, it was chosen not to search the gray literature as reviewers have not evaluated it and this would bias the study.

Information Sources and Search Strategy

The literature search conducted searches using seven databases (Cochrane, EMBASE, MEDLINE through PubMed, Web of Science, Scopus, SportDiscuss and PEDro). All databases were investigated on the same day (31 August 2021). The search strategy was designed using keywords and MESH terms related to psychometric properties, quantitative sensory testing and shoulder, appropriate for each database. The final search strategy is presented in the supplementary data (Appendix 1). The search was performed without restrictions on date or language. A researcher (A.A.R.) with experience in SRs contributed to the literature search. Additionally, the reference list from relevant articles was further checked to ensure the maximum possible search results.

Deviation from the protocol

Although it was mentioned in the protocol that the AMED database would be searched, the authors had no access. Finally, the search strategy was revised after discussion by the researchers to make it more accurate concerning the research question.

Study Selection

All retrieved articles were imported into Rayyan Application (https://rayyan.ai/) and duplicates were removed. Initially, titles and abstracts of the retrieved articles were screened separately for inclusion by two reviewers (BP and PA). The reviewers used a formulated questionnaire with the eligibility criteria. Before the onset of the screening,

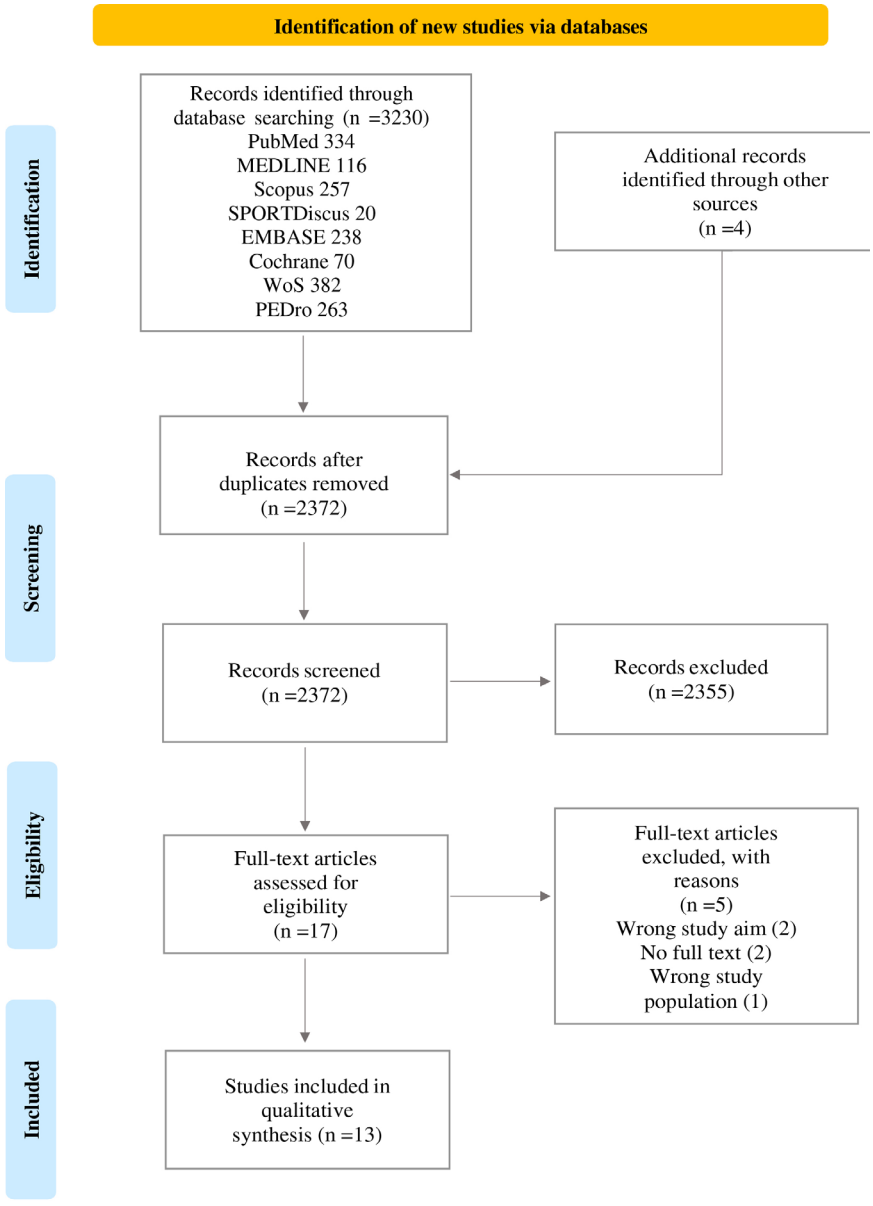

Figure 1. PRISMA flow diagram[56].

a pilot study was carried out that examined five articles to check the consistence between the raters. The agreement between the raters was excellent (k=1)[40]. Establishing clear and detailed eligibility criteria helped to enhance inter-rater agreement. In case of disagreements, a third reviewer (SK) was consulted to make the final decision. At that stage, studies that did not fit into the purpose, construct, population and type of studies were excluded. In case of uncertainty regarding the incorporation of the article, the full text was reviewed. Reading the full text, the selected articles were checked again. Where the full text was not available, an email was sent to the authors. The authors had 2 weeks to respond otherwise, the study would not be included in the review.

Data Selection Process

A data extraction sheet, based on COSMIN recommendations[41-43], was developed. In addition, fields related to the measurement protocol (rate, model, number of measurements, etc.) were added. The data extraction sheet was tested on a pilot basis in two randomly selected studies and improved accordingly. The extraction process was carried out by the lead researcher (P.B.).

Risk of Bias in Individual Studies

The methodological quality of each included study was independently evaluated using the COSMIN Risk of Bias

Table 1. Summary information of PPT studies.

Pressure Pain Threshold							
Study	**Instrument**	**Population**	**Sample Characteristics**	**Construct**	**Pressure Points of Measurement**	**Measurement properties**	**Interval time**
De Groef et al. 2016	Digital Wagner FPX Algometer (Greenwich, CT, USA)	30 H F (middle-aged)	Age Mean: 50.3 (±7.3) years [40 to 60 years] **BMI** Mean: 25.8 (±3.8) kg/m^2	PPT (The subject was asked to say "stop" when the sensation of pressure first changed to pain) **Rate of pressure**: 1 kg/second.	Points of measurement were defined by palpation for tender muscle points in the region of the upper trapezius (between the C7 spinous process and the acromion), supraspinatus (above the spine of the scapula), infraspinatus (muscle belly under the spine of the scapula), pectoralis major (under the clavicle), pectoralis minor (between the caudal edge of the 4th rib and the inferomedial aspect of the coracoid process) and serratus anterior (below the axilla, on the muscle belly which branches to the ribs).	Interrater reliability, Measurement error	5 minutes
Nascimento et al. 2019	Digital Wagner FPX Algometer (Greenwich, CT, USA)	78 participants, 52 P for unilateral SIS 26 H participants	Symptomatic Group 1 Age Mean: 29 ± 17.50 [18 to 60 years], BMI Mean: 24.26 ± 2.75 Sex: 12 M; 14 F Symptomatic Group 2 Age Mean: 35.12 ± 9.73 [18 to 60 years], BMI Mean: 24.42 ± 2.52 Sex: 16 M; 10 F Asymptomatic Group Age Mean: 28 ± 17.50 BMI Mean: 23.99 ± 3.18 Sex: 12 M; 14 F	PPT ("I am going to start applying pressure to your muscle with this instrument. When the pressure becomes uncomfortable, say "stop") **Rate of pressure**: 0.50 kgf/s	lower trapezius (in the muscle belly, halfway between the midpoint of the medial border of scapula and the spinous process of the twelfth thoracic vertebra), upper trapezius (halfway between the C7 spinous process and acromion process), infraspinatus (in the muscle belly below the midpoint of the spine of scapula), and medial deltoid (muscle belly, near the inferolateral insertion)	Intraday intrarater reliability, Intraday, interrater reliability, interday intrarater reliability, Measurement Error	Intraday interrater: 5 minutes Intraday intrarater: unclear Interday interrater: 48 hours
Wang-Price et al. 2019	Pressure Algometer, Medoc Ltd, Ramat Yishai, Israel	60 participants - 30 H adults and 30 P with neck-shoulder pain and tenderness	Asymptomatic Group: Age Mean:26.9 ± 5.7 BMI: N/A Sex: 21 F;9 M Symptomatic Group: Age Mean:29.9 ± 8.8 BMI: N/A Sex: 24 F; 6 M	**PPT** At the start of testing, a patient response unit with a red button was given to each participant. The participant was instructed to hold the response unit in the nontesting hand and press the button once the sensation of pressure changed to a sensation of pain. **Rate of pressure**: 40 kPa/s	For the middle deltoid muscle, the pressure algometer was applied at the midpoint between the insertion of deltoid and acromion. For the levator scapulae muscle, it was applied at the point 2 cm above the lower insertion of levator scapulae located in the upper medial border of the scapula. Last, for the upper trapezius muscle, it was applied at the point halfway between the midline and lateral border of the acromion	Intraday & Interday intrarater reliability, Measurement Error	Intraday intrarater: unclear Interday intrarater: 3-7 days
Vaegter et al. 2018	Pressure Algometer, Somedic Sales AB, Sweden, Horby	35 H	Mean age [range] = 23.1 ± 2.2 [20–30] years, Average BMI [range] = 23.1 ± 1.7 [20.0–27.5], 17F;18M)	PPT- (the first time the pressure was perceived as minimal pain, the subject pressed a button, and the pressure intensity defined the PPT) **Rate of pressure**: 30 kPa/s	Site 1 was located in the middle of the dominant quadriceps muscle, 15 cm proximal to the base of patella. Site 2 was located in the nondominant upper trapezius muscle, 10 cm from the acromion in direct line with the seventh cervical vertebra.	Intraday intrarater reliability, Interday intrarater reliability	Intraday intrarater: 3 min. Interday intrarater reliability: 1 week

