8-21-2019 9:00 AM

Psychometric Properties of the Brief Pain Inventory-Short Form and Revised Short McGill Pain Questionnaire Version-2 in Musculoskeletal Conditions

Samuel Ugochukwu Jumbo, *The University of Western Ontario*

Supervisor: MacDermid, Joy C., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Health and Rehabilitation Sciences

ABSTRACT

Introduction: Comprehensive pain assessment depends on the use of psychometrically valid patient-reported outcome measures (PROMs). The Brief Pain Inventory-Short Form (BPI-SF) and Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) are general-use multidimensional pain assessment tools commonly used in musculoskeletal conditions. Understanding all relevant measurement properties supports stronger decisions about PROMs.

Thesis Objectives: The overarching objective of this thesis was to determine the sufficiency of measurement evidence backing the use of the BPI-SF and SF-MPQ-2 in musculoskeletal conditions. Specifically, a systematic review was conducted to locate, summarize and compare the quality and content of psychometric evidence backing the BPI-SF and SF-MPQ-2 in musculoskeletal conditions. Based on this review, the gap in evidence regarding the reliability and agreement properties (reproducibility) of SF-MPQ-2 was examined among patients with musculoskeletal shoulder pain.

Methods: For the systematic review, we searched four databases to identify relevant citations. Two reviewers independently screened, extracted and appraised (using MacDermid and COSMIN guidelines) all psychometric reports on both tools in musculoskeletal conditions. To determine the SF-MPQ-2 reproducibility, a convenience sample of adults diagnosed with musculoskeletal shoulder pain (baseline, n=195; test-retest, n=48) completed the SF-MPQ-2 twice. Cronbach alpha (α), intraclass correlations coefficient ($ICC_{2,1}$), agreement parameters (SEM, MDC) and Bland-Altman plots were assessed.

Results: High quality evidence indicated both tools have high internal consistency (α = 0.83-0.96); and that they are moderately related (r = 0.3-0.69) to other health-related outcome measures. More studies of better quality have evaluated the BPI-SF responsiveness (n=5), retest reliability (n=3), known group validity (n=2) and structural validity (n=3), compared to the SF-MPQ-2. Our analysis of the SF-MPQ-2 reproducibility established internal consistency as satisfactory (α, 0.83-0.95), relative reliability as good (neuropathic, intermittent, and affective subscales: 1CC$_{2,1}$= 0.78 - 0.88) to excellent (total and continuous subscale scores: 1CC$_{2,1}$= 0.92 - 0.95). Agreement was within acceptable limits and there was no evidence of systematic bias.

Conclusion: A greater volume of evidence of better quality currently supports the BPI-SF although emerging evidence suggest the SF-MPQ-2 has excellent reliability and agreement properties when used to assess adults with musculoskeletal shoulder pain. Direct comparisons of the two scales in different contexts are needed.

Keywords: Brief Pain Inventory; Musculoskeletal Conditions; McGill Pain Questionnaire; Reliability; Psychometric Properties; Reproducibility; Systematic Review

LAY ABSTRACT

What is the problem? Musculoskeletal (MSK) refers to anything related to our muscles, tendons, joints and connective tissue. Pain that comes from any of these tissues is called MSK pain. MSK pain is one of the most common reasons people seek treatment from a doctor or therapist, so the better we understand this pain, the better decisions we can make about treatment. An important way of measuring MSK pain is by asking the person to give ratings for different aspects of their pain using tools called patient-reported outcome measures (or PROMS for short). Health care providers use PROMs for pain assessment because they are simple and affordable. But more important, they give accurate scores that help monitor treatment progress from the person's own view. Researchers and health care providers need to know which tools are best for MSK pain, especially when they are used to assess more than one condition, like fractures and tendonitis.

Study question: The key question in my thesis work was: is there enough good evidence that researchers and health care providers can feel confident using the Brief Pain Inventory Short-Form (BPI-SF) and Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) for measuring MSK pain?

How did we study the problem? We did a careful search of online libraries of health science research to find every study we could that told us about both tools. We recorded all the key information about how well they measured MSK pain. Then, we assessed and compared the quality of these studies, so we knew what information was best. In our second study, we checked if the SF-MPQ-2 gave us the same (reliable) scores when it was used by patients coming to see a doctor because of shoulder pain at two different times.

What did we learn? After all these studies, we concluded that the BPI-SF currently has more good quality evidence backing its use in MSK condition than the SF-MPQ-2. We are also

confident that SF-MPQ-2 will probably be a good tool for measuring MSK shoulder pain since it yields consistent scores from our evaluation.

What do we still not know? Researchers and health care providers who want to use these tools should be aware that they are not yet the ‘gold standard’. More research is needed to confirm some of their measurement properties in different kinds of MSK pain problems.

AKNOWLEDGEMENT

I sincerely glorify God, the source of inspiration, sound health, wisdom, and divine understanding, for the strength and unshaken determination to complete this project. Without you, I am nothing!

Words to thank my amazing supervisor and mentor, Dr. Joy MacDermid, are difficult to find. I am sincerely grateful for your unlimited support, dedication, ‘simplicity’, patience; for allowing me to explore yet guiding me when I go astray; for being easy to talk to, for your sense of humility, and for believing in my capabilities. You are an inspiring leader and researcher, and I am fortunate to be your student.

I would also like to appreciate my thesis advisory committee members, Dr. Tara Packham, Dr. George Athwal and Dr. Kenneth Faber for their commitment, time and dedication towards this project. Your advice and assistance during the data collection phase allowed me to reach a target sample size in good time. Also, your thoughtful feedback, input and corrections improved the quality of the work. Indeed, I am very fortunate to have learned from you all.

Thank you for your words of encouragement, unconditional love, support, prayers, and for being you, Inge Gajewski Jumbo, my best friend and wife. You were always there when the going got tough. Even when you did not clearly understand what I was saying or my frustrations, you listened, you encouraged me, and somehow, strength comes. We did it together!

Finally, my family and friends are deeply appreciated for their unwavering support, prayers, listening ears, and advices. Michael Kalu, thank you for forbearing and inspiring me to reach greater heights. This thesis is dedicated to my late Mother, Elder Mrs. Peace Uloaku Jumbo, for her unquantifiable love, sacrifices and for teaching me the fear of God. You did

not see the four walls of the university but sacrificed everything for us to be educated. Your memories live on!

CONTRIBUTIONS

Samuel Ugochukwu Jumbo, the Master of Science candidate primarily authored and lead this thesis from September 2017 to August 2019. He completed a systematic review study protocol for an independent study course, a systematic review; designed a protocol for the second thesis reproducibility study, recruited participants, collected data, analyzed and interpreted the data, drafted the manuscripts and incorporated committee members feedback.

Dr. Joy C. MacDermid supervised the thesis and provided funding for the study. Samuel and Dr. Joy MacDermid reviewed and refined the research question and study design. Dr Joy MacDermid reviewed the manuscripts. She also provided countless important feedback in office hours, student meetings and via e-mail correspondences at all stages of the thesis including at committee meetings.

Dr Tara Packham, who served as a supervisory committee member, provided support, guidance and feedback on the thesis manuscripts both in committee meetings and via e-mail correspondences. Dr. Kenneth Faber also served as a supervisory committee member and provided important feedback in committee meetings, assisted with the recruitment of participants for the second thesis manuscript, and reviewed the thesis manuscripts. Dr George Athwal was a supervisory committee member and assisted with participants recruitment and review of the thesis manuscripts.

TABLE OF CONTENTS

Chapter 4. Discussion

CHAPTER 1: Literature Review

Overview of Musculoskeletal Disorders

‘Musculoskeletal disorders’ (MSDs), as a term, describes a vast number of inflammatory and degenerative conditions affecting the muscles, tendons, ligaments, joints, peripheral nerves, and supporting blood vessels (1). More than 291 pathologies (2) have been defined as MSDs, and may include conditions with a) unambiguous pathophysiology such as, tendon inflammations (e.g. tenosynovitis, epicondylitis, bursitis), nerve compression disorders (e.g. carpal tunnel syndrome, sciatica), and osteoarthritis, or b) conditions with ambiguous or less standardized pathophysiology (e.g. myalgia, low back pain) and generalized body pain of unknown cause (1,3). MSDs predominantly affect the low back, neck, shoulder, forearm, hand, and the lower extremity (1,4).

MSDs are the most common cause of long-term pain and disability (3,5,6). MSDs impact negatively on an individual’s level of participation, quality of life, social, psychological and economic well-being (6,7). The prevalence of MSDs is high and is expected to continue to increase for several reasons including greater rates of obesity, sedentary lifestyle, and the growing ageing population (3). MSDs currently account for 21.3% of the total years lived with disability (YLDs) and globally represent the 4th largest health burden (8). Prevalence rates are higher in developed countries than in developing countries. For instance, at any one time, joint pain, swelling, or limitation of movement will affect no less than 30% of American adults in their life time and represent the leading cause of disability among adults within or below 45-year-old (6). In Ontario Canada, MSDs account for 40% of all chronic conditions, 54% of all long-term disabilities and 24% of all restricted activity days (9). MSDs such as rheumatoid arthritis, osteoarthritis, osteoporosis, spinal disorders and major limb traumas come with the greatest financial consequence on the individual and society (8). In Canada, treatment and management of MSDs directly accounts for 7 billion dollars in expenditures, including cost of research, hospital bills, medical services and professional bills, while MSDs indirectly cost 25 billion dollars from loss to disability/profitable work hours and premature death (7). Recent reports for the United States indicate MSDs cost over 125 billion dollars per annum directly and indirectly. Indeed, MSDs are pervasive burdens with influence reaching all ages, walks of life, countries and regions.

Overview of Shoulder Pain

Shoulder pain, the third most common musculoskeletal complain after back and neck pain, originates from different problems affecting the shoulder structures (10–12). As the most mobile joint of the body, the shoulder is at high risk of instability and pain. Also, the shoulder links the upper extremity to the thorax; hence, tissues including muscles, tendons, and major neurovascular structures surrounding the shoulder indirectly become potential sources of referred pain (12). Examples of conditions affecting the shoulder directly include: (a) Rotator Cuff Disorders (RCDs): a group of disorders including rotator cuff tendinopathy, impingement, sub-acromial bursitis, rotator cuff tears; (b) Glenohumeral Disorders: capsulitis (“frozen shoulder”), arthritis; (c) Acromioclavicular Diseases; (d) Infection; and (e) Traumatic Dislocation. Of these conditions, RCDs are the most common pathology affecting the shoulder joint. Other conditions that can affect the shoulder indirectly include: (a) Neck Pain; (b) Myocardial Ischemia; (c) Referred Diaphragmatic Pain; (d) Polymyalgia Rheumatica; and (e) Malignancy i.e. apical lung cancers or metastases (11).

A review of shoulder pain/complaints prevalence studies till the year 2001(13) noted substantial variation in ranges across study reports: point prevalence ranged from 7-27% (adults > 70 years) and 13.2 – 26% (adults < 70years). The annual prevalence of shoulder pain/complaints ranged from 5 – 47% (13); another review estimated prevalence of shoulder complaints at 50% in the general population (14). The annual incidence of shoulder complaints was 7% in the general population (14) but varied across different age groups at 0.9% (31-35years), 2.5% (42-46 years), 1.1% (56-60 years), and 1.6% (70-74years) (13).

Leclerc et al. (15) has summarized risk factors for shoulder pain as a mix of personal (e.g. age and gender), occupational factors (skilled, semi-skilled or unskilled jobs), lifestyle behaviors (e.g. physical inactivity) and existing comorbidities (e.g. depression, heart and sleep conditions, and obesity). While the risk of shoulder pain increases with age (15), being a female (16) and having a long history of smoking are other risk factors for shoulder pain (17). Individuals who engaged or are engaging in an unskilled job that requires the constant use of their upper limb are more likely to report shoulder pain than those in skilled and semi-skilled jobs who do not use their upper limb repetitively while performing their job responsibilities (18). Physically inactive persons are also highly predisposed to shoulder pain (19). Moreover, the presence of mental/psychological comorbidities like depression and anxiety (20), undergoing a previous shoulder surgery, or even experiencing a past injury/dislocation on the shoulder can increase the risk of persistent shoulder pain (15).

Musculoskeletal Pain as a Multidimensional Construct

The International Association for the Study of Pain (IASP) describes pain as: "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" (21). Pain is not only the most recurrent symptom in musculoskeletal disorders but also accounts for most of the accompanying burden of disease (22). Although pain was initially perceived as a unidimensional construct with a resultant emphasis on capturing intensity, overwhelming evidence has established the multidimensional nature of pain (23,24). For instance, Melzack and Casey (25) hypothesized three dimensions, including: the sensory–discriminative, motivational–affective and cognitive–evaluative. The experience of pain perception is the confluence of six dimensions: physiologic, sensory, affective, cognitive, behavioral and socio-cultural (26,27). Since pain is multidimensional in nature, comprehensive pain assessment depends on the use of validated multidimensional patient-reported outcome measures (PROMs) that can adequately capture and quantify how pain impacts on different domains.

Patient-reported Outcome Measures (PROMs)

Patient-reported outcome measures (PROMs) are standardized, validated questionnaires completed by patients to measure their perceptions of their own functional status and wellbeing (28). Although initially developed for monitoring treatment effectiveness in clinical trials, PROMs are now used to also evaluate the patients view about their symptoms, functional status, treatment and other health-related qualities of life (28–30). In the past two decades, the importance of PROMs has been more recognized and widely accepted in health care practice. This is due to a shift in understanding that the patient's perspective about their health is genuine, and as valid as findings obtained from conventional biomedical clinical tests, lab results, and the clinicians' view (29,31).

Types of PROMs and their Advantages.

PROMs can be classified into seven main types (29). The first is **disease-specific tools** e.g. Western Ontario Rotator Cuff index (32): they focus on a specific disease, and are likely more sensitive to change, and appreciated by patients. Secondly, **site or region-specific tools** e.g. Knee injury and Osteoarthritis Outcome Score (33) or Shoulder Pain and Disability Index (34) can be used to evaluate conditions affecting the region of interest, often irrespective of the origin of pathology. Third are **dimension-specific tools** e.g. the Brief Pain Inventory-Short Form (BPI-SF) (35) or Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) (36): they provide the advantages of a thorough and exhaustive assessment of a health domain and can be compared across conditions. Fourth, **generic tools** e.g. the Medical Outcome Study 36-Item Short-Form Health Survey (37) can be applied in multiple conditions when disease-specific assessments are not available for evaluation. Fifth, **summary items** are brief and take less time to complete, providing a global rating of health or disability e.g. the

Global Rating of Change scales (38). Sixth, **individualized PROMs** e.g. Patient-specific Functional scale (39): respondents using this tool are not bounded to questions but are free to define their concerns or goals for treatment. Seventh, **utility tools** e.g. EuroQoL EQ-5D (40), allow respondents to quantify their preferences and values regarding their overall health status.

Psychometric Properties to consider when selecting PROMs.

Many outcome tools can be used for pain assessment in musculoskeletal conditions and selecting the most appropriate outcome measures for clinical or research purposes has become a difficult task that requires good understanding of psychometric properties, in addition to other factors (29,41,42). Some of the important psychometric properties that need to be considered when making decisions to select a PROM includes:

1. **Validity,** the level to which a tool measures the intended construct(s) (29,43), should be satisfactorily established for a tool to be considered for evaluation purposes. Ways of estimating validity includes investigating: (a) convergent and discriminative validity, where a clear hypothesis must be provided *a priori* (44,45); (b) criterion validity, where comparison is made against an established gold standard (42,46); and (c) structural validity using Rasch modelling or factor analysis to examine the dimensionality of multidimensional tools (41,42).

Content validity is the most foundational type of validity since without it, other validity indices have little value. It is the degree to which the content of a PROM instrument reflects the construct to be measured (43) and as such, describes how representative the items of a PROM reflect the patient's perspective under evaluation. Face validity can also be evaluated as a step toward establishing content validity. It entails synthesizing the impression of experts and/or patients (tool users) on the adequacy of tool (41,47). An instrument's content must adequately reflect what it is expected to evaluate before it should be considered for use (47). This benchmark must be established before any further investigation of psychometric properties because a tool with unrepresentative items does not merit any further evaluation (29,47).

2. **Reliability,** the degree to which the measurement is free from error (43), describes the reproducibility and internal consistency of a PROM instrument. Reproducibility is a critical measurement property that needs to be determined, especially, for all pain assessment tools. It measures the degree to which repeated measurements in stable respondents provide similar results (48–50) and as such, precedes evaluation of responsiveness or validity. Good evidence demonstrating relative and absolute reliability of a tool supports reproducibility; however, both forms of reliability focus on two different questions (48,51,52).

Absolute reliability (agreement) examines how closely related the scores from repeated measurement are comparable - that is, the more closely related, the higher the reliability, which substantiates the evaluative ability a tool. Methods examining standard error of the measurement (SEM) and minimal detectable change (MDC), and inspecting the distribution of scores on Bland-Altman plots provide evidence in support of a tool's absolute reliability (48,49). On the other hand, relative reliability is the degree to which individuals maintain their position in a sample over repeated measurements (48,49,51). The intraclass correlation coefficient (a unitless measure with magnitude ranging from 0 (poor) to 1 (very good) is the most accepted means of measuring relative reliability for ratio or interval scale scores. Higher ICC scores support the ability of a tool to discriminate between subjects regardless of measurement error (51,53,54).

Test-retest reliability is assessed by testing participants on repeated occasions (55). The PROM should be administered repeatedly within in a short enough time interval that ensures patients stability yet long enough time interval that avoids learning or memory effects. Internal consistency, measured using Cronbach's alpha, is a weaker form of assessment that is commonly used to determine the degree of the interrelatedness among items (56,57).

3. **Responsiveness,** the ability of a PROM to evaluate change over time in the construct to be measured (43), refers to an instruments ability to detect clinically relevant change. Evaluative instruments, like pain assessment tools, are expected to be able to determine the presence or absence of change in status following intervention. Some of the acceptable statistical approaches utilized for determining a tool's level of responsiveness include: (a) calculating the area under the curve (AUC) using the Receiver Operator Curve (ROC curve), (b) estimating the standardize response mean (SRM) or effect sizes and, (c) estimating the level of correlation with similar outcomes (44,58). Well-defined hypotheses with magnitude and direction of change should always be provided while considering the expected effect of administering or withholding an intervention (44,45,59). Statistical methods like the use of the paired T-test for significant differences between groups or Guyatt's responsiveness ratio (60) should not be used as indicators of responsiveness (58).

4. **Interpretability,** the degree to which one can assign qualitative meaning—that is, clinical or commonly understood connotations—to an instrument's quantitative scores or change in scores (43), is often established by estimating the minimal clinically important difference (MCID) score of the tool. Interpretability makes an instrument easy to use and understand in clinical practice and assists with classification and prediction.

Systematic Reviews of Measurement Studies

Systematic reviews of measurement studies involve identifying, extracting, critically appraising and comparing 'contextual' evidence from the literature on a tool's measurement properties (61). Evidence from systematic reviews informs decisions made for or against a tool and instill user's confidence in the tool's performance. Robust systematic reviews of measurement studies rely heavily on critical appraisals to authenticate and synthesize the quality of evidence supporting PROMs measurement properties. Critical appraisals often involve examining the quality of the measurement evidence (i.e. validity, reliability, responsiveness, etc.) against established standards, and evaluating the methodological quality of the study for bias (risk of bias). In some critical appraisal tools, the feasibility/usability of a tool, the administration burden and response burden are examined as part of the appraisal process. Two popular critical appraisal tools used in measurement studies are: (a) The (**CO**nsensus-based **S**tandards for the selection of health status **M**easurement **IN**struments) (COSMIN) Methodology, which comprises the risk of bias (44,45), quality citeria checklist (44,46,49) and the modified **G**rading of **R**ecommendations **A**ssessment, **D**evelopment and **E**valuation (GRADE) (44) and, (b) MacDermid's measurement studies quality assessment checklist (41). While complimentary, they have individual strengths and weakness which need to be appreciated.

Strengths and Limitations of the COSMIN Methodology

One of the main advantages of the COSMIN methodology lies in its standardized definition of measurement properties which guarantees less confusion when extracting evidence as described by reporting authors. Also, the COSMIN examines risk of bias per report of measurement properties. Therefore, a poor outcome for reporting of one measurement property does not necessarily impact on the rating of other reports in the same article because each report is treated as a 'stand-alone' study. Finally, a comprehensive user's manual (44) is available for reviewers to consult which decreases subjectivity. However, one disadvantage of the COSMIN method is that inexperienced users will find it difficult and confusing to synthesize and complete all stages of the critical appraisal involving completing the risk of bias, quality criteria checklist and Modified GRADE level of evidence determination. Further, some of the criterion are quite arbitrary: for instance, a sample of 50-100 subjects is needed for a study to be rated adequate in reliability assessment, even though sample size calculations often suggest less. Arbitrary benchmarks or items that affect imprecision are also included as bias criterion. This can have a major impact since it is the lowest rating in a section that is selected as the overall rating.

Strengths and Limitations of the MacDermid's Appraisal Method

MacDermid's tool is focused on overall study design and quality, not the bias associated with individual measurement properties. The scaling is numbered, and a total sum can be generated which makes it easy for users to conclude on the quality of studies. This attribute allows for identification of common design flaws in the individual studies. However, the weight apportioned to quality indicators is arbitrary and does not necessarily reflect the impact of potential sources of bias. Further, the focus on study design, rather than individual measurement properties, does not directly align with the information needed to make decisions about the adequacy of individual measurement properties. In addition, more training may be required to resolve complexity or sources of disagreement between raters from the absence of standardized definitions of measurement properties. As a quality tool, it does not focus on assessing risk of bias. Quality and risk of bias are related but separate constructs.

The Brief Pain Inventory-Short Form (BPI-SF): History, Content Structure and Advantages

The Brief Pain Inventory, formerly the Wisconsin Brief Pain Questionnaire (62,63), was initially developed to provide a simple but comprehensive outcome tool for monitoring analgesic effect in cancer pain management, epidemiological studies and research. Early versions were developed with sponsorship of the National Cancer Institute (NCI) and the Cancer Unit of the World Health Organization (WHO) (63). The Brief Pain Inventory has undergone series of transformations to improve its structure, including, the addition of the 'least pain item', and reducing the number of items in its long version - to decrease responders' burden (63). While the long version is still used for clinical research purpose, the short version, commonly referred to as the Brief Pain Inventory, is employed for pain assessment in conditions including musculoskeletal disorders (63).

The content structure of the BPI-SF is based on Beecher's definition of pain dimensionality as 'sensory' and 'reactive'(64). The sensory dimension of the tool evaluates how pain severity/intensity fluctuates on four items: pain at its - 'worst', 'least', 'on average' and 'now'. The reactive dimension of the tool evaluates how responders perceive pain interference in two sub-dimensions: (a) an activity sub-dimension consisting of 3 items: 'work', 'general activity' and 'walking', and (b) an affective sub-dimension consisting of 3 items: 'relations with others', 'enjoyment of life', and 'mood'. One item, 'pain interference with sleep', stands alone and can be influenced by both sub-dimensions (63).

As the name suggests, the greatest advantage/strength of the BPI-SF is its simplicity and brevity: it takes 5-minutes to complete, yet it captures pain comprehensively. Also, the BPI-SF evaluates pain 'interference', uniquely, with items easily appreciated by patients: hence, the **I**nitiative on **M**ethods, **M**easurement, and **P**ain **A**ssessment in **C**linical **T**rials (IMMPACT) group has recommended its use in all chronic pain-related clinical trials (65). Finally, multiple language translations (63), based on standard translation processes, are available for the BPI-SF which encourages its global use.

The Revised Short McGill Pain Questionnaire-2 (SF-MPQ-2): History, Content Structure and Advantages

The Revised Short McGill Pain Questionnaire Version-2 was developed about a decade ago (1st January, 2019) after Dworkin and his team noted the absence of a single tool for neuropathic and non-neuropathic pain assessment (36). The previous Short McGill Pain Questionnaire was then expanded to the current Revised Short McGill Pain Questionnaire Version-2, to be able to simultaneously evaluate and/or discriminate neuropathic and non-neuropathic pain symptoms.

The content structure of the expanded SF-MPQ-2 consist of 22 items: 15 items of which were retained from the former version, the Short McGill Pain Questionnaire (36). The 7 new items added comprise the neuropathic subscale and were selected based on the researchers experience and the results of focus groups with chronic pain patients (36). The SF-MPQ-2 evaluates 2 dimensions of pain: (a) sensory (pain quality and intensity) and, (b) an affective dimension (emotional experience of pain). Aside its evaluative properties, it has a high discriminative property to distinguish different pain types/qualities. Its 22-items distinguish pain into 4 categories: (a) continuous (throbbing pain, cramping pain, gnawing pain, aching pain, heavy pain, and tender); (b) neuropathic (hot-burning pain, cold-freezing pain, pain caused by light touch, itching, tingling or pins and needles, and numbness); (c) affective (tiring-exhausting, sickening, fearful, punishing-cruel), and (d) intermittent (shooting pain, stabbing pain, sharp pain, splitting pain, electric-shock pain, piercing).

As its strengths, the SF-MPQ-2 does not only assess pain intensity but can also be used to distinguish pain according to its source. This makes it very useful for pain assessment when there is need to be sure of the mechanism of pain (nociceptive or neuropathic), or when there is need to quantify mixed (both neuropathic and nociceptive) pain experiences. On the down side, some of the SF-MPQ-2 pain descriptors are difficult to appreciate by patients (66,67). Also, 22 items in one questionnaire may be perceived as too long and burdensome to complete (66,67). Finally, although the Mapi Research Trust has provided computerized translations of the SF-MPQ-2, they are not based on standardized cross-cultural translations involving forward and backward translation processes.

Current Gap in the Literature

The main objective of this thesis was to explain the sufficiency of measurement evidence backing the use of the Brief Pain Inventory- Short Form and Revised Short McGill Pain Questionnaire Version-2 in pain-related musculoskeletal conditions. Currently, both tools are used frequently for musculoskeletal pain assessment both in the clinical and research setting, however, no single study has synthesized the scope of evidence supporting their measurement properties for use in MSK conditions. MSK conditions are common but diverse. Clearly understanding the measurement properties backing outcome measures can inform users choice. Hence, the absence of a comprehensive review of the evidence suggest selection by researchers/clinicians is based on reports obtained from single studies, colleagues or peers' recommendations, easy access to the tools, high recognition or even the appearance of their items/face validity (68)**.** However, we know that comprehensive pain assessment depends on the use of tools with proven context-specific evidence backing their measurement properties, because only valid tools yield dependable scores (41,42,44–46,59,69). Therefore, systematically reviewing the literature to determine the sufficiency of evidence for measurement properties underpinning the use of the BPI-SF and SF-MPQ-2 in MSK conditions is overdue and necessary.

The two research questions guiding this dissertation were as follows:

1. What is the quality and content of measurement evidence supporting the use of the Brief Pain Inventory-Short Form and Revised Short McGill Pain Questionnaire Version-2 in Musculoskeletal Conditions?
2. What is the reproducibility (reliability and agreement parameters) and internal consistency of the Revised Short McGill Pain Questionnaire Version-2 for use among patients with musculoskeletal shoulder pain?

Composition of Dissertation Papers

This dissertation consists of two papers presented in a manuscript style as Chapters two and three. Chapter two is a systematic review manuscript. Chapter three is a research study on the reproducibility of the Revised Short McGill Pain Questionnaire Version-2 among patients with shoulder pain. The systematic review (Chapter – 2) examined the quality and content of measurement evidence reported for the Brief Pain Inventory-Short Form and Revised Short McGill Pain Questionnaire Version-2 in pain-related musculoskeletal conditions. In this review, we synthesized, appraised and compared reported evidence on the measurement properties of both outcome tools. The critical appraisal involved two methods that checked the quality and risk of bias (the COSMIN guidelines) and the rigor of authors report of measurement properties (MacDermid's tools). The review findings helped us identify the gaps in the literature which informed our objective in the third chapter of the dissertation.

The third chapter of the thesis comprehensively examined the reliability and agreement properties of the Revised Short McGill Pain Questionnaire Version-2 among patients with musculoskeletal shoulder pain. In this study, the internal consistency, test-retest reliability, standard error of measurement (SEM), minimal detectable change (MDC) and Bland-Altman methods were used to assess the reproducibility of the SF-MPQ-2. The results of the study established evidence in support of the reproducibility of the SF-MPQ-2 for use among adults with shoulder pain. In summary, research in this thesis attempts to address the literature gaps in measurement evidence by systematically summarizing the available evidence on the psychometric properties of the Brief Pain Inventory-Short Form and Revised Short McGill Pain Questionnaire Version-2 and evaluating the reproducibility of Revised Short McGill Pain Questionnaire Version-2 among patients with musculoskeletal shoulder pain.

REFERENCES

1. Punnett L, Wegman DH. Work-related musculoskeletal disorders: the epidemiologic evidence and the debate. J Electromyogr Kinesiol. 2004 Feb;14(1):13–23.

2. Briggs AM, Cross MJ, Hoy DG, Sànchez-Riera L, Blyth FM, Woolf AD, et al. Musculoskeletal Health Conditions Represent a Global Threat to Healthy Aging: A Report for the 2015 World Health Organization World Report on Ageing and Health. Gerontologist. 2016 Apr;56(Suppl 2):S243–55.

3. Lidgren L. The Bone and Joint Decade and the global economic and healthcare burden of musculoskeletal disease. In: Journal of Rheumatology. 2003. p. 4–5.

4. Aptel M, Aublet-Cuvelier A, Claude Cnockaert J. Work-related musculoskeletal disorders of the upper limb. Jt Bone Spine. 2002 Dec;69(6):546–55.

5. Storheim K, Zwart J-A. Musculoskeletal disorders and the Global Burden of Disease study. Ann Rheum Dis. 2014 Jun;73(6):949–50.

6. Woolf AD, Pfleger B. Burden of major musculoskeletal conditions. Bull World Health Organ. 2003;81(9):646–56.

7. Coyte PC, Asche C V., Croxford R, Chan B. The economic cost of musculoskeletal disorders in Canada. Arthritis Care Res (Hoboken). 1998 Oct;11(5):315–25.

8. March L, Smith EUR, Hoy DG, Cross MJ, Sanchez-Riera L, Blyth F, et al. Burden of disability due to musculoskeletal (MSK) disorders. Best Pract Res Clin Rheumatol. 2014 Jun;28(3):353–66.

9. Badley EM, Rasooly I, Webster GK. Relative importance of musculoskeletal disorders as a cause of chronic health problems, disability, and health care utilization: findings from the 1990 Ontario Health Survey. J Rheumatol. 1994 Mar;21(3):505–14.

10. Karel YHJM, Scholten-Peeters GGM, Thoomes-de Graaf M, Duijn E, van Broekhoven JB, Koes BW, et al. Physiotherapy for patients with shoulder pain in primary care: a descriptive study of diagnostic- and therapeutic management. Physiotherapy. 2017 Dec;103(4):369–78.

11. Mitchell C, Adebajo A, Hay E, Carr A. Shoulder pain: diagnosis and management in primary care. BMJ. 2005 Nov;331(7525):1124–8.

12. Brox JI. Regional musculoskeletal conditions: shoulder pain. Best Pract Res Clin Rheumatol. 2003;17(1):33–56.

13. Luime J, Koes B, Hendriksen I, Burdorf A, Verhagen A, Miedema H, et al. Prevalence and incidence of shoulder pain in the general population; a systematic review. Scand J Rheumatol. 2004 Mar;33(2):73–81.

14. van der Heijden GJMG. Shoulder disorders: a state-of-the-art review. Baillieres Best Pract Res Clin Rheumatol [Internet]. 1999 Jun [cited 2018 Aug 25];13(2):287–309. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10952865

15. Leclerc A, Chastang J-F, Niedhammer I, Landre M-F, Roquelaure Y, Study Group on Repetitive Work. Incidence of shoulder pain in repetitive work. Occup Environ Med. 2004 Jan;61(1):39–44.

16. Straker LM, Smith AJ, Bear N, O'Sullivan PB, de Klerk NH. Neck/shoulder pain, habitual spinal posture and computer use in adolescents: the importance of gender. Ergonomics [Internet]. 2011 Jun 17 [cited 2019 Jun 8];54(6):539–46. Available from: https://www.tandfonline.com/doi/full/10.1080/00140139.2011.576777

17. Baumgarten KM, Gerlach D, Galatz LM, Teefey SA, Middleton WD, Ditsios K, et al. Cigarette Smoking Increases the Risk for Rotator Cuff Tears. Clin Orthop Relat Res [Internet]. 2010 Jun 13 [cited 2019 Jun 8];468(6):1534–41. Available from: http://link.springer.com/10.1007/s11999-009-0781-2

18. Hanvold TN, Wærsted M, Mengshoel AM, Bjertness E, Stigum H, Twisk J, et al. The effect of work-related sustained trapezius muscle activity on the development of neck and shoulder pain among young adults. Scand J Work Environ Health [Internet]. 2013 Jul [cited 2019 Jun 8];39(4):390–400. Available from: https://www-jstor-org.proxy1.lib.uwo.ca/stable/23558338

19. Nygren A, Berglund A, von Koch M. Neck-and-shoulder pain, an increasing problem. Strategies for using insurance material to follow trends. Scand J Rehabil Med Suppl [Internet]. 1995 [cited 2018 Aug 25];32:107–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7784832

20. Siivola SM, Levoska S, Latvala K, Hoskio E, Vanharanta H, Keinänen-Kiukaanniemi S. Predictive factors for neck and shoulder pain: a longitudinal study in young adults. Spine (Phila Pa 1976) [Internet]. 2004 Aug 1 [cited 2019 Jun 8];29(15):1662–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15284513

21. Merskey H, Bogduk N. Classification of chronic pain. Descriptions of chronic pain syndromes and definitions of pain terms. Prepared by the International Association for the Study of Pain, Subcommittee on Taxonomy. Pain Suppl. 1986;3:S1-226.

22. Treede R-D, Rief W, Barke A, Aziz Q, Bennett MI, Benoliel R, et al. Chronic pain as a symptom or a disease. Pain. 2019 Jan;160(1):19–27.

23. Clark CW, Yang JC, Tsui S-L, Ng K-F, Clark SB. Unidimensional pain rating scales: a multidimensional affect and pain survey (MAPS) analysis of what they really measure. Pain. 2002 Aug;98(3):241–7.

24. Dansie EJ, Turk DC. Assessment of patients with chronic pain. Br J Anaesth. 2013 Jul;111(1):19–25.

25. Melzack R, senses KC-T skin, 1968 undefined. Sensory, motivational, and central control determinants of pain: a new conceptual model. researchgate.net.

26. Ahles TA, Blanchard EB, Ruckdeschel JC. The multidimensional nature of cancer-related pain. Pain [Internet]. 1983 Nov 1 [cited 2018 Aug 25];17(3):277–88. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/0304395983901008

27. McGuire DB. Comprehensive and multidimensional assessment and measurement of pain. J Pain Symptom Manage [Internet]. 1992 Jul [cited 2018 Aug 25];7(5):312–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1624815

28. Dawson J, Doll H, Fitzpatrick R, Jenkinson C, Carr AJ. The routine use of patient reported outcome measures in healthcare settings. BMJ. 2010 Jan;340(jan18 1):c186–c186.

29. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess (Rockv). 1998;2(14).

30. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013 Jan;346(jan28 1):f167–f167.

31. Höfer S, Lim L, Guyatt G, Oldridge N. The MacNew Heart Disease health-related quality of life instrument: A summary. Health Qual Life Outcomes [Internet]. 2004 Jan 8 [cited 2019 May 21];2(1):3. Available from: http://hqlo.biomedcentral.com/articles/10.1186/1477-7525-2-3

32. Kirkley A, Alvarez C, Griffin S. The Development and Evaluation of a Disease-specific Quality-of-Life Questionnaire for Disorders of the Rotator Cuff: The Western Ontario Rotator Cuff Index. Clin J Sport Med [Internet]. 2003 Mar [cited 2019 Jun 8];13(2):84–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12629425

33. Roos EM, Roos HP, Lohmander LS, Ekdahl C, Beynnon BD. Knee Injury and Osteoarthritis Outcome Score (KOOS)—Development of a Self-Administered Outcome Measure. J Orthop Sport Phys Ther [Internet]. 1998 Aug 1 [cited 2019 Jun 8];28(2):88–96. Available from: http://www.jospt.org/doi/10.2519/jospt.1998.28.2.88

34. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a Shoulder Pain and Disability Index. Arthritis Care Res (Hoboken) [Internet]. 1991 Dec 1 [cited 2019 Jun 8];4(4):143–9. Available from: http://doi.wiley.com/10.1002/art.1790040403

35. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singapore. 1994 Mar;23(2):129–38.

36. Dworkin RH, Turk DC, Revicki DA, Harding G, Coyne KS, Peirce-Sandner S, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). Pain [Internet]. 2009 Jul 1 [cited 2018 Aug 25];144(1):35–42. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S0304395909001250

37. Mchorney CA, Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II . Psychometric and ClinicMchorney, Colleen A, John E Ware & Anastasia E Raczek. 2014. The MOS 36-Item Short-Form Health Survey (SF-36): II . Psychometric and Clinical Tests of Validity in Measuring. Med Care [Internet]. 2014 [cited 2019 Jun 8];31(3):247–63. Available from: https://www.jstor.org/stable/3765819

38. Kamper SJ, Maher CG, Mackay G. Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design. J Man Manip Ther [Internet]. 2009 Jul 18 [cited 2019 Jun 8];17(3):163–70. Available from: http://www.tandfonline.com/doi/full/10.1179/jmt.2009.17.3.163

39. Stratford P. Assessing Disability and Change on Individual Patients: A Report of a Patient Specific Measure. Physiother Canada [Internet]. 1995 Oct 8 [cited 2019 Jun 8];47(4):258–63. Available from: https://utpjournals.press/doi/10.3138/ptc.47.4.258

40. The EuroQol Group. EuroQol - a new facility for the measurement of health-related quality of life. Health Policy (New York) [Internet]. 1990 Dec 1 [cited 2019 Jun 8];16(3):199–208. Available from: https://www.sciencedirect.com/science/article/abs/pii/0168851090904219

41. MacDermid JC, Law M, Michlovitz S. Outcome measurement in evidence-based rehabilitation. In: Law M, MacDermid JC, editors. Evidence-based rehabilitation: A guide to practice. 3rd ed. Thorofare NJ, USA: Slack Incorporated; 2014. p. 65–104.

42. MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy practice. J Hand Ther. 2004 Apr;17(2):165–73.

43. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010 Jul;63(7):737–45.

44. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HCW, et al. COSMIN methodology for systematic reviews of Patient - Reported Outcome Measures (PROMs). 2018.

45. 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. Qual Life Res. 2018 May;27(5):1171–9.

46. Prinsen CAC, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a “Core Outcome Set” - a practical guideline. Trials. 2016 Dec;17(1):449.

47. 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. Qual Life Res [Internet]. 2018 May 17 [cited 2019 Jun 8];27(5):1159–70. Available from: http://link.springer.com/10.1007/s11136-018-1829-0

48. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol [Internet]. 2006 Oct 1 [cited 2019 Mar 10];59(10):1033–9. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S0895435606000291

49. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol [Internet]. 2007 Jan 1 [cited 2018 Aug 26];60(1):34–42. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S0895435606001740#bib41

50. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res [Internet]. 2001 [cited 2019 Mar 7];10(7):571–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11822790

51. Kottner J, Streiner DL. The difference between reliability and agreement. J Clin Epidemiol [Internet]. 2011 Jun 1 [cited 2019 Mar 10];64(6):701–2. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S0895435610004336

52. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud [Internet]. 2011 Jun 1 [cited 2019 Mar 7];48(6):661–71. Available

from: https://www.sciencedirect.com/science/article/pii/S0020748911000368

53. Schuck P. Assessing reproducibility for interval data in health-related quality of life questionnaires: Which coefficient should be used? Qual Life Res [Internet]. 2004 Apr [cited 2019 Mar 10];13(3):571–85. Available from: http://link.springer.com/10.1023/B:QURE.0000021318.92272.2a

54. Pynsent PB. Choosing Health Outcome Measures. J Bone Jt Surg [Internet]. 2001 [cited 2019 Mar 10];792–4. Available from: https://online.boneandjoint.org.uk/doi/pdf/10.1302/0301-620X.83B6.0830792

55. Streiner D, Norman G, Cairney J. Health measurement scales: a practical guide to their development and use. 5th ed. New York, USA: Oxford University Press; 2015.

56. Cronbach LJ. Test "reliability": Its meaning and determination. Psychometrika [Internet]. 1947 Mar [cited 2019 Mar 10];12(1):1–16. Available from: http://link.springer.com/10.1007/BF02289289

57. Karanicolas PJ, Bhandari M, Kreder H, Moroni A, Richardson M, Walter SD, et al. Evaluating Agreement: Conducting a Reliability Study. J Bone Jt Surgery-American Vol [Internet]. 2009 May [cited 2019 Mar 10];91(Suppl 3):99–106. Available from: https://insights.ovid.com/crossref?an=00004623-200905003-00016

58. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Med Res Methodol [Internet]. 2010 Dec 18 [cited 2019 Jun 8];10(1):22. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-10-22

59. 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. Qual Life Res. 2018 May;27(5):1147–57.

60. Guyatt G, Walter S, Norman G. Measuring change over time: Assessing the usefulness of evaluative instruments. J Chronic Dis [Internet]. 1987 Jan 1 [cited 2019 Jun 8];40(2):171–8. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/0021968187900695

61. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res. 2009 Apr;18(3):313–33.

62. Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain. 1983 Oct;17(2):197–210.

63. Cleeland CS. The Brief Pain Inventory User Guide [Internet]. 2008. Available from: https://www.mdanderson.org/documents/Departments-and-Divisions/Symptom-Research/BPI_UserGuide.pdf

64. Beecher H. Measurement of subjective responses: quantitative effects of drugs. New York, NY, US:Oxford University Press.; 1959.

65. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2005 Jan;113(1):9–19.

66. Packham TL. Outcome Measurement in Complex Regional Pain Syndrome. McMasters University; 2011.

67. Packham T, MacDermid JC, Henry J, Bain J. A systematic review of psychometric evaluations of outcome assessments for complex regional pain syndrome. Disabil Rehabil. 2012 Jun;34(13):1059–69.

68. Jumbo S, MacDermid J, Michael K, Packham TL, Athwal GS, Faber K. Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-

Form McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review Protocol. Submitted. 2018;

69. Mehta SP, MacDermid JC, Richardson J, MacIntyre NJ, Grewal R. A Systematic Review of the Measurement Properties of the Patient-Rated Wrist Evaluation. J Orthop Sport Phys Ther [Internet]. 2015 Apr [cited 2018 Dec 11];45(4):289–98. Available from: http://www.jospt.org/doi/10.2519/jospt.2015.5236

CHAPTER 2:

Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review

Samuel U. Jumbo, BMR. PT. Western University, Faculty of Health and Rehabilitation Sciences, Elborn College London, Ontario, Canada. Email: **sjumbo@uwo.ca**

Joy C. MacDermid, PT, PhD. Western University, Faculty of Health and Rehabilitation Sciences, Elborn College London, Ontario, Canada; McMaster University, School of Rehabilitation Science, 1400 Main Street West, Hamilton, Ontario, Canada; Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, Ontario, Canada. Email: **jmacderm@uwo.ca**

Michael E. Kalu, PT, MSc, PhD (S). McMaster University, School of Rehabilitation Science, 1400 Main Street West, Hamilton, Ontario, Canada. Email: **kalum@mcmaster.ca**

Tara L. Packham, OT. PhD. McMaster University, School of Rehabilitation Science, 1400 Main Street West, Hamilton, Ontario, Canada. Email: **packhamt@mcmaster.ca**

George S. Athwal, MD. FRCSC. Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, Ontario, Canada. Email: **gathwal@uwo.ca**

Kenneth J. Faber, MD, MHPE, FRCSC. Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, Ontario, Canada. Email: **kjfaber@uwo.ca**

Corresponding Author: Samuel U. Jumbo, BMR. PT. Western University, Faculty of Health and Rehabilitation Sciences, Elborn College London, Ontario, Canada. **sjumbo@uwo.ca**

Funding: This work was supported by the Canadian Pain Network funded by the Canadian Institutes for Health Research. Grant Number: CIHR FRN SCA-145102.

Institutional Review Board: Not Applicable

Statement of Financial Disclosure and Conflict of Interest: None to report.

Acknowledgements: Samuel U. Jumbo was supported by a Western University Graduate Research Scholarship. Joy C. MacDermid was supported by a Canadian Institutes of Health Research (CIHR) Chair in Gender, Work and Health and the Dr. James Roth Chair in Musculoskeletal Measurement and Knowledge Translation.

Word Count: 6888, Abstract 374

ABSTRACT

Study design: Systematic review of clinical measurement studies.

Background: The BPI-SF and SF-MPQ-2 are general-use, self-report, multidimensional pain measures frequently used in musculoskeletal (MSK) conditions. Synthesizing knowledge of their measurement properties, as assessed in MSK conditions, should provide a deeper understanding of their strengths and limitations.

Objectives: To systematically locate, critically appraise, compare and summarize clinical measurement research addressing the use of BPI-SF and SF-MPQ-2 in pain-related musculoskeletal conditions.

Methods: Four databases (Medline, CINAHL, EMBASE & SCOPUS) were systematically searched for relevant citations, each for the BPI-SF and SF-MPQ-2. We included articles reporting the psychometric properties (e.g. validity, reliability, responsiveness) and interpretability indices (e.g. minimal clinically important difference) of both tools, as assessed in mixed and specific MSK studies. Independently, two reviewers extracted data and assessed the quality of evidence with the MacDermid's measurement studies quality assessment tool and the updated **CO**nsensus-based **S**tandards for the selection of health **M**easurement **IN**struments (COSMIN) guidelines.

Results: Twenty-five articles were included (BPI-SF, n=17; SF-MPQ-2, n=8). Both tools lack reporting on their cross-cultural validities and measurement error indices (standard error of measurement, minimal detectable change). High quality studies suggest the tools are internally consistent (α = 0.83-0.96), and they associate modestly with similar outcome measures (r = 0.3-0.69). There is strong evidence suggesting the BPI-SF conforms to its two-dimensional structure in MSK studies; the SF-MPQ-2 four-factor structure was not clearly established. Seven reports of high-to-moderate quality evidence were supportive of the BPI-SF known group validity (n=2) and responsiveness (n=5) while no similar evidence was available for the SF-MPQ-2. Furthermore, the SF-MPQ-2 was more frequently associated with floor effects in MSK studies than the BPI-SF (SF-MPQ-2, 42% vs BPI-SF, 6%).

Conclusion: Although the SF-MPQ-2 presents potential, and both tools display high-quality evidence in support of their internal consistency and criterion-convergent validities, high to moderate quality evidence suggests the BPI-SF subscales have a better responsiveness, retest reliability, known group validity and structural validity than the SF-MPQ-2. Therefore, the BPI-SF is currently better for pain assessment in MSK conditions. However, methodologically sound studies are still needed for both tools' measurement properties including their cross-cultural validities, retest reliability, measurement error indices, minimal clinical important difference and clinical important difference.

Keywords: Brief Pain Inventory; McGill Pain Questionnaire; Musculoskeletal Conditions; Patient-Reported Outcome Measures; Psychometric Properties

INTRODUCTION

Musculoskeletal (MSK) conditions are among the leading causes of years lived with disability.[15,53] Pain originating from MSK conditions has a significant impact on patients' general wellbeing and commonly results in frequent visits to the emergency or outpatient department of hospitals and clinics.[3,22] Patient-reported outcome measures (PROMs) are the primary methods of assessing and monitoring the patients' pain experience, and well validated PROMs help clinicians make informed care decisions.[27,30,44] A large number of PROMs with multidimensional scales now exist for pain assessment in various conditions including MSK disorders. The Brief Pain Inventory-Short Form (BPI-SF)[8] and the Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2)[10] are examples of generic PROMs that are increasingly being used for pain assessment in musculoskeletal conditions: experts are currently presenting these tools as core outcomes for pain assessment in chronic musculoskeletal pain studies (BPI-SF)[9,23] and complex regional pain syndrome (SF-MPQ-2 neuropathic subscale).[13]

The BPI-SF contains 11-items evaluating the severity and interference of pain with daily functioning. Four items quantify patients' responses on the severity of pain at its 'worst', 'least', 'on average' and 'now.' Each of the four descriptors are anchored to a common scaling structure that has zero as 'no pain' and 10 as 'pain as bad as you can imagine'. To obtain a total pain-severity score, the mean of the 4 severity items is computed. The interference subscale of the BPI-SF captures the patients' perception of how pain impacts seven constructs: 'mood', 'enjoyment of life', 'relationship with others', 'sleep', 'general activity', 'walking ability', and 'work'. Each of the seven items are equidistantly bounded on a zero-to-10 numerical rating scale having zero as 'does not interfere' and 10 as 'completely interferes'. To obtain a total pain interference score, the mean of four or more of the interference items must be computed.[7]

The Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2)[10] was created by expanding its parent version, the Short McGill Pain Questionnaire (SF-MPQ). It concurrently evaluates the severity and characteristics/qualities of pain originating from both neuropathic and nociceptive sources. The SF-MPQ-2 has 22-items evaluating pain across four dimensions as follows: 1) continuous pain (throbbing, cramping, gnawing, aching, heavy, and tender pain); 2) intermittent pain (shooting, stabbing, sharp pain, splitting pain, electric-shock, and piercing pain); 3) neuropathic pain (hot-burning, cold-freezing, pain caused by light touch, itching, tingling or pins and needles, and numbness pain), and 4) affective pain (tiring-exhausting, sickening, fearful, and punishing-cruel). Each pain descriptor is scored on a zero (none)-to-10 (worst possible) numerical rating scale. The total pain score is the mean of the 22-items, while the total subscale scores are the mean of each of their respective descriptor cluster.[10]

Like some generic tools, the BPI-SF and SF-MPQ-2 were originally developed and tested for pain assessment in specific-disease populations. For instance, the BPI-SF was originally developed for cancer pain assessment[8,20] while the SF-MPQ-2 was purposefully expanded to include the neuropathic pain elements, and initially validated in a diverse chronic pain population with patients having conditions including neck and shoulder pain, and painful diabetic neuropathy.[10] Currently, however, both tools are deemed useful for pain assessment in other conditions, including musculoskeletal conditions. Although evidence supporting the measurement properties of both tools has continued to accumulate, there has been no systematic inquiry on their performance as *‘general-use’* pain assessment tools in musculoskeletal conditions. Previous reviews of the BPI-SF[5,6,43] and the SF MPQ-2[43] only summarized their measurement properties as examined in back pain, with no report on other MSK conditions.

However, by systematically extracting evidence from the pool of studies that have reported the BPI-SF and SF-MPQ-2 measurement properties in mixed and specific MSK populations, we expect to have a broader understanding of their measurement performance in MSK conditions. Such knowledge aids decision-making in the clinical and research setting, and helps in selecting the appropriate outcome measure for use, for example, in epidemiological pain studies where these tools are sometimes employed.[11] Furthermore, the methodological quality of a study places value on its report of measurement properties: therefore, a thorough critical appraisal of the measurement reports would help identify their risk of bias, which often modifies the strength accorded the studies' conclusions.[36]

Investigating the two questionnaires measurement properties is timely and will yield insights on how they efficiently assess pain in MSK conditions. Such knowledge can guide the preferences of the busy clinician/researcher encountering such conditions thus fostering evidence-based practice in MSK pain management. The objective of this review was to systematically locate, summarize, critically appraise, and compare the quality of measurement research utilizing the Brief Pain Inventory-Short form (BPI-SF) and the Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions.

METHODS

Design

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and checklist,[35] and the protocol manuscript[16] was registered with Prospero (CRD42018095862). The review was conducted in four steps: (a) comprehensive searches were performed in four bibliographic databases to identify relevant citations, each for the BPI-SF and SF-MPQ-2; (b) title, abstract and full-text were screened to fit pre-determined standards; (c) data on measurement properties were extracted; (d) in two phases, the MacDermid's measurement studies quality assessment tool and the COSMIN guidelines were used to assess the quality of the included studies.

Data Source and Search Strategy

Search strategies developed in consultation with a health-research librarian were run last on the 10th of July 2019 in the Medline—OVID, EMBASE—OVID, CINAHL (Cumulative Index to Nursing and Allied Health Literature), and Scopus bibliographic databases to identify relevant citations. The search concepts for measurement properties were refined from a previous comprehensive search filter validated by Terwee et al[50] and included different measurement property keywords **(see Appendix 1)**. The search for studies addressing the BPI-SF properties was not restricted by time or language. However, the search for the SF-MPQ-2 articles was restricted by time to its year of first publication (i.e. January 1st, 2009)[10] in order to limit citations related to its parent versions (SF-MPQ and MPQ) but language was not restricted.

Eligibility requirements

Inclusion Criteria:

(i) We included only studies whose purpose was to evaluate at least one of the measurement properties of the BPI-SF and SF-MPQ-2 in a sample population characterized with MSK conditions fully, or to an extent that satisfies 70-percent of the study sample size.

(ii) We included studies with available full text, published in a peer-reviewed journal in any language and conducted in a population within the age of 16 years and above.

Exclusion Criteria:

(i) We excluded all letters to the editor, review articles, book reviews and short communications, clinical protocols, case reports, animal studies, and series.

(ii) We excluded studies without explicit evidence that describe pain to be of MSK origins either in summary tables or text. Examples of such articles include studies with unclear terms used to define the sample populations' pain, such as, 'non-malignant pain' or 'chronic pain' or 'non-cancer pain' without clearly relating such pain to be of MSK origin fully, or partly.

(iii) We excluded studies conducted in disability or pain-related conditions that were primarily due to some of the following: congenital and developmental abnormalities, neoplasm, infection, surgical procedures not due to MSK conditions (e.g. coronary heart surgery, laparotomy), neurological or neuropathic pain (undefined as lumbar, cervical or thoracic radiculopathies with back pain), or HIV/AIDS pain.

Study selection and screening

Prior to the screening/selection phases, articles retrieved from each database were exported to the Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia - available at www.covidence.org), for de-duplication and study selection. In line with the eligibility requirements, the screening/selection took place in two steps: first, title and abstracts were independently screened for obviously unrelated articles. We then assessed the full text of emerging articles for congruency with the aims and eligibility requirements of the review. A further hand search of the included studies reference list was conducted, each for the BPI-SF and SF-MPQ-2, to identify any other relevant citations. All disagreement during the selection were discussed and settled among the screening authors (JS and MK).

Data extraction

We extracted data from individual studies using a structured data extraction form developed from the guide available in the second author's (JCM) work.[28] Two review authors (JS and MK) worked independently to extract data from the included studies, and a third review author was to be contacted if any disagreement arose. When reported, the following information were extracted: floor-ceiling effect, construct validity (criterion-convergent and known group), internal structure (internal consistency and structural validity [i.e. Rasch and factor analysis]), reliability (test-retest), responsiveness (AUC's, change correlation indices, standardized response mean [SRM], effect-size [ES]), interpretability properties (clinically important difference [CID], minimal clinically important difference [MCID]), measurement error indices, and measurement invariance/cross-cultural validities. Data was summarized according to the subscale of the tools evaluated. To guide our review, painful MSK conditions were defined as disorders affecting the muscles, bones, soft tissue, joints and spine.[40] We then classified the acceptable studies into populations as follows:

a) 'Mixed' MSK population studies: studies that satisfied the requirements of ≥ 70-percent sample size proportion representing MSK conditions of different mechanism or pathophysiology or described by different body regions.

b) 'Specific' MSK population studies: those conducted among homogenous MSK samples described by the body region affected or the pathophysiology/mechanism.

We also extracted information on: (a) the characteristics of the studies, for example, country where the study was conducted, language, study design, study setting, the SF-MPQ-2 and BPI-SF language/version/subscale/item used, and (b) participants characteristics including sample size, age, and sex.

Review Team Hypotheses

The following hypotheses, as suggested by the COSMIN[36,38] initiative, were defined in advance by the review team to guide the quality assessment phase.

a) Correlations between the two questionnaires and other pain/health-related outcome tools were expected to be 0.3 and above in magnitude. Correlations were then classified as low to moderate at 0.3-0.69; high at 0.7 and above.[28,32]

b) The lower bound confidence intervals of reports supporting Area Under the Curve (AUC) estimates had to be ≥ 0.7 to represent discrimination beyond chance.

c) We described the correlation indices between the change scores of other pain/health-related outcome measures and the two questionnaires as sufficient at ≥ 0.3.

However, the review team could not define hypotheses to assess authors responsiveness report based on the standardized response mean (SRM) and effect size (ES) indices. As such, article authors were expected to provide context-specific hypotheses that defined the magnitude and direction of expected change.

Quality Assessment

Quality assessment was conducted in two phases: first, the MacDermid's measurement studies quality assessment tool[26] was used to appraise the included studies. Next, the COSMIN guidelines[36,41,42,49] were used to examine methodological risk of bias and to compare reported measurement properties against benchmark quality standards. Both phases of quality assessment were deemed complementary.