Table 1. *(Cont. from previous page).*

Jones et al. 2007	Pressure Algometer, Somedic Sales AB, Sweden, Horby	19 H F	Mean Age: of 23.9 ± 5.2 years, [range: 20-39] Mean BMI: 23.6±3.5 19F	PPT -the "instant or moment that the pressure on the skin surface changed from the sensation of pressure to the sensation/ perception of pain." **Rate of pressure**: Unclear	1) the midpoint of the muscle fibers of the long head of biceps, 2) the midpoint of the anterior fibers of the deltoid, 3) the midpoint of the muscle fibers of the lateral head of the triceps, 4) the upper half of the fibers of the rhomboids major, 5) the midpoint of the posterior fibers of the deltoid, 6) the proximal one-third of the fibers of the supraspinatus (the location closest to the medial border of the scapula), 7) the upper fibers of the trapezius (medial to the superior angle of the scapula) and 8) the distal one-third of the muscle fibers of the infraspinatus.	Intraday intrarater reliability, Interday intrarater reliability	Interday intrarater: 4 consecutive days
Persson et al. 2004	Pressure Algometer, Somedic Sales AB, Sweden, Horby	27 H F	Mean age: 42 (range, 24–59) years, Mean height: 167 (range, 151–174), Mean weight: 65 (range, 52–90) kg 27F	PPT (The subjects were instructed to press the signal button, held in the dominant hand, when the sensation of "pressure" changed to one of "pain or discomfort" and the measurement ceased at that time). **Rate of pressure**: 40 kPa/s	3 points over the descending part of the trapezius (points T1, T2, T3) muscle, along a straight line from the spinous process of the 7th cervical vertebra to the lateral edge of the acromion, and 4 points over the mid-portion of the deltoid (points D4, D5, D6, D7) muscle.	Intraday intrarater reliability, Intrerday intrarater reliability, Interday interrater reliability	Intraday intrarater: 10 minutes. Intrerday intrarater: 1 day and month. Interday interrater reliability: 0-2 days
Vanderweeën et al. 1996	A pain threshold meter, model PTH-AF 2, commercially available through the Pain Diagnostic and Treatment Corporation (Great Neck, NY 11021, USA)	30 P with chronic unilateral pain	Sex: 15F, 15M	PPT (Subjects were instructed to say 'yes' as soon as the sensation of pressure changed to pain) **Rate of pressure**: 1 kg/s	14 trigger points were evaluated on both sides of the body. 8 were paravertebral and 6 in the shoulder and arm region. The location and innervation of these trigger points have been described by Travell and Simons (1983). In the extremities the 6 trigger points measured were located in the pectoralis major, supraspinatus, infraspinatus, trapezius, extensor carpi radialis brevis and the first dorsal interosseous muscle.	Intraday intrarater reliability	5 minutes
Levoska et al. 1993	model PTH-AF 2, commercially available through the Pain Diagnostic and Treatment Corporation (Great Neck, NY 11021, USA)	100 F office workers (33 P with nesk-shoulders symptoms 67 H)	Mean age: 38 (range 20-55) years Mean height: 163 (range 149-178) cm, Mean weight: 60 (range 44-115) kg 100F	PPT (subjects were asked to say "now" when the sensation stopped being pressure and began to be a definite pain) **Rate of pressure**: 1 kg/s	The measurement points of the right and left trapezius and levator scapulae muscles were chosen for measurement of pain threshold according to the theory of myofascial pain espoused by Travell and Rinzler (1952).	Interday intrarater reliability Intraday interrater reliability	Intrarater: 2 days (n=40). Interrater: unclear (n=60)

M=Male, F=Female, H= Healthy participants, P=Patients, SIS= Shoulder Impingement Syndrome, VAS= Visual Analogue Scale, PPT= Pressure Pain Threshold.

Table 2. Summary information of CPM and TS studies.

Conditioned Pain Modulation								
Study	**Population**	**Sample Characteristics**	**Test stimulus**	**Tested sites**	**Conditioning Stimulus**	**Tested sites**	**Measurement property**	**Interval time**
Valencia et al. 2013	190 H and 134 P (acute or sub-acute shoulder pain preparing to undergo shoulder surgery)	Mean Age 43,83±17,8 161 M (74H & 87P) 163F (116H & 47P)	SHPR Pathway Pain & Sensory Evaluation System, Medoc, Ramat, Yishai, Israel	the thenar eminence of the non-surgical side for the patients, and non-dominant side for the healthy subjects.	CPT 8°C	Affected Hand and Dominant Hand	Intraday Intrarater Reliability	2 minutes
Cathcart et al. 2009	20 H	9M;11F Mean Age (M) 27±6.4 Mean Age (F) 23±3.6	PPT Digital Wagner FPX Algometer (Greenwich, CT, USA)	Left arm	Ischemic pressure 20 mmHg/s VAS 3/10	Right Shoulder Upper Trapezius	Intraday Intrarater Reliability	60 minutes
Alsouhibani et al. 2019	30 H	15M; 15F Mean Age 19.3±1.5	PPT Pressure Algometer, Medoc Ltd, Ramat Yishai, Israel	Non-Dominant Right Shoulder	CPT 0±1°C	Left Foot	Intraday and Interday Intrarater Reliability	20 minutes 7 days
Marcuzzi et al. 2017	42 H	21M;21F Mean Age 30.2±10	TTS 30s heat (NRS 6/10) PPT FDK40; Wagner Instrument, Greenwich, CT	Non-Dominant Forearm Non-Dominant Upper Trapezius	CPT 10.5±1.0°C	Dominant Foot	Interday Intrarater Reliability	2 months 4 months
Temporal Summation								
Cathcart et al. 2009	20 H	9M;11F Mean Age (M) 27±6.4 Mean Age (F) 23±3.6	PPT (10 pulses)	Right Shoulder Upper Trapezius			Intraday Intrarater Reliability	60 minutes

M=Male, F=Female, H= Healthy participants, P=Patients,VAS= Visual Analogue Scale, NRS= Numeric Rating Scale, SHPR=Suprathreshold heat pain response, PPT= Pressure Pain Threshold, TTS=Thermal Test Stimulus, CPT=Cold Pressure Test.

checklist by two reviewers (P.B. and E.K.)[38]. Each study was rated as very good, adequate, doubtful, or inadequate quality. The consensus was reached through discussion and the contribution of a third reviewer (Z.D.). The overall rating of the quality of each individual study on a measurement property was determined as the lowest rating among all response options within one section, termed the 'worst score counts" principle. Then, the results of each study were rated as either sufficient (+), insufficient (-), or indeterminate (?), following the updated criteria for good measurements properties[41].