Phase 1: Quality, Comprehensiveness and Breadth of Measurement Evidence Reports

An initial calibration meeting informed the reviewers' (JS and MK) independent appraisal of the included articles. MacDermid's measurement studies quality assessment tool[26] was used to examine the quality, breadth and rigor of authors report against 12 criteria. Where applicable, each criterion receives: 0-points, if judged, 'not done/documented' OR 'substantial inadequate' OR 'inappropriate'; 1-point, if judged, 'acceptable but suboptimal'; and 2-points, if judged, 'consistent with best practice', and NA if judged 'Not Applicable' to a study. An article can receive a total score ranging from zero (lowest) to 24 (highest) points, which can be converted to a percentage that represents its total quality rating. Percentile scores were interpreted as follows: poor quality report, 0%–30%; Fair, 31%–50%; Good, 51%–70%; Very good, 71%–90%; Excellent, >90%. The reliability of the process was calculated.

Phase 2: Risk of Bias Assessment and Quality of Measurement Properties

In this phase, the same review authors (JS and MK) independently assessed the included studies for conformity with the COSMIN risk of bias (RoB)[38,41] and quality criteria checklists.[42,49] The RoB checklist consists of 10 boxes, each representing a measurement property featured in the COSMIN consensus-based taxonomy.[37] Irrespective of the terminology used to define measurement properties by individual authors, we defined all measurement properties according to the COSMIN consensus-based taxonomy of measurement properties,[37] and that determined the corresponding box to complete for the risk of bias assessment. Each of the risk of bias assessment box contains several items/questions that are scored on a 4-point rating scale as 'very good', 'adequate', 'doubtful' and 'inadequate'.[38] The lowest rating for any item/question on a study's measurement property determines its overall rating for methodological risk of bias based on the worst score count system. Of the ten boxes in the COSMIN risk of bias checklist, we ignored two boxes: PROM development and content validity, because the BPI-SF and SF-MPQ-2 were not initially validated in study populations satisfying our inclusion criteria.[36,38]

In the second phase of the COSMIN quality assessment, all the extracted data on measurement properties for each of the tools were rated against the good measurement property quality criteria, as available in the COSMIN User's Manual Version-1.[36,42,49] Based on the category of the measurement property reported (e.g. internal consistency, test-retest reliability), studies were compared and rated against the COSMIN quality criteria as: **'Sufficient' (+)**, if within the benchmark quality criteria; '**Indeterminate' (?)**, if there was an inadequate report to compare to the benchmark quality criteria, and; **'Insufficient' (-)**, if the reported measurement property was below the benchmark quality criteria.

Evidence Synthesis

To determine the level of evidence supporting the two questionnaires measurement properties, we applied the Modified GRADE **(G**rades of **R**ecommendation, **A**ssessment, **D**evelopment, and **E**valuation) as described by the COSMIN initiative.[36] In our synthesis, more attention was given to the consistency of reports contained in studies with 'sufficient' **(+)** quality ratings (which is the acceptable quality criteria rating) and prone to lesser risk of bias: that is, with 'very good' or 'adequate' risk of bias rating. First, we pooled/summarized the results extracted on each measurement property, per tool. For example, the estimated intraclass correlation coefficient (ICC) for the BPI-SF and SF-MPQ-2 was summarized using a weighted average. Cronbach alpha for internal consistency was summarized using the observed range of occurrences across the studies with a low risk of bias. Proportions were used to summarize the number of correlations within the low-to-moderate range (rho = 0.3-0.69), and 'high' range (rho ≥ 0.7). For structural validity, responsiveness and known group validity, we considered the number of studies with a low risk of bias rating available on each of those measurement properties, per tool, and used narrative synthesis to summarize their findings. Conclusions were formulated using the Modified GRADE level of evidence approach while considering the risk of bias, inconsistency, imprecision and indirectness of the pooled results, under the assumption that all the captured conditions were MSK conditions, regardless of their type (mixed or specific).[38,41] The same pair of reviewer (JS and MK) conducted both phases of the quality assessments and synthesis, independently, in line with the review teams hypotheses. The review authors met and discussed their ratings until consensus was reached. No further clarification was needed from a third author.

RESULTS

Figure 1 summarizes the review screening and selection processes. Our searches identified 1267 unique citations after de-duplication. We reviewed 92 articles in full text after the title and abstract screening phase. In total, 25 articles satisfied the review teams criteria.[1,4,11,12,14,17–22,24,25,29,31,33,39,45–48,51,52,54,55] Seventeen (17) articles assessed the BPI-SF in 16 studies, while 8-articles evaluated the SF-MPQ-2 in 7-studies**.** Each of the tools had two articles (BPI-SF[21,22] and SF-MPQ-2[11,51]) that reported from one study population and their results were merged in this review. Some studies only focused on a particular subscale of the BPI-SF,[12,31,55] or did not assess the BPI-SF as a primary outcome,[14,19,21,45] whereas all the SF-MPQ-2 subscales were always examined as primary outcomes aside one that assessed only the neuropathic subscale.[39]

The characteristics of the MSK populations assessed in the included studies are summarized in **Table 1.** Eight studies evaluated different measurement properties of the BPI-SF[4,12,19,21,22,24,47,48] in Mixed-MSK population studies; three examined the SF-MPQ-2.[1,25,29] One or more measurement properties of the BPI-SF were assessed in Specific-MSK population studies with conditions including fibromyalgia,[31] back pain,[20,46,54] knee pain,[45] hip pain,[18] osteoporosis,[14] and arthritis[20,33,55] while subscales of SF-MPQ-2 were examined among knee pain,[29,52] complex regional pain syndrome (CRPS),[39] and acute back pain[11,51] patients. Although a wide range of measurement properties were assessed for both tools, studies reporting measurement error indices (MDC and SEM), cross-cultural validities, and interpretability properties (CID and MCID) were scarce.

Quality of the Included Studies

All the included studies were first inspected with MacDermid's quality assessment tool and **Table 2** summarizes the total quality score of each article. The BPI-SF articles received quality ratings within 50% to 86%, with a mean of 67% whereas the SF-MPQ-2 displayed higher quality ratings within 63% to 92% (mean 78%). This difference in quality rating in favor of the SF-MPQ-2 was not surprising because it was consistently assessed as a primary outcome in the included studies. The BPI-SF, however, was often utilized as a comparator or secondary outcome in 4 (23%) of the included studies[14,19,45,54] which impacted negatively on its mean quality rating for some criteria assessed with the MacDermid's quality assessment tool.

However, some quality issues were common to both tools: a) the absence of sample-size calculation or rationalization, b) imprecise or unclear hypotheses, c) limited or insufficient description of test procedures to an extent that allows replication of methods, d) absence or minimal reporting of error estimates (confidence intervals, SEM), and e); over-exaggerated or out-of-context conclusions/recommendations **(see Table 2).** Raters' agreement was excellent, as indicated by a high inter-rater reliability for the summary of their quality rating scores (ICC 0.87; 95% CI, 0.79–0.92) and unweighted kappa for individual item scores (0.75). All disagreements between the raters at this phase were clarified by the second author (JCM).

The results of the second phase of the quality assessment according to the COSMIN guidelines are summarized in the subsequent headings below, alongside evidence on the extracted measurement properties. **Tables 3 to 5** contain details on the BPI-SF measurement properties, the MSK population they represent, and their quality rating as assessed, first, with the COSMIN risk of bias (RoB), then, as benchmarked against the COSMIN quality criteria. **Table 6** summarizes the same details for the SF-MPQ-2**.** Finally, **Table 7** summarizes the result of the level of evidence synthesis for both tools, as per the modified GRADE.

Floor /Ceiling Effects, MCID and CID

Floor/ceiling effects occurs when a significant number of responders select scores that concentrate at the lowest (floor) and on the highest (ceiling) limits of a PROM.[49] When compared to the BPI-SF, subscales of the SF-MPQ-2 were more frequently associated with floor effects in studies assessing MSK conditions. Only Kapstad and colleaques[18] found significant flooring on the BPI-SF severity [21%] and interference [28%] subscales among total hip replacement patients 1 year post-op **(Table 4).** In contrast, three studies[1,11,25] reported flooring on the SF-MPQ-2 subscales: Lovejoy et al.[25] found significant flooring in a 'mixed' study population on the affective (28%), intermittent (15.1%) and neuropathic (12.4%) subscales of SF-MPQ-2; Adelmanesh et al[1] reported less significant floor effects (3.5-8.7%) while Dworkin et al.[11] noted some floor effects on the affective subscale (15%) among patients with acute back pain **(Table 6).**

Only one well-powered study (n= 1411)[31] reported MCID and CID within 2.09-2.89 for the BPI-SF severity and average pain item among fibromyalgia patients **(Table 4)**. No study has investigated the MCID or CID of the SF-MPQ-2.

Test-retest reliability

Few studies have examined the retest reliability of the BPI-SF and SF-MPQ-2 in MSK conditions. Three studies assessed the BPI-SF, but pooled evidence from two studies[4,55] displayed a high level of evidence supporting the interference subscale retest-reliability; the weighted ICC was 'sufficient' at 0.83. One report[4] of 'adequate' RoB rating was available in support of BPI-SF severity subscale retest reliability (ICC, 0.83) but further assessment with the modified GRADE suggest it represents a low level of evidence due to its low sample size (n=71) and RoB rating **(Table 3).** One study was rated 'doubtful' on the BPI-SF because the stability of patients was not certain.[33] Four studies examined the SF-MPQ-2 retest reliability: only one[17] received an 'adequate' RoB ratings, with ICC scores within 0.73-0.90 but represented a low level of evidence on the COSMIN Modified GRADE due to its small sample size **(Table 6 & 7)**. The remaining three studies[1,29,52] were rated 'doubtful' because of uncertainties pertaining to patients stability; for instance, inappropriate retest intervals that were long enough for change to occur (3 months)[52] or short enough (7-hours) to allow recall bias[1] were reported **(Table 6)**.

In conclusion, although not consistent across all its subscales, the COSMIN modified GRADE suggests a **‘**moderate’ level of evidence of ‘sufficient’ quality currently supports the BPI-SF retest reliability which is relatively stronger than the evidence for the SF-MPQ-2, with ‘very low’ level of evidence of ‘insufficient’ quality.

Internal consistency

Internal consistency was the most frequently reported form of reliability among the reviewed articles, and a high level of evidence supports both tools internal consistency in MSK studies. For the BPI-SF, eight studies with low RoB ratings (7 = ‘very good’; 1 = ‘adequate’)[4,12,18,20,33,48,55] reported Cronbach alpha within 0.82 - 0.96 **(Table 3)**. Of the five studies that examined the SF-MPQ-2, four[11,17,25,29] had ‘very good’ ratings for RoB, with Cronbach alpha within 0.88-0.96 (total score), and 0.75-0.92 (subscales scores). The only inadequate study failed to confirm the dimensionality of a translated version of the SF-MPQ-2 before reporting Cronbach alpha for the total scale score[1] **(Table 6).** In summary, the COSMIN Modified GRADE suggest a ‘high’ level of evidence of ‘sufficient’ quality supports both tools internal consistency in MSK conditions.

Structural Validity (hypothesis testing)

Studies have addressed questions on the multidimensional structure of the BPI-SF and SF-MPQ-2 in various MSK conditions. Seven reports with ‘very good’ RoB ratings examined the structural validity of the BPI-SF using factor analysis: four studies[4,20,33,48] were classified as ‘indeterminate’ because authors did not report details comparable to the COSMIN quality criteria; the remaining three studies[12,24,46] were ‘sufficient’ and displayed a high level of evidence in support of the BPI-SF two-factor structure in MSK. Two articles suggested a two-factor solution explaining the BPI-SF severity and interference subscales was more optimal than a three or one factor solution among mixed[24] and low back pain[46] patients, respectively. Also, Farrere et al.[12] confirmed the Portuguese interference subscale conformed with a one-factor solution, as originally hypothesized for the English version **(Table 4).**

Four studies examined the SF-MPQ-2 using factor analysis; however, pooled evidence from two high quality studies (‘very good’ RoB rating) displayed a conflicting quality of evidence in support of the SF-MPQ-2 factor structure in MSK condition. Although Dworkin et al.[11] proposed a four-factor solution for the SF-MPQ-2, all but the ‘neuropathic subscale’ did not confirm their proposed hypothesis. The study received a ‘very good’ risk of bias rating but was ‘insufficient’ when compared against the COSMIN quality criteria. Conversely, Lovejoy and co.[25] showed that factor analysis indices (root mean square error of approximation [RMSEA], Tucker-Lewis index [TLI], comparative fit index [CFI], Akaike information criterion [AIC]) favoured a four-factor solution over a one-factor solution **(Table 6)**; the study was ‘very good’ on RoB assessment and ‘sufficient’ against the COSMIN quality criteria. Two studies were not included in the evidence synthesis: the first was ‘inadequate’ on RoB assessment because it was underpowered (less than 100 participants)[29] while the other study[1] was ‘indeterminate’ from the lack of comparable details with the COSMIN criteria.

Turner and colleagues[52] used a ‘very good’ methodology in their Rasch analysis of the SF-MPQ-2 structure. Their findings suggest the SF-MPQ-2 is structurally unstable for use among patients with knee pain: the total scale score exhibited some form of dimensionality and differential item functioning. Furthermore, although the continuous, intermittent and affective subscales were unidimensional, some item misfit (1, 8 and 9) and disordered response thresholds were noted **(Table 6).** Packham et al.[39] also examined the SF-MPQ-2 neuropathic subscale structural stability using Rasch analysis among complex regional pain syndrome patients (CRPS). No signs of misfit or dimensionality was reported, and the level of item difficulty was adequate. Although the study received a ‘very good’ RoB rating, it was underpowered (n= 57) and thus, of ‘insufficient’ quality according to COSMIN standards.

Overall, synthesis based on the COSMIN Modified GRADE suggest a ‘high’ level of evidence of ‘sufficient’ quality supports the structural validity of the BPI-SF; hence, it comparatively better than the SF-MPQ-2 with ‘high’ level of evidence but of ‘conflicting’ quality **(Table 7).**

Known-groups validity

Authors have tested a spectrum of hypotheses to confirm the ability of the two outcome measures to differentiate known groups, but very few reports came from studies with good methodological approaches. Of the ten hypotheses tested with the BPI-SF, only four reports received ‘very good’ RoB ratings: Stubbs et al[47] demonstrated the ability of the BPI-SF to discriminate elderly patients with mixed MSK conditions into known groups of recurrent fallers and non-fallers, beyond chance (AUC, 0.72-0.73); however, two more hypotheses posited in the same study about the BPI-SF ability to differentiate fallers from none-fallers failed (AUC < 0.7).[47] Six additional hypotheses[20,46,55] examined the discriminative ability of the BPI-SF to categorize back pain and osteoarthritis patients into different groups of varying pain severities (Mild, Moderate, Severe, etc.). The findings were rated ‘inadequate’ on RoB assessment and ‘indeterminate’ against the COSMIN quality criteria from the use of poor statistical approaches and unclear definitions of known groups **(Table 4),** which was not different from our findings among the studies that reported on the SF-MPQ-2 known group validity[11,25] **(Table 6).**

To conclude, the COSMIN Modified GRADE indicates a ‘moderate’ level of evidence of ‘sufficient’ quality supports the BPI-SF known group validity in MSK studies while a ‘very low’ level of evidence of an ‘indeterminant’ quality supports the SF-MPQ-2 known group validity in MSK conditions **(Table 7).**

Criterion-convergent validity

Criterion-convergent validity was the most investigated psychometric property on the two questionnaires. A sizable number of health-related outcome measures (n=22) correlated with the BPI-SF and SF-MPQ-2 in different MSK populations, and all but 2 studies received 'very good' and 'sufficient' ratings on RoB assessment and against the COSMIN quality criteria. Furthermore, established correlations were mostly low-to-moderate in magnitude (BPI-SF, 78%; SF-MPQ-2, 67%) **(Table 4 & 6)**.

When disease/region specific PROMs like the Western Ontario and McMaster Universities Arthritis Index, Oswestry Disability Index, Roland Morris Disability Questionnaire were associated with subscales of the BPI-SF[18,20,22,33,46,48,55] and SF-MPQ-2, [17] relationships were predominantly low-to-moderate (r = 0.3-0.69). Similarly, low-to-moderate (r = 0.3-0.67) correlations were observed with mental/psychological status questionnaires: for instance, the SF-MPQ-2 correlated with the Beck Depression Inventory (BDI-II), the Generalized Anxiety Disorder-7 scale (GAD-7) and the Hospital Anxiety and Depression Scale,[11,25] while the BPI-SF subscales demonstrated an association with the Hospital Anxiety and Depression Scale.[12,20]

When generic PROMs were associated with the two questionnaires, correlations were mostly low-to-moderate (0.3-0.69) **(Table 4 & 6)**; however, a few high associations ($r \geq 0.7$) were seen in Mixed-MSK studies between subscales of the BPI-SF and the Short Form Health Questionnaire (SF 36)[20] and the Chronic Pain Grade scale.[20,22] Similarly, the SF-MPQ-2 correlated highly ($r \geq 0.7$) with the pain Numeric Rating Scale and the Pain Disability Index among complex regional pain syndrome patients;[39] the Visual Analogy Scale[1] and the Multidimensional Pain Index[25] in a mixed-MSK population, and with the SF-MPQ-2 parent versions (MPQ and SF-MPQ) among patients with knee pain.[29] In summary, a 'high' level of evidence of 'sufficient' quality supports both tools criterion-convergent validities in MSK populations, and established correlations were predominantly low-to-moderate in magnitude (*r* = 0.3-0.69) although a few higher correlations were seen with generic tools in mixed studies.

Responsiveness

The responsiveness of the BPI-SF was more extensively investigated than the SF-MPQ-2 in MSK conditions. Eighteen reports[14,18–21,45,48,54] were available on the BPI-SF responsiveness, but only five[14,19,21,54] received 'very good' RoB and 'sufficient' quality ratings. Two good quality studies, based on the construct approach, reported low-to-moderate correlations (0.30-0.69) between change scores of the BPI-SF, and the generic EQ-5D[54] (low back pain) and disease-specific Osteoporosis Quality of Life Questionnaire[14] (osteoporosis patients). The receiver operator curve (ROC) approach was adopted in the remaining three 'good' quality studies: in one report, the BPI-SF subscale was able to discriminate improved low back pain patients beyond chance (AUC ≥ 0.7)[54]; the other two reports in Mixed-MSK populations supported the ability of the BPI-SF to discriminate patients that experienced 'any improvement' from those with 'moderate improvement'.[19,21] However, the interference subscale did not satisfy the review teams criteria (AUC > 0.7).[19]

Nine reports[18–21,45,54] based on the effect size (ES) and standardized response mean (SRM) approach to responsiveness, were rated 'adequate' on RoB assessment but 'indeterminate' against the quality criteria. All nine reports lacked clearly stated hypotheses that defined the magnitude and direction of expected change: for instance, it was impossible to tell if the reported ES or SRM supported the effectiveness of administered interventions, or the responsiveness of the outcome measure to capture change as expected.[36] Four reports from three studies[20,48,55] were also rated 'inadequate' and 'insufficient' from their poor design or use of suboptimal statistical approaches, such as Guyatt statistic, co-variances, t-test statistic or significant P-values **(Table 5).**

Four studies[1,11,25,39] addressed the SF-MPQ-2 responsiveness in MSK conditions: one[39] among CRPS patients was 'adequate' on RoB assessment but 'indeterminant' against the COSMIN quality standards because the authors did not define the expected change directions or magnitude. The remaining three studies were 'inadequate' and 'indeterminant' because authors employed suboptimal statistical approaches **(Table 6).** In summary, evidence synthesis according to the COSMIN Modified GRADE suggest the BPI-SF has a 'high' level of evidence of 'sufficient' quality in support of its responsiveness from 5 studies. On the contrary, a 'very low' level of evidence of 'indeterminant' quality was seen across seven studies for the responsiveness of the SF-MPQ-2.

DISCUSSION

Our systematic review indicates that better quality of evidence currently supports the psychometric properties of the BPI-SF over the SF-MPQ-2. Although not a head-to-head comparison, evidence synthesis based on COSMIN guidelines[36–38,41,42,49] suggest both tools have high-quality evidence supporting their internal consistency and criterion-convergent validities. However, sufficient evidence of high-to-moderate quality only supports the BPI-SF responsiveness, retest reliability, known group validity and structural validities as compared to the SF-MPQ-2 in MSK conditions **(Table 7).** In addition, more articles described floor effects on the SF-MPQ-2 than was reported with the use of the BPI-SF.

Two different, but complementary approaches to appraising the quality of measurement evidence of outcome tools were employed in this review. Although quality ratings favoured the SF-MPQ-2 over the BPI-SF when the structured clinical quality assessment tool was used, quality assessed according to the COSMIN guidelines favoured the BPI-SF. The SF-MPQ-2 only received higher quality rating in the initial phase of assessment because it was often the primary focus of psychometric investigations, not because the reports were void of bias. However, more studies were available on the BPI-SF and a reasonable number assessed measurement properties with sound methodological approaches which favoured the BPI-SF quality ratings when assessed with the COSMIN guidelines. This finding suggest that the quality or comprehensiveness of a study report does not rule out the tendencies of bias in the evidence; however, insufficient reporting can expose otherwise good evidence to bias.

Many of the included studies (42%) reported remarkable floor effects on select subscales of the SF-MPQ-2. This may relate to the quality/characteristic of pain captured on the 22-item descriptors of the SF-MPQ-2.[10] Not all characteristics of pain may be similarly represented in MSK conditions and the variation in patient scores on the SF-MPQ-2 subscales may selectively reflect unique pain phenotypes within MSK conditions. This ultimately predisposes some of the subscales of the SF-MPQ-2 to floor effect. For instance, high mean scores on the neuropathic pain subscale is unlikely if participants perceive pain as nociceptive. Similarly, it is unlikely for very resilient participants to report high scores on the SF-MPQ-2 affective subscale. Therefore, the floor effect on subscales of the SF-MPQ-2 does not necessarily represent redundancy but reflect its discriminative nature. Future studies may explore this speculation.

The available high-quality evidence on internal consistency and structural validity supports the stability of the BPI-SF. Cronbach alpha was satisfactory and the original 2-factor structure (severity and interference), as hypothesized during the BPI-SF development, was reproduced.[7,8] However, it was notable that several studies[4,20,33,48] had an indeterminate quality assessment rating because authors failed to report comparable indices even though they mostly suppose the BPI-SF conforms with a 2-factor solution **(Table 4)**. There is yet to be a Rasch analysis on the BPI-SF structural validity in a well-represented MSK population; the single Rasch paper identified during screening[53] failed to satisfy our inclusion criteria ≥ 70% threshold for MSK participants. Therefore, future authors should review the COSMIN quality criteria[36] when planning and documenting studies on the structural validity of the BPI-

SF, and a Rasch analysis should be performed in a population that sufficiently represents a spectrum of musculoskeletal conditions.

For the SF-MPQ-2, internal consistency was established, but a conflicting (+/-) quality of evidence displayed with its use in different MSK conditions. This happens periodically with multidimensional tools and may suggest that the questionnaire's factor structure varies with the context of use.[2] Mixed results of the item response examination of the SF-MPQ-2 using Rasch methods has been documented. Packham and colleagues[36] utilized 'very good' methods to examine the SF-MPQ-2 neuropathic subscale with the Rasch. Although their evidence supported the independent use of the neuropathic scale for CRPS assessment, the study was rated insufficient according to the COSMIN quality criteria because of its small sample size (n=57). In contrast, Turner and colleagues[52] showed that the SF-MPQ-2 structure did not fit with the Rasch as assessed among a representative sample of knee pain subjects. Overall, it is too early to draw conclusions from the conflicting evidence on the SF-MPQ-2 structural validity in MSK conditions until more high-quality evidence emerges from studies conducted in different MSK populations. Pooled evidence from such studies could be synthesized to provide conclusive evidence in future reviews on the SF-MPQ-2.

Evidence for convergent-criterion related validities have been reported for both assessment tools in studies of high quality. It was noteworthy that the reported correlation with disease-specific/regional tools were consistently low-to-moderate (rho = 0.3-0.69) in most of the included studies. This could be from differences in concepts used to describe pain in generic and specific tools. While generic tools are structured to capture pain in multiple conditions, specific/regional tools are designed to elicit responses on patients' pain in homogenous conditions. Although not entirely clear, these differences could underpin why correlations were only low-to-moderate when the tools were used in homogenous conditions.

Test-retest reliability was among the least examined measurement properties across the articles: estimates on measurement error (SEM and MDC) were completely lacking. Only the interference subscale of the BPI-SF has been shown to have satisfactory intraclass correlation coefficient in support of its test-retest reliability and has received a high-quality rating for use in MSK conditions. Both the BPI-SF severity subscale and the SF-MPQ-2 questionnaire lack high quality studies on their retest reliability. Some recurring flaws identified during the risk of bias assessment included: a) non-specification of the ICC model used in study analysis, preferably the two-way random effect model,[36] b) low sample sizes, c) employment of suboptimal retest-intervals, and d) the unverified assumption of participant stability. Therefore, well-powered high-quality studies that measure the test-retest reliability of the BPI-SF severity subscale and the SF-MPQ-2 are needed. Again, authors should consider adhering to the COSMIN reporting guidelines, and vague descriptions of patient stability should be avoided; if necessary, a Global Rating of Change score can be used to confirm stability in doubtful situations. Finally, robust estimates on the measurement error indices (SEM and MDC) of both tools should be pursued in different MSK populations.

One of the most important findings from this review relates to evidence backing the two questionnaires responsiveness and known group validity in MSK conditions. Although high quality evidence supports the BPI-SF responsiveness in MSK conditions,[14,19,21,54] studies that assessed responsiveness using the standardized response mean (SRM) and effect size (ES) were ‘indeterminant’ because authors did not provide well-defined hypotheses with clear magnitude and direction of expected change.[36,38] The expected effectiveness or ineffectiveness of interventions differ and hypotheses testing responsiveness have to be precise and based on the context. For instance, smaller outcome measure change score indices are predicted in a study of a specific low intensity intervention than a high intensity intervention in the same population. Although the reported change scores would differ, both would generate support for the ability of the measure to detect change if the change scores were proportional to the intervention intensity. It is possible that authors ignored the importance of documenting this evidence, or, such hypotheses were not defined initially in study protocols because most computations of ES and SRM responsiveness indices came from RCTs which had a secondary aim of examining outcome measurement performance.[19,21,54,55] For known group validity, we only observed a moderate level of evidence in support of the BPI-SF since the available evidence of sufficient quality focused on older adults (indirectness). Future studies examining the responsiveness of BPI-SF, based on the ES and SRM approach need to provide precise hypotheses that specify the magnitude and direction of expected change. Furthermore, high quality studies are still needed in diverse age groups to explore known group validity of the BPI-SF. Unfortunately, we are unsure of the evidence on the known group validity and responsiveness of the SF-MPQ-2 in MSK population because studies of 'very low' and indeterminant quality rating were found for both properties in MSK conditions. This gap should be addressed in high-quality studies, bearing in mind that the SF-MPQ-2 was primarily developed/expanded to be able to discriminate patients by the quality of pain they experience,[10] and it is highly important that clinicians/researchers are sure that a pain assessment tool is able to detect a change in the patient's condition.[2,28,32]

Valid and accurate MCID and CID estimates are important for outcomes evaluation, prognostication and communication among health care professionals.[28,32,44] Unfortunately, only one well-powered study has reported the MCID and CID indices of the BPI-SF severity subscale and the average pain item among fibromyalgia using the anchor-based approach.[45] This lone report, however, is inadequate to permit informed decision-making considering the diversity of MSK conditions with multiple conditions requiring estimates on their MCID and CID, and the lack of estimate for the interference subscale. Even worst, no study has assessed the SF-MPQ-2 MCID and CID in any MSK population. Multiple studies aimed at estimating the MCID and CID of the two questionnaires are urgently recommended since these estimates determine the level of significance associated with treatments when clinicians/researchers use the tools.[2,27,28]

In the present review, different ‘mixed’ and ‘specific’ MSK conditions were investigated with the two questionnaires but it was obvious that none of the studies exclusively reported the measurement properties of the tools in upper extremity conditions (e.g. carpal tunnel syndrome, tennis elbow, shoulder pain/dysfunction) and neck-related MSK conditions. Nonetheless, categories of MSK conditions such as neck and shoulder pain are common and frequently present to clinicians. Future studies should consider investigating the two questionnaires measurement properties in these classes of MSK conditions. Moreover, it will be of great benefit to the clinician managing mixed or peculiar upper extremity and neck MSK pathologies to be able to use outcome measures that not only assess the multidimensional nature of pain, but at the same time yield scores comparable across different studies.

LIMITATIONS

Our study has some limitations. First, our inclusion criteria considered only studies reporting measurement properties from sample populations with ≥70% MSK sufferers. Therefore, relevant reports of individuals with MSK disorders may have been overlooked or omitted because the reports did not meet our defined inclusion criteria. Second, we were unable to conduct a meta-analysis on the reported measurement properties across the included studies. This was due to gross differences in study methodology, MSK population characteristics, and the time intervals adopted in individual studies. However, our summary tables and narrative synthesis should be comprehensive enough to allow the clinician/researcher to understand the measurement properties accompanying the tools in peculiar MSK conditions, while at the same time, having an idea of their general performance in MSK conditions.

Third, in comparing the measurement properties of the two questionnaires, based on the COSMIN Modified GRADE approach, we did not acknowledge the differences in their culturally adapted/translated versions. While clinometric experts discourage this, it may not be problematic in this review because the evidence backing the culturally adapted/translated versions were mostly similar to the original versions. For example, culturally adapted/translated versions and original versions exhibited similar factor structures, their internal consistency estimates were within a similar range, and when cross-cultural adaptations were performed, authors confirmed compliance with standardized procedures to ensure content reflected the same concepts with the tools' original versions. Nonetheless, specific details on the adapted/translated versions measurement properties and their quality ratings are available for consideration in our result **Tables 3 to 6**.

CONCLUSION

Although the SF-MPQ-2 presents potential, a greater volume of better-quality evidence was found in support of the BPI-SF measurement properties, including its responsiveness, retest reliability, known group validity and structural validities, which suggest it is currently better for pain assessment in MSK condition. Further investigation of (a) the retest reliability of the BPI-SF severity subscale; (b) the SF-MPQ-2 structural validity, known group validity, retest reliabilities, and responsiveness; and (c) the two questionnaires cross-cultural validities, interpretability properties (MCID and CID), and measurement error indices (SEM and MDC) is needed in multiple MSK studies of high quality.

REFERENCES

1. Adelmanesh F, Jalali A, Attarian H, et al. Reliability, Validity, and Sensitivity Measures of Expanded and Revised Version of the Short-Form McGill Pain Questionnaire (SF-MPQ-2) in Iranian Patients with Neuropathic and Non-Neuropathic Pain. *Pain Med*. 2012;13(12):1631-1638. doi:10.1111/j.1526-4637.2012.01517.x.

2. Bouffard J, Bertrand-Charette M, Roy J-S. Psychometric properties of the Musculoskeletal Function Assessment and the Short Musculoskeletal Function Assessment: a systematic review. *Clin Rehabil*. 2016;30(4):393-409. doi:10.1177/0269215515579286.

3. Briggs AM, Cross MJ, Hoy DG, et al. Musculoskeletal Health Conditions Represent a Global Threat to Healthy Aging: A Report for the 2015 World Health Organization World Report on Ageing and Health. *Gerontologist*. 2016;56(Suppl 2):S243-S255. doi:10.1093/geront/gnw002.

4. Celik EC, Yalcinkaya EY, Atamaz F, et al. Validity and reliability of a Turkish Brief Pain Inventory Short Form when used to evaluate musculoskeletal pain. *J Back Musculoskelet Rehabil*. 2017;30(2):229-233. doi:10.3233/BMR-160738.

5. Chapman JR, Norvell DC, Hermsmeyer JT, et al. Evaluating Common Outcomes for Measuring Treatment Success for Chronic Low Back Pain. *Spine (Phila Pa 1976)*. 2011;36(21 SUPPL.):S54-S68. doi:10.1097/BRS.0b013e31822ef74d.

6. Chiarotto A, Terwee CB, Ostelo RW. Choosing the right outcome measurement instruments for patients with low back pain. *Best Pract Res Clin Rheumatol*. 2016;30(6):1003-1020. doi:10.1016/j.berh.2017.07.001.

7. Cleeland CS. The Brief Pain Inventory- User Guide. https://www.mdanderson.org/documents/Departments-and-Divisions/Symptom-

Research/BPI_UserGuide.pdf. Published 2008.

8. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. *Ann Acad Med Singapore*. 1994;23(2):129-138. doi:10.1016/0029-7844(94)00457-O.

9. Dworkin RH, Turk DC, Farrar JT, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain*. 2005;113(1):9-19. doi:10.1016/j.pain.2004.09.012.

10. Dworkin RH, Turk DC, Revicki DA, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). *Pain*. 2009;144(1):35-42. doi:10.1016/j.pain.2009.02.007.

11. Dworkin RH, Turk DC, Trudeau JJ, et al. Validation of the Short-Form McGill Pain Questionnaire-2 (SF-MPQ-2) in Acute Low Back Pain. *J Pain*. 2015;16(4):357-366. doi:10.1016/j.jpain.2015.01.012.

12. Ferreira-Valente MA, Pais-Ribeiro JL, P. Jensen M. Further Validation of a Portuguese Version of the Brief Pain Inventory Interference Scale. *Clínica y Salud*. 2012;23(1):89-96. doi:10.5093/cl2012a6.

13. Grieve S, Perez RSGM, Birklein F, et al. Recommendations for a first Core Outcome Measurement set for complex regional PAin synd1. Grieve S, Perez RSGM, Birklein F, et al. Recommendations for a first Core Outcome Measurement set for complex regional PAin syndrome Clinical sTudies (COMPACT). 2017. 2017. doi:10.1097/j.pain.0000000000000866.

14. Group O. Measuring Quality of Life in Women with Osteoporosis. *Osteoporos Int*. 1997;7(5):478-487. doi:10.1007/PL00004151.

15. Hoy D, March L, Brooks P, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Ann Rheum Dis*. 2014;73(6):968-974.

doi:10.1136/annrheumdis-2013-204428.

16. Jumbo S, MacDermid J, Michael K, Packham TL, Athwal GS, Faber K. Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review Protocol. *Submitted*. 2018.

17. Kachooei AR, Ebrahimzadeh MH, Erfani-Sayyar R, Salehi M, Salimi E, Razi S. Short Form-McGill Pain Questionnaire-2 (SF-MPQ-2): A Cross-Cultural Adaptation and Validation Study of the Persian Version in Patients with Knee Osteoarthritis. *Arch bone Jt Surg*. 2015;3(1):45-50. doi:10.22038/ABJS.2015.3827.

18. Kapstad H, Rokne B, Stavem K. Psychometric properties of the Brief Pain Inventory among patients with osteoarthritis undergoing total hip replacement surgery. *Health Qual Life Outcomes*. 2010;8(1):148. doi:10.1186/1477-7525-8-148.

19. Kean J, Monahan PO, Kroenke K, et al. Comparative Responsiveness of the PROMIS Pain Interference Short Forms, Brief Pain Inventory, PEG, and SF-36 Bodily Pain Subscale. *Med Care*. 2016;54(4):414-421. doi:10.1097/MLR.0000000000000497.

20. Keller S, Bann CM, Dodd SL, Schein J, Mendoza TR, Cleeland CS. Validity of the Brief Pain Inventory for Use in Documenting the Outcomes of Patients With Noncancer Pain. *Clin J Pain*. 2004;20(5):309-318. doi:10.1097/00002508-200409000-00005.

21. Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K. Comparative Responsiveness of Pain Outcome Measures Among Primary Care Patients With Musculoskeletal Pain. *Med Care*. 2010;48(11):1007-1014. doi:10.1097/MLR.0b013e3181eaf835.

22. Krebs EE, Lorenz KA, Bair MJ, et al. Development and Initial Validation of the PEG, a Three-item Scale Assessing Pain Intensity and Interference. *J Gen Intern Med*. 2009;24(6):733-738. doi:10.1007/s11606-009-0981-1.

23. Kroenke K, Krebs EE, Turk D, et al. Core Outcome Measures for Chronic Musculoskeletal Pain Research: Recommendations from a Veterans Health Administration Work Group. *Pain Med*. January 2019. doi:10.1093/pm/pny279.

24. Lapane KL, Quilliam BJ, Benson C, Chow W, Kim M. One, Two, or Three? Constructs of the Brief Pain Inventory Among Patients With Non-Cancer Pain in the Outpatient Setting. *J Pain Symptom Manage*. 2014;47(2):325-333. doi:10.1016/j.jpainsymman.2013.03.023.

25. Lovejoy TI, Turk DC, Morasco BJ. Evaluation of the Psychometric Properties of the Revised Short-Form McGill Pain Questionnaire. *J Pain*. 2012;13(12):1250-1257. doi:10.1016/j.jpain.2012.09.011.

26. MacDermid JC, Law M, Michlovitz S. Outcome measurement in evidence-based rehabilitation. In: Law M, MacDermid JC, eds. *Evidence-Based Rehabilitation: A Guide to Practice*. 3rd ed. Thorofare NJ, USA: Slack Incorporated; 2014:65-104.

27. MacDermid JC, Stratford P. Applying evidence on outcome measures to hand therapy practice. *J Hand Ther*. 2004;17(2):165-173. doi:10.1197/j.jht.2004.02.005.

28. MacDermid JC, Walton DM, Avery S, et al. Measurement Properties of the Neck Disability Index: A Systematic Review. *J Orthop Sport Phys Ther*. 2009;39(5):400-C12. doi:10.2519/jospt.2009.2930.

29. Maruo T, Nakae A, Maeda L, et al. Validity, Reliability, and Assessment Sensitivity of the Japanese Version of the Short-Form McGill Pain Questionnaire 2 in Japanese Patients with Neuropathic and Non-Neuropathic Pain. *Pain Med*. 2014;15(11):1930-

1937. doi:10.1111/pme.12468.

30. McGuire DB. Comprehensive and multidimensional assessment and measurement of pain. *J Pain Symptom Manage*. 1992;7(5):312-319. doi:10.1016/0885-3924(92)90064-O.

31. Mease PJ, Spaeth M, Clauw DJ, et al. Estimation of minimum clinically important difference for pain in fibromyalgia. *Arthritis Care Res (Hoboken)*. 2011;63(6):821-826. doi:10.1002/acr.20449.

32. Mehta SP, MacDermid JC, Richardson J, MacIntyre NJ, Grewal R. A Systematic Review of the Measurement Properties of the Patient-Rated Wrist Evaluation. *J Orthop Sport Phys Ther*. 2015;45(4):289-298. doi:10.2519/jospt.2015.5236.