Results

Study Selection

A total of 3.234 articles were retrieved from the search, of which 862 were removed as duplicates. Manual searches yielded 4 additional studies. Considering title and abstract screening 2.355 articles were excluded. Seventeen full-text articles were evaluated for eligibility. It was necessary to contact the authors asking for the full text of two articles, but with no response therefore, the articles were rejected. Furthermore, three articles were excluded because they had a different purpose. Finally, twelve studies[44-55] were included for qualitative synthesis, which included three different tests (PPT, CPM and TS) from the QST protocol. The selection procedures are summarized in Figure 1.

According to the selected articles for the PPT test have been checked test-retest reliability, inter-rater reliability, construct validity and measurement error. Test-retest reliability and measurement error were evaluated for the CPM and only test-retest reliability was rated for the TS. Details of the protocols for the individual studies are listed in Tables 1-2.

Table 3. Risk of bias assessment

Authors	Intra-rater Reliability		Inter-rater Reliability		Measurement Error		Construct Validity	
	COSMIN RoB	COSMIN Rating	COSMIN RoB	COSMIN Rating	COSMIN RoB	COSMIN Rating	COSMIN RoB	COSMIN Rating
De Groef et al. 2016 (PPT)			Doubtful		Doubtful			
Nascimento et al. 2019 (PPT)	Doubtful		Doubtful		Doubtful			
Wang-Price et al. 2019 (PPT)	Doubtful				Doubtful		Very good	
Vaegter et al. 2018 (PPT)	Doubtful				Doubtful			
Jones et al. 2007 (PPT)	Inadequate							
Persson et al. 2004 (PPT)	Doubtful		Inadequate		Doubtful			
Vanderweeën et al. 1996 (PPT)	Doubtful							
Levoska et al. 1993 (PPT)	Doubtful		Doubtful					
Valencia et al. 2013 (CPM)	Doubtful							
Cathcart et al. 2009 (CPM)	Doubtful							
Cathcart et al. 2009 (TS)	Doubtful							
Alsouhibani et al. 2019 (CPM)	Doubtful							
Marcuzzi et al. 2017 (CPM)	Doubtful				Doubtful			

=sufficient, =indeterminate, =insufficient

Pressure Pain Threshold

Study Characteristics

The PPT was assessed by eight studies[44-51]. Among all included records, four studies investigated asymptomatic participants[44,47,48,51], one study included patients (chronic unilateral shoulder pain)[49], and three studies investigated both[45,46,50] (Table 1). More specifically, one study included patients with Shoulder Impingement Syndrome (SIS)[45] and two studies included patients with neck-shoulder pain[46,50]. Four studies included only women[44,48,50,51] and four both genders[45-47,49].

The sample size ranged from 27 to 35 in four studies[44,47-49], one study had a small sample (12 and 19 participants), and three studies had a sample ranged from 60 to 100 participants[45,46,50]. Participants ranged in age from 18 to 60 years, but some studies used middle-aged (40-60 years)[44] or younger[18-39] volunteers[47,51]. In all studies[44-55] involving healthy and patients, the mean age of healthy participants was lower than that of patients.

Equipment and pressure rate

In the selected studies, four different models of digital pressure algometer were used: i) Somedic Sales AB, Sweden, Horby[47,48,51], ii) Digital Wagner FPX Algometer (Greenwich,

CT, USA)[44,45], iii) model PTH-AF 2, commercially available through the Pain Diagnostic and Treatment Corporation (Great Neck, NY 11021, USA)[49,50], iv) Pressure Algometer, Medoc Ltd, Ramat Yishai, Israel[46].

In four studies[44,45,49,50] the patients verbally informed the examiner about the interruption of the measurement whereas in four studies[46-48,51] the patients pressed a button that was connected to the algometer. The algometers had different units of measurements (kPa/s or kg/s). Therefore, the pressure rate was varied as shown in Table 1. The most common tested sites in shoulder were upper trapezius[44-51], deltoid[45,46,48,51], supraspinatus[44,48,49,51], levator scapulae[46,50], infraspinatus[44-46,48,51] and major pectoralis[44,48].

Evaluation of methodological quality

The results of the evaluation of the methodological quality separately for each study, according to the COSMIN checklist[38], are presented in Table 3. As mentioned in the methodology, the worst score determined the overall result. Some studies received multiple evaluations for their methodology, as they included multiple measurement properties. Seven studies assessed intrarater (test-retest) reliability[45-51], four studies assessed interrater reliability[44,45,48,50], five studies evaluated measurement error[44-48] and only one study assessed construct validity (Table 3).

Regarding intrarater reliability, all studies were judged to be doubtful[45-51]. In more detail, item 1 was rated as doubtful in five studies and as very good in two, item 2 was rated as doubtful in two and as very good in five studies, item 3 was rated as doubtful in four studies and as very good in three studies. Items 4 and 6 were rated as doubtful in six studies and as very good in only one, item 5 was rated as doubtful in five studies, as very good in one and as inadequate in one. Item 7 was rated as adequate in four studies, as very good in two studies, as doubtful in one study. More details are listed in the supplementary data (Appendix 2).

Four studies evaluated the reliability of PPT[44,45,48,50] between different raters with the overall score as doubtful in three (3 out of 4) studies[44,45,50] and inadequate in one (1 out of 4) study[48]. Item 1 was rated as doubtful in three (3 out of 4) studies and adequate in one (1 out of 4), items 2, 4 and 5 were rated as very good in half of the studies and doubtful in the rest. Furthermore, items 3 and 6 were rated as doubtful in three (3 out of 4) studies and as very good in one, item 7 was rated as very good in two (2 out of 4) studies, as adequate in one (1 out of 4) and as inadequate in one (1 out of 4). More details are listed in the supplementary data (Appendix 2).

Measurement error was estimated in five studies[44-48], of which the methodology quality was considered doubtful. Item 1 was rated as doubtful in four (4 out of 5) studies and as adequate in one, item 2 was rated as very good in four (4 out of 5) studies and as doubtful in one whereas items 3,4, 5 and 7 were rated as doubtful in three studies (3 out of 5) and as very good in two. Item 6 was rated as doubtful in four studies (4 out of 5), as very good in one study. More details are listed in the supplementary data (Appendix 2).

Construct validity, using the known groups methods, was assessed in one study[46]. The methodological quality was assessed as very good because an adequate description of the characteristics of the subgroups was provided and the statistical method was appropriate for the hypothesis to be tested (Appendix 2).

Evaluation of measurement properties

Across all included studies, ICC was calculated to estimate reliability, except for one[50] which were rated as indeterminate. Five studies were valued as sufficient (ICC≥0.7) and one as insufficient (ICC<0.7) (Table 3). Two studies[45,51] had excellent reliability results, one of which included patients. All studies which assessed the measurement error were considered indeterminate since Minimum Important Change (MIC) was not provided. As the body of evidence consisted of studies with doubtful and inadequate methodological quality, the reliability (intra and inter-rater) and the measurement error of PPT are classified as unknown. Construct validity was assessed in one study[46]. It was checked the hypothesis that there is a significant difference between individuals with and without neck-shoulder pain in the PPTs. The hypothesis was true for all muscles except for the upper trapezius in the prone position. Although there are no other studies examining construct validity (comparison between subgroups), there is an indication that the PPTs at the middle deltoid and levator scapulae in the seated and prone position and at the upper trapezius in the seated position are able to discriminate between the individuals with and without neck-shoulder pain.

Conditioned Pain Modulation

Study Characteristics

The reliability of CPM on the shoulder was investigated in four studies[52-55] in healthy or patients with shoulder pain (Table 2). More specifically, three studies[53-55] evaluated the reliability in healthy individuals (n=20-42) with the test stimulus applied to the shoulder area and one study[52] included healthy participants (n=190) and patients with acute and subacute shoulder pain (n=134).