33. Mendoza T, Mayne T, Rublee D, Cleeland C. Reliability and validity of a modified Brief Pain Inventory short form in patients with osteoarthritis. *Eur J Pain*. 2006;10(4):353-353. doi:10.1016/j.ejpain.2005.06.002.

34. Mendoza TR, Chen C, Brugger A, et al. The utility and validity of the modified brief pain inventory in a multiple-dose postoperative analgesic trial. *Clin J Pain*. 2004;20(5):357-362.

35. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med*. 2009;6(7):e1000097. doi:10.1371/journal.pmed.1000097.

36. Mokkink LB, Prinsen CAC, Patrick DL, et al. *COSMIN Methodology for Systematic Reviews of Patient - Reported Outcome Measures (PROMs)*.; 2018.

37. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*. 2010;63(7):737-745.

doi:10.1016/j.jclinepi.2010.02.006.

38. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res*. 2018;27(5):1171-1179. doi:10.1007/s11136-017-1765-4.

39. Packham TL, Bean D, Johnson MH, et al. Measurement Properties of the SF-MPQ-2 Neuropathic Qualities Subscale in Persons with CRPS: Validity, Responsiveness, and Rasch Analysis. *Pain Med*. October 2018. doi:10.1093/pm/pny202.

40. Perruccio A V., Yip C, Power JD, Canizares M, Badley EM. Discordance Between Population Impact of Musculoskeletal Disorders and Scientific Representation: A Bibliometric Study. *Arthritis Care Res (Hoboken)*. 2019;71(1):56-60. doi:10.1002/acr.23583.

41. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1147-1157. doi:10.1007/s11136-018-1798-3.

42. Prinsen CAC, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a “Core Outcome Set” - a practical guideline. *Trials*. 2016;17(1):449. doi:10.1186/s13063-016-1555-2.

43. Ramasamy A, Martin ML, Blum SI, et al. Assessment of Patient-Reported Outcome Instruments to Assess Chronic Low Back Pain. *Pain Med*. 2017;18(6):1098-1110. doi:10.1093/pm/pnw357.

44. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102-109. doi:10.1016/j.jclinepi.2007.03.012.

45. Risser RC, Hochberg MC, Gaynor PJ, D’Souza DN, Frakes EP. Responsiveness of the

Intermittent and Constant Osteoarthritis Pain (ICOAP) scale in a trial of duloxetine for treatment of osteoarthritis knee pain. *Osteoarthr Cartil*. 2013;21(5):691-694. doi:10.1016/j.joca.2013.02.007.

46. Song C-Y, Lin S-F, Huang C-Y, Wu H-C, Chen C-H, Hsieh C-L. Validation of the Brief Pain Inventory in Patients With Low Back Pain. *Spine (Phila Pa 1976)*. 2016;41(15):E937-E942. doi:10.1097/BRS.0000000000001478.

47. Stubbs B, Eggermont L, Patchay S, Schofield P. Older adults with chronic musculoskeletal pain are at increased risk of recurrent falls and the brief pain inventory could help identify those most at risk. *Geriatr Gerontol Int*. 2015;15(7):881-888. doi:10.1111/ggi.12357.

48. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the brief pain inventory for chronic nonmalignant pain. *J Pain*. 2004;5(2):133-137. doi:10.1016/j.jpain.2003.12.005.

49. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42. doi:10.1016/j.jclinepi.2006.03.012.

50. Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res*. 2009;18(8):1115-1123. doi:10.1007/s11136-009-9528-5.

51. Turk DC, Dworkin RH, Trudeau JJ, et al. Validation of the Hospital Anxiety and Depression Scale in Patients With Acute Low Back Pain. *J Pain*. 2015;16(10):1012-1021. doi:10.1016/j.jpain.2015.07.001.

52. Turner K V, Moreton BM, Walsh DA, Lincoln NB. Reliability and responsiveness of measures of pain in people with osteoarthritis of the knee: a psychometric evaluation.

Disabil Rehabil. 2017;39(8):822-829. doi:10.3109/09638288.2016.1161840.

53. Vos T, Barber RM, Bell B, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015;386(9995):743-800. doi:10.1016/S0140-6736(15)60692-4.

54. Whynes DK, McCahon RA, Ravenscroft A, Hodgkinson V, Evley R, Hardman JG. Responsiveness of the EQ-5D Health-Related Quality-of-Life Instrument in Assessing Low Back Pain. *Value Heal*. 2013;16(1):124-132. doi:10.1016/j.jval.2012.09.003.

55. Williams VSL, Smith MY, Fehnel SE. The Validity and Utility of the BPI Interference Measures for Evaluating the Impact of Osteoarthritic Pain. *J Pain Symptom Manage*. 2006;31(1):48-57. doi:10.1016/j.jpainsymman.2005.06.008.

Figure 1: Flow diagram of the systematic review according to PRISMA

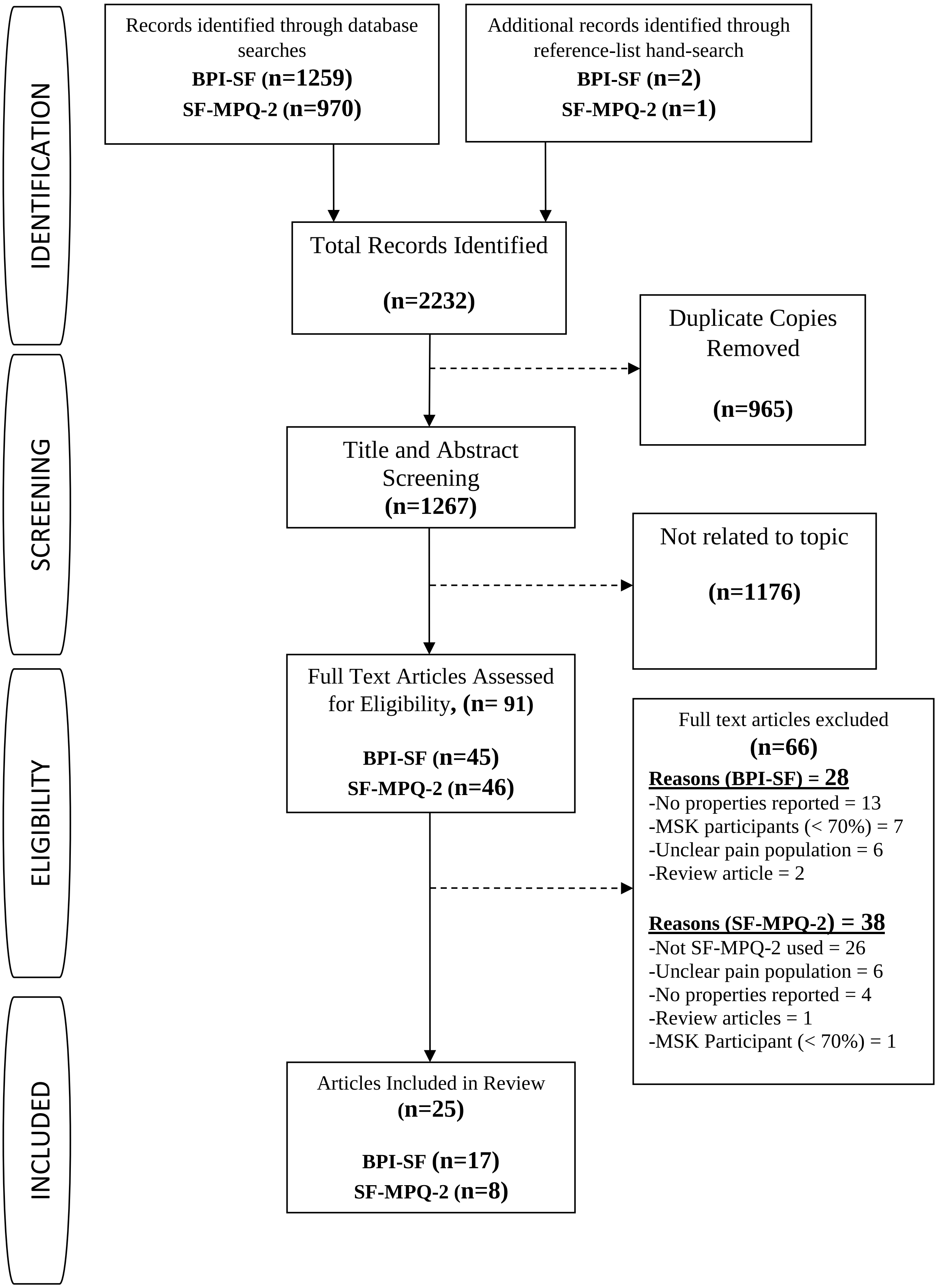

IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Records identified through database searches
BPI-SF (n=1259)
SF-MPQ-2 (n=970)
Additional records identified through reference-list hand-search
BPI-SF (n=2)
SF-MPQ-2 (n=1)
Total Records Identified
(n=2232)
Duplicate Copies Removed
(n=965)
Title and Abstract Screening
(n=1267)
Not related to topic
(n=1176)
Full Text Articles Assessed for Eligibility, (n= 91)
BPI-SF (n=45)
SF-MPQ-2 (n=46)
Full text articles excluded
(n=66)
Reasons (BPI-SF) = 28
-No properties reported = 13
-MSK participants (< 70%) = 7
-Unclear pain population = 6
-Review article = 2
Reasons (SF-MPQ-2) = 38
-Not SF-MPQ-2 used = 26
-Unclear pain population = 6
-No properties reported = 4
-Review articles = 1
-MSK Participant (< 70%) = 1
Articles Included in Review
(n=25)
BPI-SF (n=17)
SF-MPQ-2 (n=8)

TABLE 1. CHARACTERISTICS OF STUDIES ADDRESSING PSYHCOMETRICS OF THE BPI-SF AND SF-MPQ-2

First Author (Year)
Country;
Setting;
Version
	Population	**Design**	**Sample Characteristics**	**Measurement Properties Evaluated**	**Intervention/Retest**

Celik (2017)[4] Turkey; 3-outpatient physiotherapy dept.; **BPI-SF** (Turkish version)	Patients: Turkish speaking patients previously on routine outpatient physiotherapy MSK distribution: 13.8% Upper extremity; 20.9% Lower extremity; 65.3% Spine	Cross sectional	N = 287 Sex = Males, 26.9% Age = M, 49.72 (SD 12.92) yrs.	Reliability Validity	Intervention: Usual care Retest: Baseline, general baseline & 7days, n=71-test-retest.
Group O. (1997)[14] Canada; Rheumatologist practice (Canada) and 4 metabolic bone diseases practice USA; **BPI-SF** (English)	Patients: Patients who had at least one osteoporosis-induced vertebral fracture with a clinical diagnosis of chronic back pain and (or) osteoarthritis	Longitudinal cohort	N = 226 Sex = Females, 100% Age = 50yrs and above	Responsiveness	Intervention: Oestrogen replacement therapy; Cyclic Etidronate; Calcium; Calcitonin Retest: Baseline, 2weeks and 6months
Ferreira-Valente (2012)[12] Portugal; 7 health institutions; **BPI-SF** (Portuguese version *focused on Interference scale	Patients: Participants recruited had a history of chronic MSK lasting 2years (71%) and 10years (38.2%)	Cross-sectional Study	N = 214 Sex = Females, 66.1% Age = M, 60.18 (SD 14.87) yrs.	Validity Reliability	Intervention: non Retest: Baseline.

Kapstad (2010)[18]
Norway;
6 hospitals in 3 counties;
BPI-SF (Norwegian version)
	Patients: Patients on wait-list for total hip replacement surgery and had satisfactory proficiency of the Norwegian language	Prospective cohort	N = 250
			Sex = Females, 70%
			Age = M, 68.7 (SD 9.9) yrs.
				Validity
				Reliability
				Responsiveness
				Floor or ceiling effect
					Intervention: Total hip replacement surgery
					Retest: Baseline and 1-year post-surgery

Kean (2016)[19] USA; 5 primary care centres for Veteran in Indianapolis; **BPI-SF** (English)	Source: Pooled data from a study on veterans with moderate to severe persistent musculoskeletal pain MSK distribution: Fibromyalgia, wide spread pain, pain at the joints, limbs, back and neck	RCT	N = 244 Sex = Females, 83% Age = M, 55.1 yrs.	Responsiveness	Intervention: Not described Retest: Baseline and 3 months interval

Keller(2004)[20]
USA;
10 primary care centres;
BPI-SF (English)
	Patients: Patients had a primary diagnosis of osteoarthritis, rheumatoid arthritis, low back pain (on workers' compensation) and Low Back Pain (not on workers' compensation)	Cross sectional	N = 250
			Sex (not reported)
			Age (not reported)
				Validity
				Reliability
				Responsiveness
					Intervention: Routine treatment
					Retest: Baseline and patients next follow-up visit

Krebs (2009)[22] and (2010)[21]
USA;
10 primary health centres;
BPI-SF (English)
	Source: Pooled data from 3 studies. The authors compared the responsiveness of 4 outcome measures
	MSK distribution: Majorly at the lower back, hip and knee region
		RCT;
		Longitudinal Cohort
			N = 427
			Sex = 53.7%
			Age = M, 59.1 (SD 13) yrs.
				Responsiveness
				Validity
					Intervention: Depression medication + Pain-self management VS Usual care
					Retest: Baseline and 12 months

Lapane (2014)[24]
USA;
48 clinical sites;
BPI-SF (English)
	Source: Pooled data from a registry designed to provide detailed prospective pain assessment of oxycodone users
	MSK distribution: Above 70% of population with mixed chronic MSK pain including fibromyalgia, arthritis, back and neck pain
		Prospective cohort	N = 741
			Sex= Females, 59.6%
			Age= M, 49.8 (SD 13.1) yrs.
				Validity	Intervention: Prescribed Oxycodone prescription
					Retest: Baseline

CONTINUED.

Mease (2011)[31] USA; not specified; **BPI-SF** (English) *focused on average pain and severity pain scores	Source: Pooled data from 4 RCTs. Participants all had a diagnosis of fibromyalgia and were randomized into control and placebo groups	RCT	N = 1411 Sex = Females, 94.9% Age = M, 50.3 (SD 10.44) yrs.	Interpretability (MCID and CID)	Intervention: Control group - duloxetine Placebo group – sham Retest: Baseline and 12 weeks

Mendoza (2006)[33]
USA;
(not specified);
BPI-SF; 10 items Modified short form (English)
	Source: Pooled data from 2 RCT studies. Participants suffering from Hip OA (study 1) and knee OA (study 2) were recruited and randomized into control and placebo groups	RCT	N = 467 & 1019
			Sex = Males ≈ 35.3%
			Age ≈ M, 62.3 yrs.
				Validity
				Reliability
					Intervention: Valdecoxib + Naproxen
					Retest: Baseline, 1 week, and 2 weeks

Risser (2013)[45] USA; Not specified; **BPI-SF** (English) *focused on average pain scale	Source: Pooled data from 2 RCT studies. Participants had knee pain in the last 3months for a duration of at least 14 days/1-month	RCT	N = 524 Sex = Females, 57.1% Age = M, 61 yrs.	Responsiveness	Intervention: Control NSAID + Duloxetine Placebo: Sham Retest: Baseline and 24hours

Song (2016)[46]
Taiwan;
Physiotherapy department;
BPI-SF (Chinese version)
	Patients: Participants recruited had a diagnosis of Low back pain from conditions including spondylolysis, spondylolisthesis, herniated intervertebral disc, scoliosis and sciatica	Cross-sectional	N = 271
			Sex = Males, 119; Females, 152
			Age = M, 57.1 (SD 16.2) yrs.
				Validity	Intervention: None
					Retest: Baseline

Stubbs (2015)[47]
United Kingdom;
10 elderly homes;
BPI-SF (English)
	Patient: Older adults were surveyed in elderly homes on falls resulting from chronic MSK pain. They were divided into 2 groups: study group (with fall), and control group (without fall or MSK) to be compared	Cross sectional	N = 298
			Sex = females ≈ 67.4%
			Age = M, 76.6 (SD 8.5) yrs.
				Validity	Intervention: None
					Retest: Baseline

Tan (2003)[48] USA; Chronic pain unit of Veteran affairs medical centre; **BPI-SF** (English)	Patient: Participants were referred from several specialities, including surgery, rheumatology, physical medicine and rehabilitation MSK distribution: About 50% reported pain of multiple sites (including back pain), and 28.8% reported primary back pain	Cross sectional	N = 440 Sex = Males, 91.8% Age = M, 54.9 yrs.	Validity Reliability Responsiveness	Intervention: Usual care Retest: Baseline and follow-up visits

Whynes (2013)[54]
United Kingdom;
Nottingham - setting (Not described);
BPI-SF (English)
	Source: Pooled data was from participants in a study of epidural steroid injections to alleviate low back pain	RCT	N = 37
			Sex (not reported)
			Age (not reported)
				Responsiveness	Intervention: Epidural injection
					Retest: Baseline, 7days, and 12 weeks

Williams (2006)[55] USA; Multiple Medical centre; **BPI-SF** (English) *Focus on interference scale	Source: Pooled data from 2 studies conducted on participants with a primary diagnosis of OA at any region and history of moderate to severe pain for at least 1 month	RCT	N = 106 & 239 Sex = Females, 62.5 & 72.9% Age ≈ M, 63.1 (SD 9.8) yrs.	Validity Reliability Responsiveness	Intervention: Controlled release Oxycodone Retest: Study 1- Baseline, 7days, 14day; Study 2- Baseline, 15,30,45,60, 90 days; Test retest- Day 7 – Day 14

CONTINUED.