Procedure

Two studies[53,55], which included healthy participants, used PPT as a test stimulus, one study[52] used Suprathreshold heat pain response (SHPR) and in one study[54] two different stimuli (PPT and TTS) were applied (Table 2). Only one conditioning stimulus was used in all studies. The most common stimulus (3 out of 4 studies) was Cold Pressure Test (CPT)[52,54,55] which was performed with different temperatures (0-11.5°C) (Table 2). One study used ischemic pressure[53] with an increasing rate of 20 mmHg/s in order to induce a pain score of 3 on VAS. The body sites, where the conditioning stimuli was applied,

were the dominant hand[52], dominant foot[54], left foot[55], or the right trapezius[53] in the healthy individuals and the affected hand in the patients[52]. Both genders were included equally in all studies (Table 2).

Methodological quality evaluation of the studies

Four doubtful-quality studies assessed the psychometric properties of CPM (52-55) (Table 3). Of the four studies, only one[54] investigated the measurement error and no study evaluated the interrater reliability. Regarding the intrarater reliability, item 1 was rated as adequate in half of the studies and as doubtful in the rest. Most studies (3 out of 4) were rated as very good in item 2 and only one study as doubtful, item 3 was rated as very good in two studies, items 4 and 5 were rated as doubtful in three studies and only one study was rated as very good. All studies were rated as doubtful in item 6 half of the studies were rated as very good in item 7 and as adequate in the rest. Having an overall score 'doubtful', only one study[54] estimated the measurement error. Items 2, 3, 4, 5, 7 were rated as very good but items 1 and 6 as doubtful. More details are listed in the supplementary data (Appendix 2).

Evaluation of measurement properties

Against the updated criteria for good measurement properties[39,41], the reliability results of all studies were rated as insufficient (ICC<0.7) and in one study[54], which assessed the measurement error, was valued as indeterminate because the MIC was not provided (Table 3). Based on the above, the reliability and measurement error of the CPM are classified as unknown.

Temporal Summation

Study Characteristics

Only one study[53] used the TS test in healthy individuals (Table 2). Nine men and eleven women participated with mean age 27±6.4 and 23±3.6 respectively[53].

Procedure

PPT was performed (10 pulses) in the right shoulder of individuals and the intrarater reliability was estimated. The interval time between measurements was 60 minutes (Table 2).

Methodological quality evaluation of the study

The methodological quality of the study was assessed as doubtful (Table 3) due to items 4, 5 and 6. Items 2, 3, and 7 were rated as very good and item 1 as adequate. More details are listed in the supplementary data (Appendix 2).

Evaluation of measurement properties

The reliability was estimated based on ICC and it was evaluated as insufficient (ICC<0.7) (Table 3). Therefore, due to the methodological quality of the study, no conclusions can be drawn about the reliability of the TS.

Discussion

To our knowledge, this is the first SR investigating the evidence for psychometric properties of QST on the shoulder in healthy participants and patients with SP. According to the results, the level of evidence varied because of the inconsistency across methodologies of the studies. Different test protocols including tests sites, button use, type of algometer, and sample characteristics, were used. The majority of studies had moderate to low methodological quality, which leads to limitations in the generalizability of the data and the drawing of clear conclusions. Therefore, no information can be obtained on the selection of the most appropriate QST procedures for the shoulder in healthy and patients with SP.

Of the twelve included studies[44-55], only five included patients (n=279)[45,46,49,50,52]. The patients were suffering from different types of pain (acute, sub-acute and chronic pain) and pathologies (shoulder impingement syndrome, neck-shoulder pain, unilateral shoulder pain), which can be a possible explanation for the variability in the results. In the current review, it appears that the patients showed higher relative and absolute reliability compared to the healthy individuals even in different positions and between different assessment days. Previous studies have shown that QST depends on the participants' current psychosocial state[57]. Factors such as motivation, mood, menstruation and compliance, may play a key role in the stability of results[57]. Only two studies[49,51] provided evidence for the psychosocial stability of patients (Item 1).

Although there are recommendations for standardization of procedures and conduct of high-quality studies, the QST tests lack standardization, resulting in questionable reliability results[58,59]. One factor that has affected external validity was the limited number of studies with different raters. Only four studies[44,45,48,50] evaluated the reliability between raters, none of which showed good methodological quality. This SR found no CPM and TS interrater reliability studies. The training of the examiners in the measurement process may affect the results of the QST, however, only two studies[44,45] reported the training process of the raters. Across all included studies, only three studies[44,45,54] reported that examiners were blind in evaluating patients (Item 4 and 5). Other significant errors observed in the included studies were the lack of randomization in the evaluation of the scores or examiners and the short time between repeated measurements. A period of 3-5 minutes was considered insufficient in a previous study, as pain threshold may need more time to restore to baseline[60].

Previous studies have shown that using CPM with CPT at 12°C has better reliability compared to temperatures ranging from 0-10°C. Furthermore, CPT has been reported to be one of the most effective stimuli to induce CPM, especially when

combined with pressure[36], but these conclusions could not be drawn from the current systematic review. On the other hand, environmental conditions could be considered as another topic of discussion. Noise, temperature and humidity should be constant when reliability studies are performed. Only a few studies[48,49,53,54] have reported environmental conditions with an emphasis mainly on noise and temperature.

Although measurement error is an important psychometric property because it informs whether a change in score is real or caused by measurement error, as it turned out, it has not been calculated in an appropriate way (lack of a calculation formula) to provide valid information. No study identified the MIC. Future studies should focus on the evaluation of measurement error and MIC as it is important information for clinical therapists in order to assess intervention effect.

According to the above, a meta-analysis could not be supported due to the variety of methodologies (test sites, stimuli, tools), which did not allow stratification in healthy participants and patients. It was estimated that further stratification could reduce the statistical power and therefore it was avoided.

Future recommendations

Based on the results of the present SR, the authors give some recommendations for future studies. First, blinding of raters to QST and the use of appropriate statistical methods (eg ICC2,1 random effect model for agreement) are important parameters for reliability studies, as referred to in the COSMIN assessment tool[37]. Furthermore, the time interval between repetitions should exceed 15 minutes and factors such as mood, motivation, menstruation, which affect the stability of the participants, but also environmental conditions should be taken into account. Adequate and appropriate training of raters and the use of standardized protocols for evaluating pressure points and stimuli application is imperative. Finally, more studies are needed to assess psychometric properties such as known-group validity, measurement error, minimal significant change, and interrater reliability.

Limitations

This SR included only published studies, so the results may be overestimated (Publication bias). However, this was an agreed decision in order to increase the internal validity of our findings, since grey literature has not undergone a peer-review process and the deriving data and conclusions may be of doubtful quality. Furthermore, the studies were evaluated only in English, thus limiting the total number of studies.

Clinical Impact

The present study demonstrated that the evidence level of the reliability and validity of the QST is at a very early stage. This means that health care practitioners should be very careful when using QST in patient assessment and decision making.

Conclusions

Taking into account the limitations of this systematic review, the reliability seems to be good to excellent for the PPT tests, however, the variation in protocols and the moderate to poor methodological quality of the studies do not allow for extrapolation of clear results. Information on the training of the raters, the stability of the environmental conditions and the blinding of the examiners could significantly improve the quality of the studies and contribute to the generalizability of the results. Further studies are required for the CPM and TS tests in patients with shoulder pain. Assessing the psychometric properties of QST is crucial as it will contribute to the development of appropriate tools for the assessment of patients with chronic musculoskeletal pain and the improvement of the mechanism-based approach to pain management.

Authors' contributions

Study concept and design: P. Bilika, E. Kapreli, N. Strimpakos, A. Paliouras, K. Savvoulidou, A. Arribas-Romano.

Acquisition of data: P. Bilika, A. Arribas-Romano, A. Paliouras, K. Savvoulidou, E. Kapreli, N. Strimpakos.

Data interpretation: E. Kapreli, Z. Dimitriadis, N. Strimpakos, P. Bilika, A. Paliouras, K. Savvoulidou.

Drafting of manuscript and critical revision P. Bilika, E. Kapreli, E. Billis, Z. Dimitriadis, N. Strimpakos, A. Arribas-Romano.

Study Supervision: E. Kapreli, E. Billis, Z. Dimitriadis, N. Strimpakos.

Acknowledgments

The authors would like to thank all the participating experts for their commitment and the time spent, without which this review would not have been possible.

References

1. Longo UG, Facchinetti G, Marchetti A, Candela V, Risi Ambrogioni L, Faldetta A, et al. Sleep Disturbance and Rotator Cuff Tears: A Systematic Review. Medicina (Kaunas) 2019;55(8):453.
2. Minns Lowe CJ, Moser J, Barker K. Living with a symptomatic rotator cuff tear 'bad days, bad nights': a qualitative study. BMC Musculoskeletal Disorders 2014;15:228.
3. Cho CH, Jung SW, Park JY, Song KS, Yu KI. Is shoulder pain for three months or longer correlated with depression, anxiety, and sleep disturbance? Journal of Shoulder and Elbow Surgery 2013;22(2):222-8.
4. Badcock LJ, Lewis M, Hay EM, McCarney R, Croft PR. Chronic shoulder pain in the community: a syndrome of disability or distress? Annals of the rheumatic diseases 2002;61(2):128-31.
5. van der Windt DA, Koes BW, Boeke AJ, Devillé W, De Jong BA, Bouter LM. Shoulder disorders in general practice: prognostic indicators of outcome. The British journal of general practice: the journal of the Royal College of General Practitioners 1996;46(410):519-23.
6. Croft P, Pope D, Silman A. The clinical course of shoulder

pain: prospective cohort study in primary care. BMJ 1996;313(7057):601-2.
7. Luime JJ, Koes BW, Hendriksen IJ, Burdorf A, Verhagen AP, Miedema HS, et al. Prevalence and incidence of shoulder pain in the general population; a systematic review. Scandinavian journal of rheumatology. 2004;33(2):73-81.
8. Mallen CD, Peat G, Thomas E, Dunn KM, Croft PR. Prognostic factors for musculoskeletal pain in primary care: a systematic review. The British journal of general practice : the journal of the Royal College of General Practitioners. 2007;57(541):655-61.
9. Artus M, Campbell P, Mallen CD, Dunn KM, van der Windt DAW. Generic prognostic factors for musculoskeletal pain in primary care: a systematic review. BMJ Open 2017;7(1):e012901.
10. Kuijpers T, van der Windt DAWM, van der Heijden GJMG, Bouter LM. Systematic review of prognostic cohort studies on shoulder disorders. Pain 2004; 109(3):420-31.
11. Hidalgo-Lozano A, Fernandez-de-las-Penas C, Alonso-Blanco C, Ge HY, Arendt-Nielsen L, Arroyo-Morales M. Muscle trigger points and pressure pain hyperalgesia in the shoulder muscles in patients with unilateral shoulder impingement: a blinded, controlled study. Experimental Brain Research 2010;202(4):915-25.
12. Coronado RA, Simon CB, Valencia C, George SZ. Experimental pain responses support peripheral and central sensitization in patients with unilateral shoulder pain. The Clinical Journal of Pain 2014;30(2):143-51.
13. Paul TM, Soo Hoo J, Chae J, Wilson RD. Central hypersensitivity in patients with subacromial impingement syndrome. Archives of Physical Medicine and Rehabilitation 2012;93(12):2206-9.
14. Alburquerque-Sendin F, Camargo PR, Vieira A, Salvini TF. Bilateral myofascial trigger points and pressure pain thresholds in the shoulder muscles in patients with unilateral shoulder impingement syndrome: a blinded, controlled study. The Clinical Journal of Pain 2013;29(6):478-86.
15. Nijs J, George SZ, Clauw DJ, Fernández-de-las-Peñas C, Kosek E, Ickmans K, et al. Central sensitisation in chronic pain conditions: latest discoveries and their potential for precision medicine. The Lancet Rheumatology 2021;3(5):e383-e92.
16. Manion J, Waller MA, Clark T, Massingham JN, Neely GG. Developing Modern Pain Therapies. Front Neurosci 2019;13:1370.
17. Clark J, Nijs J, Yeowell G, Goodwin PC. What Are the Predictors of Altered Central Pain Modulation in Chronic Musculoskeletal Pain Populations? A Systematic Review. Pain Physician 2017;20(6):487-500.
18. Clark JR, Nijs J, Yeowell G, Holmes P, Goodwin PC. Trait Sensitivity, Anxiety, and Personality Are Predictive of Central Sensitization Symptoms in Patients with Chronic Low Back Pain. Pain practice: the official journal of World Institute of Pain 2019;19(8):800-10.
19. Nijs J, Polli A, Willaert W, Malfliet A, Huysmans E, Coppieters I. Central sensitisation: another label or useful diagnosis? Drug Ther Bull 2019;57(4):60-3.
20. Nijs J, Leysen L, Vanlauwe J, Logghe T, Ickmans K, Polli A, et al. Treatment of central sensitization in patients with chronic pain: time for change? Expert Opinion on Pharmacotherapy 2019;20(16):1961-70.
21. Pennella D, Giagio S, Maselli F, Giovannico G, Roncone A, Fiorentino F, et al. Red flags useful to screen for gastrointestinal and hepatic diseases in patients with shoulder pain: A scoping review. Musculoskeletal Care 2022.
22. Ruscheweyha R, Marziniaka, M, Stumpenhorsta F, Reinholza J, Knechta S. Pain sensitivity can be assessed by self-rating: Development and validation of the Pain Sensitivity Questionnaire. Pain 2009;146(1-2):65-74.
23. Treede RD. The role of quantitative sensory testing in the prediction of chronic pain. Pain 2019;160 Suppl 1:S66-s9.
24. Arendt-Nielsen L. Central sensitization in humans: assessment and pharmacology. Handbook of Experimental Pharmacology 2015;227:79-102.
25. Arendt-Nielsen L, Morlion B, Perrot S, Dahan A, Dickenson A, Kress HG, et al. Assessment and manifestation of central sensitisation across different chronic pain conditions. Eur J Pain 2018;22(2):216-41.
26. Chong PST, Cros DP. Technology literature review: Quantitative sensory testing. Muscle Nerve 2004; 29(5):734-47.
27. Uddin Z, MacDermid JC. Quantitative Sensory Testing in Chronic Musculoskeletal Pain. Pain Medicine (Malden, Mass) 2016;17(9):1694-703.
28. Hallin RG, Torebjörk HE, Wiesenfeld Z. Nociceptors and warm receptors innervated by C fibres in human skin. Journal of Neurology, Neurosurgery, and Psychiatry 1982;45(4):313-9.
29. Perl ER. Myelinated afferent fibres innervating the primate skin and their response to noxious stimuli. The Journal of Physiology 1968;197(3):593-615.
30. Sander HW. Sensory Testing, Quantitative. In: Aminoff MJ, Daroff RB, editors. Encyclopedia of the Neurological Sciences (Second Edition). Oxford: Academic Press; 2014.
31. Arendt-Nielsen L, Yarnitsky D. Experimental and clinical applications of quantitative sensory testing applied to skin, muscles and viscera. The journal of pain: official journal of the American Pain Society 2009;10(6):556-72.
32. Mackey IG, Dixon EA, Johnson K, Kong JT. Dynamic Quantitative Sensory Testing to Characterize Central Pain Processing. Journal of visualized experiments: JoVE 2017;16(120):54452.
33. Georgopoulos V, Akin-Akinyosoye K, Zhang W, McWilliams DF, Hendrick P, Walsh DA. Quantitative sensory testing and predicting outcomes for musculoskeletal pain, disability, and negative affect: a systematic review and meta-analysis. Pain 2019;160(9):1920-32.