Adelmanesh (2012)[1]
IRAN;
Tertiary Pain & Rehab clinic;
SF-MPQ-2 (Persian)
	Patient: Mixed sub-acute and chronic pain patients; Persian-speaking; 74-patients with diabetic neuropathic; 184-patient with pain including myofascial pain, epicondylitis, knee and neck OA, Low back pain	Cross sectional	N = 258
			Sex = Female, 55%
			Age = M, 42.53 (SD 11.93) yrs.
				Validity
				Reliability
				Responsiveness
					Intervention:
					a.) Diabetic neuropathic patients: pre-gabalin, shoe modification education, and physiotherapy)
					b.) MSK patients: Physical therapy, Acupuncture, NSAIDS
					Retest: Baseline and 3-weeks; Test-retest, baseline and 7-hours

Dworkin (2014)[11] & Turk (2015)[51] USA; Research setting of an RCT study, not specified; **SF-MPQ-2** (English)	Patients: Acute low back pain patients with radiating pain to at least one leg. Pain duration was within 30days	RCT	N = 664 Sex= Males, 50% Age= M, 45 yrs.	Validity Reliability Responsiveness Floor or Ceiling effect	Intervention: Oxycodone and Tanpotalone Retest: Baseline & 10days

Kachooei (2014)[17]
IRAN;
Knee pain clinic;
SF-MPQ-2 (Persian)
	Patient: Knee OA patients above 20years with a diagnosis of Knee pain for at least 6months. Knee OA confirmed via X-ray	Cross sectional	N = 100
			Sex = 80, Male; Female, 20
			Age = M, 53 yrs.
				Validity
				Reliability
					Intervention: None
					Retest: Baseline & 3days

Lovejoy (2012)[25]
USA;
Recruitment site including Hepatology clinics, Veteran Affairs Clinics, Primary care setting, Mental Health Classes;
SF-MPQ-2 (English)
	Patient: Mixed chronic pain population including MSK conditions like neck/joint pain (76%); low back pain (59%); Rheumatism (53%); and fibromyalgia	Cross sectional	N = 214
			Sex = 93%, Male
			Age = M, 54.4 yrs.
				Validity
				Reliability
				Floor or ceiling effect
					Intervention: None
					Retest: Baseline

Maruo (2014)[29] JAPAN; Two pain clinics, University Hospital Ortho & Neuro Surgery Depts.; **SF-MPQ-2** (Japanese version)	Patients: Mixed Japanese speaking chronic pain patients MSK distribution: Knee and hip OA, lumbar and cervical radiculopathy	Cross sectional	N = 96 Sex = Female, 51% Age: =M, 66 yrs.	Reliability Validity	Intervention: None Retest: Baseline & 3month
Packham (2018)[39] NEW ZEALAND; Different community & hospital clinics; **SF-MPQ-2** (English) ***Focus on Neuropathic subscale**	Patients: English speaking adults with type-1 CRPS, as per, IASP classification, affecting any limb; no prior history of CRPS	Longitudinal cohort	N = 59 Sex = Female, 72.9% Age = M, 48.2 (SD 13.3) yrs.	Rasch analysis Validity Reliability Responsiveness	Intervention: Not described Retest: Baseline, 6 and 12 months

Turner (2017)[52]
UNITED KINGDOM;
University clinics, Private settings in Nottinghamshire County;
SF-MPQ-2(English)
	Patients: Patients had primary diagnoses of Knee OA. Patients were excluded if they had surgery 3-months before the study	Rasch analysis	N = 255; Follow Up = 113
			Sex = Male, 42.4%
			Age = M, 68 (SD 9.6) yrs.
				Rasch analysis
				Reliability
				Responsiveness
					Intervention: Knee replacement surgery
					Retest: Baseline and 6months

TABLE 2: QUALITY OF STUDY REPORTS (ARRANGED HIGHEST TO LOWEST)

	ITEM EVALUATION CRITERIA (**SEE KEY BELOW**)												
REFERENCED STUDY	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**Total (%)**
BPI-SF													
Lapane (2014)[24]	2	2	2	0	1	NA	2	2	2	2	2	2	86
Stubbs (2015)[47]	2	2	1	0	2	NA	1	2	2	2	2	2	82
Kapstad (2010)[18]	1	2	1	2	2	NA	2	2	2	2	1	1	82
Keller (2004)[20]	2	2	2	2	1	NA	2	2	2	1	1	2	79
Mease (2011)[31]	2	2	1	0	1	NA	1	2	2	2	2	2	77.
Ferreira-Valente (2012)[12]	2	1	2	2	1	NA	1	1	2	2	1	1	73
Williams (2006)[55]	2	1	1	2	1	1	1	1	2	1	2	2	70
Song (2016)[46]	2	2	1	0	1	NA	2	1	2	1	2	1	68
Celik (2017)[4]	1	2	1	2	1	2	1	1	2	2	0	1	67
Tan (2003)[48]	2	1	1	1	1	NA	1	1	2	1	1	1	59
Group O. (1997)[14]	0	2	1	0	2	NA	1	2	2	2	1	0	59
Whynes (2013)[54]	1	1	1	0	0	NA	2	2	2	2	2	0	59
Mendoza (2006)[33]	2	1	1	2	1	1	0	1	2	1	1	1	58
Krebs (2009 and 2010)[21,22]	0	1	1	1	1	NA	1	1	2	2	1	1	55
Kean (2016)[19]	0	2	1	0	1	NA	1	1	2	2	2	0	55
Risser (2013)[45]	0	1	1	0	1	NA	0	1	2	2	2	0	50
SF-MPQ-2													
Dworkin (2014)[11] & Turk (2015)[51]	2	2	1	2	1	NA	2	2	2	2	2	2	92
Lovejoy (2012)[25]	2	2	1	2	1	NA	2	2	2	2	2	2	91
Packham (2018)[39]	2	2	1	2	0	NA	2	2	2	1	2	2	82
Maruo (2014)[29]	1	2	1	2	1	2	1	1	2	2	2	2	79

Turner (2017)[52]	2	2	1	2	2	0	1	1	2	1	2	1	71
Kachooei (2014)[17]	1	2	1	2	0	NA	2	1	2	2	2	0	68
Adelmanesh (2012)[1]	1	2	1	2	1	0	2	1	2	1	1	1	65
Percentage of studies that received 2 points per criterion	54	67	12	54	16	33	41	41	100	62	58	34	

KEY:

1. Thorough literature review to define research question
2. Description of setting and participants (inclusion/exclusion criteria)
3. Specific Hypothesis
4. Appropriate scope of psychometric properties
5. Sample size
6. Follow-up/retention
7. The authors referenced specific procedures for administration, scoring, and interpretation of procedures
8. Measurement techniques were standardization and significantly void of bias
9. Data were presented for each hypothesis
10. Selection of appropriate statistical test
11. Use of benchmarks and confidence interval
12. Valid conclusion and clinical recommendation

Total score = (sum of subtotals ÷ 24 × 100). If for a specific paper an item is deemed NA (Not Applicable), then, Total score = (sum of subtotals ÷ (2 × number of Applicable items) × 100)

Quality Summary: Poor (0%–30%), Fair (31%–50%), Good (51%–70%), Very good (71%–90%), Excellent (> 90%)

TABLE 3: SUMMARY OF BPI-SF TEST RETEST RELIABILITY, INTERNAL CONSISTENCY AND QUALITY RATING

Psychometric properties	**Authors and Extracted Data**	**Population**	**Risk of Bias**	**Quality rating**
<u>Internal consistency</u>;	<u>95% confidence interval Cronbach alpha (α) coefficients</u>: • Turkish version; Pain severity = 0.84; Pain interference = 0.89 **(Celik et al.)**[4]	Mixed-MSK	Very good	+
	• Pain severity = 0.85; Pain interference = 0.88 **(Tan et al.)**[48]	Mixed-MSK	Very good	+
	• Pain severity and pain interference (Mood and activity) scales = 0.86 - 0.96 **(Mendoza et al.)**[33]	Mixed-MSK	Adequate	+
	• Portuguese version; Pain interference = 0.91 **(Ferreira-Valente et al.)**[12]	Mixed-MSK	Very good	+
	• Pain interference = 0.82 - 0.89 **(William et al.)**[55]	Specific-arthritis	Very good	+
	• Norwegian version; Pain severity = 0.87; Pain interference = 0.88 **(Kapstad et al.)**[18]	Specific-Hip pain	Very good	+
	• Arthritis patients; pain interference = 0.95; pain severity = 0.89 **(Keller et al.)**[20]	Specific- arthritis	Very good	+
	• Lower back pain; pain interference = 0.93; pain severity = 0.82 **(Keller et al.)**[20]	Specific-back pain	Very good	+
	Pooled evidence (range): BPI-SF severity and interference: 0.82-0.96, from 8 studies of very good to adequate quality **GRADE of evidence: High**			
Test- retest	• Sample size (n)= 71; Stable on no treatment; retest interval=7days; ICC = 0.84, Interference; 0.88, severity **(Celik et al.)**[4]	Mixed-MSK	adequate	+
	• Sample size (not specified); Patient status = assumed stable on medication; retest interval= daily, for 7 days: a. Pain severity, ***Pearson correlation(r)***= 0.67-0.88. b. Mood-related Interference ***Pearson correlation (r)*** = 0.68-0.93. c. Activity related interference, ***Pearson correlation (r)*** = 0.70-0.91 **(Mendoza et al.)**[33]	Specific-arthritis	doubtful	?
	• Sample size (n)=43; Assumed stable on high pain medication; 1 week interval (7days-14days): ICC = 0.81, interference **(William et al.)**[55]	Specific-Arthritis	Adequate	+

	a.) Pooled evidence: Severity subscale, n=71, ICC = 0.84, from one study of adequate quality **GRADE of evidence = Low.** **b.) Pooled evidence (Weighted average): Interference subscale n=114, ICC = 0.83, from two studies of adequate quality** **GRADE of evidence = High**
Quality rating Key: + = Sufficient; - = Insufficient; ? = Indeterminate; NA = not applicable **Levels of risk of bias ratings: Very good; Adequate; Doubtful; Inadequate**	

TABLE 4: SUMMARY OF BPI-SF MCID & CID, FLOOR-CEILING EFFECT, VALIDITIES AND QUALITY RATINGS

Psychometric properties	**Authors and Extracted Data**	**Population**	**Risk of bias**	**Quality Rating**
Ceiling and floor effect	• Floor effect – 1-year (After Total hip replacement): 24%, interference scale; 21%, Severity scale **(Kapstad et al.)**[18]	Specific-Hip OA	**NA**	**NA**
Clinically Important Difference (CID)	Against Anchor – PGI-I 1/7 (among fibromyalgia patients) • BPI Severity scale: 2.79 (36.9% improvement) • BPI Average pain scale: 2.82 (43.4% improvement) **(Mease et al.)**[31]	Specific-Fibromyalgia	**NA**	**NA**
Minimal Clinically Important Difference (MCID)	Against Anchor – PGI-I 1/7 (among fibromyalgia patients) • Severity = 2.16 points, (34.2% improvement) • Average pain score = 2.09 points (32.3% improvement) **(Mease et al.)**[31]	Specific-Fibromyalgia	**NA**	**NA**
(i.) Known group validities	***<u>Detected difference between</u>***: **(a)** <u>Interference scale only:</u> • Study 1 Pain-level (arthritis): Low to moderate pain, M, 4.68 (SD, 2.0); High pain, M, 6.6 (SD, 1.9) (t_{130} = -5.66, P <0.0001) **(William et al.)**[55]	Specific-Arthritis	inadequate	?
	• Study 2 Pain-level (Low back pain): Low to moderate pain, M, 5.36 (SD, 1.7); High pain, M, 6.33 (SD, 1.7) (t_{104} = -2.73, P < 0.01) **(William et al.)**[55]	Specific-Arthritis	inadequate	?
	• Differentiate level of disability in Low back pain: stratified by ODI; discriminates between Mild, M, 1.61 (SD, 1.27) and Moderate, M, 3.20 (SD, 1.78); Mild, M, 1.61 (SD, 1.27) and Severe, M, 4.37 (SD, 1.69); Moderate, M, 3.20 (SD, 1.78) and Severe, M, 4.37 (SD, 1.69) (p value 0.01 - 0.001) **(Song et al.)**[46]	Specific-Back pain	inadequate	?
	• Disability (CMP older adults): AUC = 0.663; >4.5 (Fallers from non-faller); AUC, 0.684; Mean, 4.7 (Recurrent fallers from single/non-fallers) **(Stubbs et al.)**[47]	Mixed-MSK	Very good	-
	• Disability (CMP older adults): AUC, 0.724 (95% CI 0.630–0.818); Mean, 4.6 (Recurrent fallers from non-fallers) **(Stubbs et al.)**[47]	Mixed-MSK	Very good	+
	(b) <u>Severity subscale only</u>**:** • Disability (CMP older adults): AUC, 0.665; Mean, >5.1(fallers from non-fallers); AUC, 0.679, Mean, >5.3 (recurrent fallers from single/non-fallers) **(Stubbs et al.)**[47]	Mixed-MSK	Very good	-
	• Disability (CMP older adults): AUC, 0.731 (95% CI 0.635–0.826); Mean, >5.1 (recurrent fallers from non-fallers) **(Stubbs et al.)**[47]	Mixed-MSK	Very good	+
	• Differentiate level of disability in Low back pain: stratified by ODI; discriminates between	Specific-Back pain	inadequate	?

	Mild M,2.47 (SD 1.52) and moderate, M, 3.48 (SD, 1.63); mild, M, 2.47 (SD, 1.52) and severe, M, 3.94 (SD, 1.66); (p value 0.01 - 0.001) **(note:** failed to differentiate moderate and severe ***p-value = 0.089****)* **(Song et al.)**[46]			
	(c) Interference and severity scale**:** • BPI-SF differentiates arthritis patients into varying pain severity as stratified on the Anchor CPG: **Total scale** MANOVA: F (6,182) = 17.58, P < 0.0001 **Severity** ANOVA: F (3,92) = 19.01, *P* < 0.0001 **Interference** ANOVA: F(3,92) = 39.39, *P* < 0.0001 **(Keller et al.)**[20]	Specific-arthritis	inadequate	?
	• BPI-SF differentiates Low back pain patients into varying pain severity as stratified on the Anchor CPG: **Total scale** MANOVA: F (6,204) = 14.66, *P* < 0.0001**;** **BPI Severity** ANOVA: F (3,103) = 12.47, *P* < 0.0001; **BPI Interference** ANOVA: F(3,103) = 33.82, *P* < 0.0001**(Keller et al.)**[20]	Specific-low back pain	inadequate	?
	Pooled evidence: 2 hypotheses confirmed in sufficient studies with "very good" quality rating. **GRADE of evidence: Moderate**			
(ii) Convergent/ Criterion Validity	Change in Predicted Direction: Change in BPI scores baseline to 7days, in concordance, with Patient global assessment of arthritis rating of change from baseline to 14 days) **(Mendoza et al.)**[33] Severity subscale: -2.35, Improved; -1.08, No change (CI 3.26 P < 0.001) Activity-interference subscale: 0.76, Improved; -2.32, No change (CI 2.56 P < 0.001) Mood-interference subscale: -1.19, improved; -2.21, No change (CI 3.95 P < 0.011)	Mixed-MSK	Doubtful	-
	Correlation with other scales reported as moderate (0.30-0.69): Severity subscale only: • VAS[T]& SF-BPI-Worst pain **(Celik et al.)**[4]	Mixed-MSK	Adequate	-
	• SF-36 (Physical function, mental, Role [emotional & physical], social function, vitality, general health), CPG-disability, HAD-disability, RMDQ-disability **(Keller et al.)**[20]	Mixed-MSK	Very good	+
	• RMDQ-disability; CPG disability **(Kreb et al. 2009)**[22]	Mixed-MSK	Very good	+
	Interference subscale only: • SF-12 Mental health and physical function, NRS[P], HAD[P]-Activity & Disability **(Ferreira-Valente et al.)**[12]	Mixed-MSK	Very good	+
	• WOMAC, (pain, stiffness, physical function, pain at night in bed), ALQ, sleep quality, number of night awakening **(William et al.)**[55]	Specific-Arthritis	Very good	+

	• WOMAC[n] (stiffness); SF-36 (role physical, role emotional, and general health) **(Kapstad et al.)**[18]	Specific-Hip OA	Very good	+
	• CPG intensity **(Krebs et al. 2009)**[22]	Mixed-MSK	Very good	+
	Both Interference and severity scale: • ODI[c] **(Song et al.)**[46]	Specific-Back pain	Very good	+
	• VAS (pain), WOMAC (pain, physical function and stiffness) **(Mendoza et al.)**[34]	Specific-Arthritis	Very good	+
	• FMI, SF-36 body pain **(Krebs et al. 2009)**[22]	Mixed-MSK	Very good	+
	• WOMAC[n] [Pain, Physical function); SF-36 (physical function, bodily pain, vitality, social function and mental function) **(Kapstad et al.)**[18]	Specific-Hip OA	Very good	+
	• RMDQ-disability **(Tan et al.)**[48]	Mixed-MSK	Very good	+
	Correlation with other scales reported as High (≥0.70): Interference subscale only**:** • SF-36 (Vitality, mental health, physical & social function), CGP-disability; RMDQ-disability **(Keller et al.)**[20]	Mixed-MSK	Very good	+
	• CPG disability, RMQD-disability **(Krebs et al. 2009**)[22]	Mixed-MSK	Very good	+
	Severity subscale only: • CPG intensity **(Krebs et al. 2009)**[22]	Mixed-MSK	Very good	+
	Interference and severity subscale**:** • SF-36-body pain; CPG-Intensity**(Keller et al.)**[20]	Mixed-MSK	Very good	+
	• PEG, Overall pain stress **(Krebs et al. 2009)**[22]	Mixed-MSK	Very good	+
	Pool evidence: 15 PROs examined: 94% hypothesis confirmed; 24 (64%) hypothesis @ rho, 0.3-0.69;10 hypothesis @ rho ≥0.7 **Grade of evidence: high**			
Structural validity; Factor Analysis	**Support 1 factor:** • Confirmatory Factor Analysis; supports 1 factor structure (assessed only the Portuguese BPI-SF interference); χ2 (14) =72.54, (p < 0.001), CFI = 0.91; SRMR = 0.06; CFI = 0.91; SRMR 0.06 **(Ferreira-Valente et al.)**[12]	Mixed-MSK	Very good	+

	Support 2 factor: - Confirmatory factor analysis with principle promax rotation; 56% variance accounted; 2 factor structure (interference items & pain intensity) **(Celik et al.)**[4]	Mixed-MSK	Very good	?
	- Exploratory Factor analysis; principle promax rotation; 2 factor structure with 67% variance accounted; eigenvalues > 1 (6.9 and 1.2);interitem correlation 0.59-0.88 **(Keller et al.)**[20]	Mixed-MSK	Very good	?
	- Compared a one, two & three factor model; 2 factor model (pain and interference) yielded best fit with CFI 0.99, CI 0.04-0.07. RMSEA for 2 factor model was 0.05 compares to 1 factor model (0.17) and 3 factor model (0.12) **(Lapane et al.)**[24]	Mixed-MSK	Very good	+
	- Supports 2 factor structure; confirmatory factor analysis; yielded 2 factor models (pain and interference) with RMSEA 0.09, GFI 0.91, CFI 0.92; however, NFI - 0.89 and AGFI - 0.87 (> 0.90 required) and SRMR 0.39 **(Song et al.)**[46]	Specific-Back pain	Very good	+
	- Confirmatory factor analysis; promax rotation; support two factor (interference and pain) accounting for 63.6% variance; eigenvalue 5.62 & 1.38 for interference and pain intensity respectively **(Tan et al.)**[48]	Mixed-MSK	Very good	?
	Support 3 factor: - Modified version of SF-BPI; confirmatory factor analysis; Oblique rotation; 3 factor structures (pain, mood-interference and activity-interference); 86% variance accounted; eigenvalue range 0.9 - 5.1; in both studies, items "sleep" and "enjoyment of life" did not load properly and were dropped **(Mendoza et al.)**[33]	Specific-Arthritis	Very good	?
	Pooled evidence: "severity" and "interference" factor structures explained in 3 sufficient studies with "very good" quality **GRADE of evidence: High**			

Key: ODI, *Oswestry Disability Index;* **PEG;** *3-item SF-BPI ("pain average," "interference with enjoyment of life," and "interference with general activity");* **CPG;** Chronic Pain Grade; **SF-36,***Short Form-36 Health Status Questionnaire***; SF-12,** *Short Form-12 Health Status Questionnaire*; **RMDQ**, *Roland Morris Disability Questionnaire;* **FMI**, *Functional Morbidity Index;* **WOMAC**, *Western Ontario and McMaster Universities Osteoarthritis Index*; **HAD;** *Hospital Anxiety and* Depression Scale; **ALQ,** *Activities and Lifestyle Questionnaire;* **VAS,** *Visual Analogy Scale;* **NRS,** *Numerical Rating scale*, **CMP**, *Chronic Musculoskeletal Pain;* ***t,*** *Turkish;* ***c***, *Chinese;* **pg**, *Portuguese version;* ***n***, *Norwegian;* **AGFI,** *adjusted goodness-of-fit index;* ***χ2,*** *chi-squared*; **CFI,** *comparative fit index;* **df,** *degree of freedom;* **GFI,** *goodness-of-fit index;* **NFI,** *normed fit index;* **RMSEA,** *root mean squared error of approximation;* **SRMR,** *standardized root mean square residual;* **CI**, *Confidence interval;* **AUC**, *Area Under the ROC Curve***; PGI-I**, *Patient global impression of improvement*