34. Braun M, Bello C, Riva T, Hönemann C, Doll D, Urman RD, et al. Quantitative Sensory Testing to Predict Postoperative Pain. Current Pain and Headache Reports 2021;25(1):3.
35. Gwilym SE, Oag HC, Tracey I, Carr AJ. Evidence that central sensitisation is present in patients with shoulder impingement syndrome and influences the outcome after surgery. The Journal of Bone and Joint Surgery British Volume 2011;93(4):498-502.
36. Nuwailati R, Bobos P, Drangsholt M, Curatolo M. Reliability of conditioned pain modulation in healthy individuals and chronic pain patients: a systematic review and meta-analysis. Scandinavian Journal of Pain 2022;22(2):262-278.
37. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021;372:n160.
38. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Medical Research Methodology 2020;20(1):293.
39. Bilika P, Savvoulidou K, Paliouras A, Dimitriadis Z, Billis E, Strimpakos N, et al. Psychometric properties of quantitative sensory testing focusing on healthy and patients with shoulder pain: a systematic review protocol. International Journal of Clinical Trials 2021;8(3):7
40. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica 2012;22(3):276-82.
41. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation. 2018;27(5):1147-57.
42. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation 2018;27(5):1171-9.
43. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation 2018;27(5):1159-70.
44. De Groef A, Van Kampen M, Vervloesem N, Clabau E, Christiaens MR, Neven P, et al. Inter-rater reliability of shoulder measurements in middle-aged women. Physiotherapy 2017;103(2):222-30.
45. Nascimento J, Alburquerque-Sendín F, Vigolvino LP, Oliveira WF, Sousa CO. Absolute and Relative Reliability of Pressure Pain Threshold Assessments in the Shoulder Muscles of Participants With and Without Unilateral Subacromial Impingement Syndrome. Journal of Manipulative and Physiological Therapeutics 2020;43(1):57-67.
46. Wang-Price S, Zafereo J, Brizzolara K, Mackin B, Lawson L, Seeger D, et al. Psychometric Properties of Pressure Pain Thresholds Measured in 2 Positions for Adults With and Without Neck-Shoulder Pain and Tenderness. Journal of Manipulative and Physiological Therapeutics 2019;42(6):416-24.
47. Vaegter HB, Lyng KD, Yttereng FW, Christensen MH, Sørensen MB, Graven-Nielsen T. Exercise-Induced Hypoalgesia After Isometric Wall Squat Exercise: A Test-Retest Reliabilty Study. Pain medicine (Malden, Mass) 2019;20(1):129-37.
48. Persson AL, Brogårdh C, Sjölund BH. Tender or not tender: test-retest repeatability of pressure pain thresholds in the trapezius and deltoid muscles of healthy women. Journal of Rehabilitation Medicine 2004;36(1):17-27.
49. Vanderweeën L, Oostendorp RA, Vaes P, Duquet W. Pressure algometry in manual therapy. Manual Therapy 1996;1(5):258-65.
50. Levoska S, Keinänen-Kiukaanniemi S, Bloigu R. Repeatability of measurement of tenderness in the neck-shoulder region by a dolorimeter and manual palpation. The Clinical Journal of Pain 1993;9(4):229-35.
51. Jones DH, Kilgour RD, Comtois AS. Test-retest reliability of pressure pain threshold measurements of the upper limb and torso in young healthy women. The journal of pain: official journal of the American Pain Society 2007;8(8):650-6.
52. Valencia C, Kindler LL, Fillingim RB, George SZ. Stability of conditioned pain modulation in two musculoskeletal pain models: investigating the influence of shoulder pain intensity and gender. BMC Musculoskeletal Disorders 2013;14:182.
53. Cathcart S, Winefield AH, Rolan P, Lushington K. Reliability of temporal summation and diffuse noxious inhibitory control. Pain Research & Management 2009;14(6):433-8.
54. Marcuzzi A, Wrigley PJ, Dean CM, Adams R, Hush JM. The long-term reliability of static and dynamic quantitative sensory testing in healthy individuals. Pain 2017;158(7):1217-23.
55. Alsouhibani A, Vaegter HB, Hoeger Bement M. Systemic Exercise-Induced Hypoalgesia Following Isometric Exercise Reduces Conditioned Pain Modulation. Pain medicine (Malden, Mass) 2019;20(1):180-90.
56. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. Annals of Internal Medicine 2009;151(4):264-9.
57. Siao P, Cros DP. Quantitative sensory testing. Physical medicine and rehabilitation clinics of North America 2003;14(2):261-86.
58. Yarnitsky D, Bouhassira D, Drewes AM, Fillingim RB,

Granot M, Hansson P, et al. Recommendations on practice of conditioned pain modulation (CPM) testing. European journal of pain (London, England) 2015;19(6):805-6.
59. Moloney NA, Hall TM, Doody CM. Reliability of thermal quantitative sensory testing: a systematic review. Journal of Rehabilitation Research and Development 2012;49(2):191-207.
60. Lewis GN, Rice DA, McNair PJ. Conditioned pain modulation in populations with chronic pain: a systematic review and meta-analysis. The journal of pain: official journal of the American Pain Society 2012;13(10):936-44.

Appendix 1. Search strategy for each database (31 August 2021).

PubMed Search Formula vía NLM. Results: 334
("Psychometrics"[Mesh] OR "Reproducibility of Results"[Mesh] OR "observer variation"[Mesh] OR psychometr* OR clinimetr* OR valid* OR "content validity" OR "construct validity" OR "structural validity" OR hypothesis-testing OR "criterion validity" OR "predictive validity" OR sensitivity OR specificity OR "measurement error" OR reliab* OR "intratester reliability" OR "intertester reliability" OR "test-retest reliability" OR "absolute reliability" OR responsiveness OR "standard error of measurement" OR reproducibility OR test-retest OR agreement OR "smallest real difference" OR "minimal detectable change" OR "minimal important difference" OR "clinically important difference" OR "meaningful change" OR repeatability OR accuracy OR precision OR consistency OR stability) AND ("Central Nervous System Sensitization"[Mesh] OR "Pain Threshold"[Mesh] OR "Sensory Thresholds"[Mesh] OR QST OR "quantitative sensory testing" OR hyperalgesia OR hyperaesthesia OR allodynia OR "pressure pain threshold" OR "temporal summation" OR "conditioned pain modulation" OR PPT OR CPM OR "electrical perception threshold" OR "heat pain threshold" OR "wind up") AND ("Shoulder"[Mesh] OR "Shoulder Pain"[Mesh] OR shoulder*)
MEDLINE Complete Search Formula (EBSCO). Results: 116
("Psychometrics"[Mesh] OR "Reproducibility of Results"[Mesh] OR "observer variation"[Mesh] OR psychometr* OR clinimetr* OR valid* OR "content validity" OR "construct validity" OR "structural validity" OR hypothesis-testing OR "criterion validity" OR "predictive validity" OR sensitivity OR specificity OR "measurement error" OR reliab* OR "intratester reliability" OR "intertester reliability" OR "test-retest reliability" OR "absolute reliability" OR responsiveness OR "standard error of measurement" OR reproducibility OR test-retest OR agreement OR "smallest real difference" OR "minimal detectable change" OR "minimal important difference" OR "clinically important difference" OR "meaningful change" OR repeatability OR accuracy OR precision OR consistency OR stability) AND (¨Central Nervous System Sensitization¨[Mesh] OR "Pain Threshold"[Mesh] OR "Sensory Thresholds"[Mesh] OR QST OR "quantitative sensory testing" OR hyperalgesia OR hyperaesthesia OR allodynia OR "pressure pain threshold" OR "temporal summation" OR "conditioned pain modulation" OR PPT OR CPM OR "electrical perception threshold" OR "heat pain threshold" OR "wind up") AND ("Shoulder"[Mesh] OR "Shoulder Pain"[Mesh] OR shoulder*)
SPORTDiscus Search Formula (EBSCO). Results: 30
("Psychometrics"[Mesh] OR "Reproducibility of Results"[Mesh] OR "observer variation"[Mesh] OR psychometr* OR clinimetr* OR valid* OR "content validity" OR "construct validity" OR "structural validity" OR hypothesis-testing OR "criterion validity" OR "predictive validity" OR sensitivity OR specificity OR "measurement error" OR reliab* OR "intratester reliability" OR "intertester reliability" OR "test-retest reliability" OR "absolute reliability" OR responsiveness OR "standard error of measurement" OR reproducibility OR test-retest OR agreement OR "smallest real difference" OR "minimal detectable change" OR "minimal important difference" OR "clinically important difference" OR "meaningful change" OR repeatability OR accuracy OR precision OR consistency OR stability) AND ("Central Nervous System Sensitization"[Mesh] OR "Pain Threshold"[Mesh] OR "Sensory Thresholds"[Mesh] OR QST OR "quantitative sensory testing" OR hyperalgesia OR hyperaesthesia OR allodynia OR "pressure pain threshold" OR "temporal summation" OR "conditioned pain modulation" OR PPT OR CPM OR "electrical perception threshold" OR "heat pain threshold" OR "wind up") AND ("Shoulder"[Mesh] OR "Shoulder Pain"[Mesh] OR shoulder*)
Cochrane Library Search Formula. Results: 70
ID Search Hits #1 MeSH descriptor: [Psychometrics] explode all trees 2877 #2 MeSH descriptor: [Reproducibility of Results] explode all trees 11017 #3 MeSH descriptor: [Observer Variation] explode all trees 1953 #4 (psychometr* OR clinimetr* OR valid* OR "content validity" OR "construct validity" OR "structural validity" OR hypothesis-testing OR "criterion validity" OR "predictive validity" OR sensitivity OR specificity OR "measurement error" OR reliab* OR "intratester reliability" OR "intertester reliability" OR "test-retest reliability" OR "absolute reliability" OR responsiveness OR "standard error of measurement" OR reproducibility OR test-retest OR agreement OR "smallest real difference" OR "minimal detectable change" OR "minimal important difference" OR "clinically important difference" OR "meaningful change" OR repeatability OR accuracy OR precision OR consistency OR stability):ti,ab,kw 190842 #5 #1 OR #2 OR #3 OR #4 191429 #6 MeSH descriptor: [Central Nervous System Sensitization] explode all trees 32 #7 MeSH descriptor: [Pain Threshold] explode all trees 1744 #8 MeSH descriptor: [Sensory Thresholds] explode all trees 3191 #9 (QST OR "quantitative sensory testing" OR hyperalgesia OR hyperaesthesia OR allodynia OR "pressure pain threshold" OR "temporal summation" OR "conditioned pain modulation" OR PPT OR CPM OR "electrical perception threshold" OR "heat pain threshold" OR "wind up"):ti,ab,kw 5364 #10 #6 OR #7 OR #8 OR #9 7740 #11 MeSH descriptor: [Shoulder] explode all trees 604 #12 MeSH descriptor: [Shoulder Pain] explode all trees 1011 #13 (shoulder*):ti,ab,kw 13008 #14 #11 OR #12 OR #13 13008 #15 #5 AND #10 AND #14 70

Scopus Search Formula vía ELSEVIER: Results: 257
TITLE-ABS-KEY ("Psychometrics" OR "Reproducibility of Results" OR "observer variation" OR psychometr* OR clinimetr* OR valid* OR "content validity" OR "construct validity" OR "structural validity" OR hypothesis-testing OR "criterion validity" OR "predictive validity" OR sensitivity OR specificity OR "measurement error" OR reliab* OR "intratester reliability" OR "intertester reliability" OR "test-retest reliability" OR "absolute reliability" OR responsiveness OR "standard error of measurement" OR reproducibility OR test-retest OR agreement OR "smallest real difference" OR "minimal detectable change" OR "minimal important difference" OR "clinically important difference" OR "meaningful change" OR repeatability OR accuracy OR precision OR consistency OR stability) AND TITLE-ABS-KEY ("Central Nervous System Sensitization" OR "Pain Threshold" OR "Sensory Thresholds" OR qst OR "quantitative sensory testing" OR hyperalgesia OR hyperaesthesia OR allodynia OR "pressure pain threshold" OR "temporal summation" OR "conditioned pain modulation" OR ppt OR cpm OR "electrical perception threshold" OR "heat pain threshold" OR "wind up") AND TITLE-ABS-KEY ("Shoulder" OR "Shoulder Pain" OR shoulder*)
Wos Core Collection Searh Formula vía ELSEVIER: Results: 382
#1 TS=("Psychometrics" OR "Reproducibility of Results" OR "observer variation" OR psychometr* OR clinimetr* OR valid* OR "content validity" OR "construct validity" OR "structural validity" OR hypothesis-testing OR "criterion validity" OR "predictive validity" OR sensitivity OR specificity OR "measurement error" OR reliab* OR "intratester reliability" OR "intertester reliability" OR "test-retest reliability" OR "absolute reliability" OR responsiveness OR "standard error of measurement" OR reproducibility OR test-retest OR agreement OR "smallest real difference" OR "minimal detectable change" OR "minimal important difference" OR "clinically important difference" OR "meaningful change" OR repeatability OR accuracy OR precision OR consistency OR stability) #2 TS=(“Central Nervous System Sensitization” OR “Pain Threshold” OR “Sensory Thresholds” OR qst OR “quantitative sensory testing” OR hyperalgesia OR hyperaesthesia OR allodynia OR “pressure pain threshold” OR “temporal summation” OR “conditioned pain modulation” OR ppt OR cpm OR “electrical perception threshold” OR “heat pain threshold” OR “wind up”) #3 TS=(“Shoulder” OR “Shoulder Pain” OR shoulder*) # 4 #1 AND #2 AND #3
EMBASE Search Formula: Results: 238
No. **Query** **Results** **238** **#4** **#1** AND **#2** AND **#3**
124,453 **#3** '**shoulder**' OR '**shoulder pain**' OR **shoulder***
77,793 **#2** '**central nervous system sensitization**' OR '**pain threshold**' OR '**sensory thresholds**' OR **qst** OR '**quantitative sensory testing**' OR **hyperalgesia** OR **hyperaesthesia** OR **allodynia** OR '**pressure pain threshold**' OR '**temporal summation**' OR '**conditioned pain modulation**' OR **ppt** OR **cpm** OR '**electrical perception threshold**' OR '**heat pain threshold**' OR '**wind up**'
5,149,608 **#1** '**psychometrics**'/exp OR '**psychometrics**' OR '**reproducibility of results**'/exp OR '**reproducibility of results**' OR '**observer variation**'/exp OR '**observer variation**' OR **psychometr*** OR **clinimetr*** OR **valid*** OR '**content validity**'/exp OR '**content validity**' OR '**construct validity**'/exp OR '**construct validity**' OR '**structural validity**'/exp OR '**structural validity**' OR '**hypothesis testing**'/exp OR '**hypothesis testing**' OR '**criterion validity**'/exp OR '**criterion validity**' OR '**predictive validity**'/exp OR '**predictive validity**' OR '**sensitivity**'/exp OR sensitivity OR '**specificity**'/exp OR **specificity** OR '**measurement error**'/exp OR '**measurement error**' OR **reliab*** OR '**intratester reliability**' OR '**intertester reliability**'/exp OR '**intertester reliability**' OR '**test-retest reliability**'/exp OR '**test-retest reliability**' OR '**absolute reliability**' OR '**responsiveness**'/exp OR **responsiveness** OR '**standard error of measurement**'/exp OR '**standard error of measurement**' OR '**reproducibility**'/exp OR **reproducibility** OR '**test retest**' OR '**agreement**'/exp OR **agreement** OR '**smallest real difference**' OR '**minimal detectable change**'/exp OR '**minimal detectable change**' OR '**minimal important difference**'/exp OR '**minimal important difference**' OR '**clinically important difference**' OR '**meaningful change**' OR '**repeatability**'/exp OR **repeatability** OR '**accuracy**'/exp OR **accuracy** OR '**precision**'/exp OR **precision** OR '**consistency**'/exp OR **consistency** OR '**stability**'/exp OR **stability.**

PEDro Search Formula. Results: 263	
#1 Psychometrics Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: No records found **#2 Reproducibility of Results** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 2 **#3 observer variation** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 1 **#4 psychometr*** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 1 **#5 clinimetr*** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#6 valid*** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 77 **#7 content validity** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#8 construct validity** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#9 structural validity** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#10 hypothesis-testing** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 3 **#11 criterion validity** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#12 predictive validity** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#13 sensitivity** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 60 **#14 specificity** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 10 **#15 measurement error** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0	**#16 intratester reliability** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#17 intertester reliability** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#18 test-retest reliability** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 2 **#19 absolute reliability** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#20 responsiveness** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 3 **#21 standard error of measurement** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#22 reproducibility** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 2 **#23 test-retest** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 4 **#24** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: **#25 agreement** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 17 **#26 smallest real difference** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 0 **#27 minimal detectable change** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 4 **#28 minimal important difference** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 16 **#29 clinically important difference** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 32 **#30 meaningful change** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 12

#31 repeatability Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results: 1 **#32 accuracy** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results:20 **#33 precision** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results:12	**#34 Consistency** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results:2 **#35 Stability** Body part: upper arm, shoulder and shoulder girdle Match all search terms (AND) Results:29

Appendix 2. COSMIN Risk of Bias for PPT studies. *(Ratings: (V) Very good; (A) Adequate; (D) Doubtful; (I) Inadequate; N/A R1: Rater 1 R2: Rater 2 C: Consensus).*

Intrarater Reliability	Study 1 De Groef et al. 2016			Study 2 Nascimento et al. 2019			Study 3 Wang-Price et al. 2019			Study 4 Vaegter et al. 2018			Study 5 Jones et al. 2007			Study 6 Persson et al. 2004			Study 7 Vanderweeën et al. 1996			Study 8 Levoska et al. 1993		
Design requirements				R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C
1 Stability of the patients				D	D	D	D	D	D	**D**	**D**	**D**	V	V	V	**D**	**D**	**D**	V	V	V	**D**	**D**	**D**
2 Time interval				D	D	D	V	V	V	V	V	V	V	V	V	V	V	V	**D**	**D**	**D**	V	V	V
3 Similarity of measurement condition				D	D	D	D	D	D	V	V	V	D	D	D	V	V	V	V	V	V	D	D	D
4 Administation without knowledge of scores or values				V	V	V	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
5 Score assignment or determination without knowledge of the scores or values				V	V	V	D	D	D	D	D	D	I	I	I	D	D	D	D	D	D	D	D	D
6 Other important flaws				V	V	V	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D
Statistical Methods																								
7 For continuous scores: ICC				V	V	V	A	D	D	A	A	A	A	A	A	V	V	V	A	A	A	A	A	A
8 For ordinal scores: Kappa				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9 For dichotomous/nominal scores: Kappa for each category against the other categories				-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FINAL RATING (Lowest score of items)				**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**I**	**I**	**I**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**

Interrater Reliability	Study 1 De Groef et al. 2016			Study 2 Nascimento et al. 2019			Study 6 Persson et al. 2004			Study 8 Levoska et al. 1993		
Design requirements	R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C
1 Stability of the patients	A	A	A	A	**D**	**D**	D	D	D	**D**	**D**	**D**
2 Time interval	**D**	**D**	**D**	V	V	V	V	V	V	D	D	D
3 Similarity of measurement condition	D	D	D	**D**	D	D	V	V	V	D	D	D
4 Administation without knowledge of scores or values	V	V	V	V	V	V	D	D	D	D	D	D
5 Score assignment or determination without knowledge of the scores or values	V	V	V	V	V	V	D	D	D	D	D	D
6 Other important flaws	D	D	D	V	V	V	D	D	D	D	D	D
Statistical Methods												
7 For continuous scores: ICC	V	V	V	V	V	V	**I**	**I**	**I**	A	A	A
8 For ordinal scores: Kappa	-	-	-	-	-	-	-	-	-	-	-	-
9 For dichotomous/nominal scores: Kappa for each category against the other categories	-	-	-	-	-	-	-	-	-	-	-	-
FINAL RATING (Lowest score of items)	**D**	**D**	**D**	**D**	**D**	**D**	**I**	**I**	**I**	**D**	**D**	**D**

Intrarater Reliability	Study 1 De Groef et al. 2016			Study 2 Nascimento et al. 2019			Study 3 Wang-Price et al. 2019			Study 4 Vaegter et al. 2018			Study 6 Persson et al. 2004		
Design requirements	R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C
1 Stability of the patients	A	A	A	A	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**
2 Time interval	**D**	**D**	**D**	V	V	V	V	V	V	V	V	V	V	V	V
3 Similarity of measurement condition	D	D	D	**D**	D	D	D	D	D	V	V	V	V	V	V
4 Administation without knowledge of scores or values	V	V	V	V	V	V	D	D	D	D	D	D	D	D	D
5 Score assignment or determination without knowledge of the scores or values	V	V	V	V	V	V	D	D	D	D	D	D	D	D	D
6 Other important flaws	D	D	D	V	V	V	D	D	D	D	D	D	D	D	D
Statistical Methods															
7 For continuous scores: SEM, SDC, LoA or CV calculated?	V	V	V	V	V	V	V	V	V	A	A	A	D	A	A
8 For dichotomous/nominal/ ordinal scores: Percentage specific (e.g. positive and negative) agreement	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FINAL RATING (Lowest score of items)	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**

Hypotheses testing for construct validity (comparison between subgrous)	Study 3 Wang-Price et al. 2019		
Design requirements	R 1	R2	C
1 Adequate description of important characteristcs of the subgroups	**V**	**V**	**V**
Statistical Methods			
2 Appropriate statistical method for the hypothesis to be tested	V	V	V
3 Other important flaws	V	V	V
FINAL RATING (Lowest score of items)	**V**	**V**	**V**

COSMIN Risk of Bias for CPM and TS studies. *(Ratings: (V) Very good; (A) Adequate; (D) Doubtful; (I) Inadequate; N/A R1: Rater 1 R2: Rater 2 C: Consensus).*

Intrarater Reliability	**Study 9** Valencia et al. 2013			**Study 10** Cathcart et al. 2009			**Study 11** Alsouhibani et al. 2019			**Study 12** Marcuzzi et al. 2017		
Design requirements	R 1	R2	C	R 1	R2	C	R 1	R2	C	R 1	R2	C
1 Stability of the patients	A	A	A	A	A	A	**D**	**D**	**D**	**D**	**D**	**D**
2 Time interval	**D**	**D**	**D**	V	V	V	V	V	V	V	V	V
3 Similarity of measurement condition	D	D	D	V	V	V	D	D	D	V	V	V
4 Administation without knowledge of scores or values	D	D	D	**D**	**D**	**D**	D	D	D	V	V	V
5 Score assignment or determination without knowledge of the scores or values	D	D	D	D	D	D	D	D	D	V	V	V
6 Other important flaws	D	D	D	D	D	D	D	D	D	D	D	D
Statistical Methods												
7 For continuous scores: ICC	A	A	A	V	V	V	A	A	A	V	V	V
8 For ordinal scores: Kappa	-	-	-	-	-	-	-	-	-	-	-	-
9 For dichotomous/nominal scores: Kappa for each category against the other categories	-	-	-	-	-	-	-	-	-	-	-	-
FINAL RATING (Lowest score of items)	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**	**D**
Interrater Reliability												
Design requirements												
1 Stability of the patients												
2 Time interval												
3 Similarity of measurement condition												
4 Administation without knowledge of scores or values												
5 Score assignment or determination without knowledge of the scores or values												
6 Other important flaws												
Statistical Methods												
7 For continuous scores: ICC												
8 For ordinal scores: Kappa												
9 For dichotomous/nominal scores: Kappa for each category against the other categories												
FINAL RATING (Lowest score of items)												
Measurement Error										**Study 12** Marcuzzi et al. 2017		
Design requirements										R1	R2	C
1 Stability of the patients										**D**	**D**	**D**
2 Time interval										V	V	V
3 Similarity of measurement condition										V	V	V
4 Administation without knowledge of scores or values										V	V	V
5 Score assignment or determination without knowledge of the scores or values										V	V	V
6 Other important flaws										D	D	D
Statistical Methods												
7 For continuous scores: SEM, SDC, LoA or CV calculated?										V	V	V
8 For dichotomous/nominal/ordinal scores: Percentage specific (e.g. positive and negative) agreement										-	-	-
FINAL RATING (Lowest score of items)										**D**	**D**	**D**