Quality rating Key: + = sufficient; - = insufficient; ? = indeterminate; NA = not applicable

Levels of Risk of Bias: Very good; Adequate; Doubtful; Inadequate

TABLE 5: SUMMARY OF BPI-SF RESPONSIVENESS AND QUALITY RATINGS

Psychometric properties	**Authors and Extracted Data**	**Population**	**Risk of bias**	**Quality rating**
Correlation of change scores	Cross sectional change correlation **(Group O.)**[14] OQLD, physical= 0.72 and Symptom= 0.81 (baseline to 2 weeks) Longitudinal change correlation BPI-SF Severity: OQLD, physical= 0.38; Symptom= 0.48 (2 weeks to 12 months)	Specific- Osteoporosis	Very good	+
	Cross sectional change correlation**(Whynes et al.)**[54] BPI Severity: 0.52, 0DI; -0.59, EQ-5D-Index; -0.48, ED-5D-VAS BPI interference: 0.61, 0DI; -0.63, EQ-5D-Index; -0.40, ED-5D-VAS Similar pattern across other scales Weekly change correlation BPI Severity: -0.57, EQ-5D-Index; -0.56, ED-5D-VAS; 0.70, ODI BPI Interference: -0.58, EQ-5D-Index; -0.50, ED-5D-VAS; 0.65, ODI Similar pattern across other scales	Specific-Back pain	Very good	+
	Sensitivity to change in patient status **(Krebs et al. 2010)**[21] AUC (Any improvement): Severity = 0.81-0.83; Interference = 0.70 - 0.78 AUC (Moderate Improvement): Severity = 0.81-0.85; Interference = 0.67-0.77 Better or similar to PEG, CPG, RMDQ, and SF-36-body pain	Mixed-MSK	Very good	+
	AUC (interference) = 0.80 **(Whynes et al.)**[54]	Specific-Back pain	Very good	+
	Sensitivity to change in patient status **(Kean et al.)**[19] • AUC (Severity): 0.727, any improvement; 0.737, moderate improvement • AUC(total): 0.727, any improvement; 0.743, moderate improvement Better than PEG, SF-36 bodily pain, and PROMIS PI-6b SF; PROMIS P-SF (27 and 57)	Mixed-MSK	Very good	+
	• AUC (interference): 0.677, any improvement; 0.694, Moderate improvement. **(Kean et al.)**[19]	Mixed MSK	Very good	-
	Using MANOVA statistic, BPI significantly identified change in patients' conditions (with 2-SEM signalling change on BPI) among: (a.) Arthritis (against HAD): [Wilks' Lambda F(4,194) = 4.84, P < 0.001] (b.) Low back pain (against RMDQ): [Wilks' Lambda F (4,198) = 10.77, P < 0.0001] **(Keller et al.)**[20]	Specific- Arthritis Specific-Back pain	Inadequate Inadequate	- -

<table>
<tr><td></td><td>BPI responsiveness to detect interventional change across visits; Interval (27.73 days) <b>(Tan et al.)</b>[48]
<table>
<tr><th></th><th>Mean Change</th><th></th><th></th><th></th></tr>
<tr><th></th><th>1st visit</th><th>2nd visit</th><th>3rd visit</th><th>T-test/CI</th></tr>
<tr><td>BPI interference</td><td>7.42</td><td>6.71</td><td>6.46</td><td>2.52-5.33
P > 0.01-0.001</td></tr>
<tr><td>BPI Severity</td><td>7.07</td><td>6.63</td><td>6.14</td><td>1.12-4.66
P > 0.01-0.001</td></tr>
</table>
</td><td>Mixed-MSK</td><td>Inadequate</td><td>-</td></tr>
<tr><td><b><u>ES and SRM</u></b></td><td>Improvement rating similar with RMDQ and PEG; higher than generic outcomes like CPG and SF 36-body pain.
SRM: better, -1.02; Worst, 0.37; Same, -0.18 <b>(Krebs et al. 2010)</b>[21]</td><td>Mixed-MSK</td><td>Adequate</td><td><b>?</b></td></tr>
<tr><td></td><td>Sensitivity to change; anchor (patient reported global change- item 1/7) 3-months; compared to SF-36 body pain, PEG and PROMIS PI-6b SF; PROMIS P-SF (27 and 57) <b>(Kean et al.)</b>[19]
• SRM (Severity): Better, 0.71; Worst, -0.47; Same, 0.13
• SRM (Interference): Better, 0.94; Worst, 0.03; Same, 0.38
• SRM (Total): Better, 0.94; Worst, -0.22; Same, 0.31</td><td>Mixed-MSK</td><td>Adequate</td><td><b>?</b></td></tr>
<tr><td></td><td>SRM (interference and severity) = 0.90; higher than ODI (0.82) and EQ-5D (0.63-0.83) <b>(Whynes et al.)</b>[54]</td><td>Specific-Back pain</td><td>Adequate</td><td><b>?</b></td></tr>
<tr><td></td><td>ES: Total RCT study = 0.64;
Back pain population = 0.60;
Hip/knee OA = 0.69 <b>(Keller et al.)</b>[20]</td><td>Mixed-MSK
Specific-Back pain
Specific- Arthritis</td><td>Adequate
Adequate
Adequate</td><td><b>?</b>
<b>?</b>
<b>?</b></td></tr>
<tr><td></td><td>ES: 0.53, BPI average pain item; Anchor, PRPI (0/10 scale); BPI average pain item similar to ICOAP and (Anchor) PRPI but higher than WOMAC pain and ICOAP <b>(Risser et al.)</b>[45]</td><td>Specific-Knee pain</td><td>Adequate</td><td><b>?</b></td></tr>
<tr><td></td><td>12 months Responsiveness to intervention improvement in 2 different conditions <b>(Keller et al.)</b>[20]
<table>
<tr><th></th><th></th><th>SRM</th><th></th></tr>
<tr><th>Categories</th><th>Improved</th><th>Same</th><th>Declined</th></tr>
<tr><td><b><u>Arthritis</u></b></td><td></td><td></td><td></td></tr>
<tr><td>BPI Intensity</td><td>-0.87</td><td>-0.55</td><td>0.01</td></tr>
<tr><td>BPI Interference</td><td>-0.84</td><td>-0.33</td><td>0.16</td></tr>
<tr><td><b><u>Low back pain</u></b></td><td></td><td></td><td></td></tr>
<tr><td>BPI Intensity</td><td>-1.09</td><td>-0.40</td><td>0.28</td></tr>
</table>
</td><td>Specific- Arthritis</td><td>Adequate</td><td><b>?</b></td></tr>
</table>

<table>
<tr><td></td><td><table><tr><td>BPI Interference</td><td>-1.13</td><td>-0.56</td><td>0.43</td></tr></table></td><td>Specific- back pain</td><td>Adequate</td><td><b>?</b></td></tr>
<tr><td></td><td>

- Magnitude of overall condition change (ES, SRM and RI), 12 months post THR **(Kapstad et al.)**[18]

<table><tr><th>Indicators</th><th>BPI interference</th><th>BPI pain intensity</th></tr><tr><td>ES</td><td>1.71</td><td>1.57</td></tr><tr><td>SRM</td><td>1.52</td><td>1.61</td></tr><tr><td>RI</td><td>2.05</td><td>2.03</td></tr></table></td><td>Specific-Hip pain</td><td>Adequate</td><td><b>?</b></td></tr>
<tr><td></td><td>

- Magnitude of change on BPI scales; Anchor SF-36 perceived global change (1-7 scale); similarly responsive as WOMAC (physical and stiffness), and more sensitive than other SF-36 subscales in THR patient) **(Kapstad et al.)**[18]

<table><tr><th>Scale</th><th colspan="3"><b>Improved group</b></th><th colspan="3"><b>Unchanged group</b></th></tr><tr><td></td><td>ES</td><td>SRM</td><td>RI</td><td>ES</td><td>SRM</td><td>RI</td></tr><tr><td>BPI Intensity</td><td>1.70</td><td>1.71</td><td>2.17</td><td>1.00</td><td>1.36</td><td>1.39</td></tr><tr><td>BPI interference</td><td>1.80</td><td>1.81</td><td>2.16</td><td>1.27</td><td>1.40</td><td>1.56</td></tr></table></td><td>Specific-Hip pain</td><td>Adequate</td><td><b>?</b></td></tr>
<tr><td></td><td>

- Using Guyatt statistic, magnitude of change (BPI-Interference scale) between Control and Placebo **(Williams et al.)**[55]

<table><tr><th>Intervals</th><th>Study 1: ES</th><th>Study 2: ES</th></tr><tr><td>0-7days</td><td>0.46</td><td>-</td></tr><tr><td>0-14days</td><td>1.06</td><td>-</td></tr><tr><td>0-30days</td><td>-</td><td>0.84</td></tr><tr><td>0-90days</td><td>-</td><td>0.95</td></tr></table></td><td>Specific-Arthritis</td><td>inadequate</td><td>-</td></tr>
<tr><td></td><td colspan="4"><b>Pooled evidence: 5 reports with “very good” quality rating available on responsiveness of the BPI-SF
GRADE of evidence: High</b></td></tr>
<tr><td colspan="5"><b><i>Key:</i> AUC</b>, <i>Area Under the ROC Curve</i><b>; CGP,</b> <i>Chronic Pain Grade</i><b>; WOMAC,</b> <i>Western Ontario and McMasters Universities osteoarthritis index</i><b>; ICOAP,</b> <i>Intermittent and Constant Osteoarthritis Pain</i>; <b>PRPI,</b> <i>patient rated pain index; ;</i> <b>PEG,</b> <i>3-item SF-BPI (“pain average,” “interference with enjoyment of life,” and “interference with general activity”);</i> <b>RMDQ</b>, <i>Roland Morris Disability Questionnaire</i>; <b>EQ-5D,</b> <i>European Quality of Life Instrument (Version 5D)</i><b>; PROMIS PI-6b-SF</b>, <i>short-form</i> Patient-Reported Outcomes Measurement Information System-6b <i>(Interference scale);</i> <b>ODI,</b> <i>Oswestry Disability Index;</i> <b>PROMIS P-SF (27 and 57</b><i>),</i> <i>short-form-Patient-Reported Outcomes Measurement Information System Profile (29 and 57 item versions);</i><b>SF-36,</b><i>Short Form-36 Health Status Questionnaire</i>; <b>OQLD,</b> <i>Osteoporosis Quality of Life Questionnaire</i>; <b>ES</b>, <i>Effect size</i>; <b>SRM</b>, <i>Standardized Response Mean</i>; <b>RI</b>, <i>Responsiveness Index</i>
<b>Quality rating Key: + = sufficient; - = insufficient; ? = indeterminate; NA = not applicable</b>
<b>Levels of Risk of Bias: Very good; Adequate; Doubtful; Inadequate</b></td></tr>
</table>

TABLE 6. SUMMARY OF SF-MPQ-2 MEASUREMENT PROPERTIES AND THEIR QUALITY RATINGS IN MSK				
Psychometric Properties	**Extracted Data and First Author**	**MSK-Category**	**Risk of bias**	**Quality Rating**
Floor and ceiling effect:	• Floor effect: Affective Scale, 15.1%; Neuropathic Scale, 12.5% **(Dworkin et al.)**[11]	Specific-Back pain	**NA**	**NA**
	• Floor effect: Affective, 28.5%; Neuropathic , 12.4%; Intermittent, 15.1% **(Lovejoy et al.)**[25]	Mixed-MSK	**NA**	**NA**
	• Floor effect: Continuous, 4.6%; Intermittent, 4.2%; Neuropathic, 1.9%; Affective, 8.7; Total, 3.1% **(Adelmanesh et al.)**[1]	Mixed-MSK	**NA**	**NA**
Hypothesis testing: (i). Criterion/Convergent Validity	**Correlation with other measures reported as moderate (Pearson or Spearman rho=0.3-0.69):** *Continuous scale:* • WOMAC^{P}; SF 36^{p} (PH,BP,GH,RE,MH,MCS &PCS) **(Kachooei et al.)**[17]	Specific-Knee pain	Very good	+
	• BPI severity (worst, least, average and current); NRS-average back pain, NRS-average leg pain; NRS-current back pain; NRS-current leg pain; HAD-total; HAD-anxiety **(Dworkin, 2014 and Turk, 2015)**[11,51]	Specific-Back pain	Very good	+
	• MPI-interference; PDI; GAD-7; BDI-II **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+
	• VAS^{J}; SF-MPQ^{J} (sensory, affective, total); LF-MPQ^{J} (sensory, affective, evaluative, total) **(Maruo et al.)**[29]	Mixed-MSK	Very good	+
	Intermittent scale: • WOMAC^{P}; SF 36^{p} (PF, BP, VT, GH, RE,MCS) **(Kachooei et al.)**[17]	Specific-Knee pain	Very good	+
	• BPI severity (worst, least, average and current); NRS-average back pain, NRS-average leg pain; NRS-current back pain; NRS-current leg pain **(Dworkin, 2014 and Turk, 2015)**[11,51]	Specific-Back pain	Very good	+
	• MPI-interference; MPI-Severity; PDI; GAD-7; BDI-II **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+
	• VAS^{J}; SF-MPQ^{J} (sensory, affective, total); LF-MPQ^{J} (sensory, affective, evaluative, total) **(Maruo et al.)**[29]	Mixed-MSK	Very good	+
	Neuropathic scale: • WOMAC^{P}; SF 36^{p} (PF, BP, PCS) **(Kachooei et al.)**[17]	Specific-Knee pain	Very good	+
	• BPI severity (least, average and current); NRS-average leg pain; NRS-current leg pain; HAD-total; HAD-anxiety **(Dworkin, 2014 and Turk, 2015)**[11,51]	Specific-Back pain	Very good	+
	• MPI-interference; MPI-Severity; PDI; GAD-7; BDI-II **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+

	• VAS^{J}; SF-MPQ^{J} (sensory, affective, total); LF-MPQ^{J} (sensory, affective, evaluative, total) **(Maruo et al.)**[29]	Mixed-MSK	Very good	+
	• CSS (**Packham et al.**)[39]	Specific-CRPS	Very good	+
	Affective scale: • WOMAC^{P}; SF 36^{p} (PF, RE, BP, VT, MH,MCS) **(Kachooei et al.)**[17]	Specific-Knee pain	Very good	+
	• BPI severity (least and average); NRS-average leg pain; NRS-current leg pain, HAD-Total, HAD-Anxiety **(Dworkin, 2014 and Turk, 2015)**[11,51]	Specific-Back pain	Very good	+
	• MPI-interference; MPI-Severity; PDI; GAD-7; BDI-II **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+
	• SF-MPQ^{J} (sensory, total); LF-MPQ^{J} (sensory, affective, evaluative, total) **(Maruo et al.)**[29]	Mixed-MSK	Very good	+
	Total scale: • WOMAC^{P}; SF 36^{p} (PF,BP, GH, VT, RE,MH, PCS, MCS) **(Kachooei et al.)**[17]	Specific-Knee pain	Very good	+
	• BPI severity (worst, least, average and current); NRS-average back pain, NRS-average leg pain; NRS-current back pain; NRS-current leg pain; HAD-total; HAD-anxiety **(Dworkin, 2014 and Turk, 2015)**[11,51]	Specific-Back pain	Very good	+
	• MPI-interference; PDI; GAD-7; BDI-II **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+
	• VAS^{J}; SF-MPQ^{J} (sensory, affective, total); LF-MPQ^{J} (affective, evaluative) (**Maruo et al.)**[29]	Mixed-MSK	Very good	+
	• CSS (**Packham et al.**)[39]	Specific-CRPS	Very good	+
	Correlation with other tools reported as High (Pearson or Spearman rho= ≥ 0.7): *Continuous scale:* • MPI-severity **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+
	Affective scale: • SF-MPQ^{J}-affective (**Maruo et al.**)[29]	Mixed-MSK	Very good	+
	Neuropathic scale: • NRS-pain, PDI **(Packham et al)**[39]	Specific-CRPS	Very good	+
	Total scale: • MPI-severity **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+
	• NRS-pain, PDI **(Packham et al)**[39]	Specific-CRPS	Very good	+
	• VAS^{P}**(Adelmanesh et al.)**[1]	Mixed-MSK	Very good	+
	• LF-MPQ^{J}-total, LF-MPQ^{J}- sensory) **(Maruo et al.)**[29]	Mixed-MSK	Very good	+

	Pooled evidence: 14 PRO Comparators; 75 Hypothesis, rho = 0.3-0.69; 6 Hypothesis, rho = ≥ 0.7 **GRADE of evidence: High**			
(ii.) Known group validity	**SF-MPQ-2 Total OR subscales:** • Discriminant patient stratified by QTFC scale: 3, 4, & 6 **Only total scale score extracted** QTFC category 3: total scale- M, 3.97 (SD, 2.03) QTFC category 4 or 6: total scale- M, 4.49 (SD, 2.04) p-value = 0.001 **(Dworkin et al.)**[11]	Specific-back pain	inadequate	?
	• Differentiate patients by number of reported pain sites, were those with higher SF-MPQ-2 pain scores indicated more pain sites. **Only total scale score extracted**: One pain site: M, 2.44 (SD 2.14); Two-three pain site: M, 2.97 (SD 2.13) Four or more pain site: M, 3.81 (SD 2.36) p-value = 0.05 **(Lovejoy et al.)**[25]	Mixed-MSK	inadequate	?
	• Discriminate patients stratified on MPI scale into: **Only total scale score extracted**: None/mild: M, 1.16 (SD, 1.61) Moderate: M, 3.08 (SD 1.68) Severe: M, 5.55 (SD 2.00) p-value = 0.05 **(Lovejoy et al.)**[25]	Mixed-MSK	inadequate	?
	• Differentiate patients stratified on PPI scale into: Mild pain: M, 33.81, (SD 14.16) p-value = 0.041 Discomforting: 45.60 (SD 16.00) p-value = 0.028 Distressing: M 53.62 (SD 18.78) p-value = 0.032 Horrible: M, 58.49 (SD 18.97) p-value = 0.027 **(Adelmanesh et al.)**[1]	Mixed-MSK	inadequate	?
	Pooled evidence: 4 hypotheses tested in studies with "inadequate" and "insufficient" quality rating. **GRADE of evidence: Very low**			
Structural validity; Rasch analysis	**SF-MPQ-2 Total OR subscales:** • SF-MPQ-2 Continuous scale: Item 8 and 9 misfit; item 10 displays uniform DIF for gender; passed unidimensionality test; removal of item 9 returns stability across structures; differential item functioning present among gender group • SF-MPQ-2 Intermittent scale: Passed unidimensionality test; items 2 and 3 misfit; No DIF for gender • SF MPQ-2 Neuropathic Scale: No item misfit; No dependency of item; scale passed unidimensionality test • SF-MPQ-2 Affective scale: Item 15 misfit; passed unidimensionality test; items had disordered response threshold that was not resolved; did not fit Rasch model	Specific-Knee pain	Very good Very good Very good Very good Very good	- + + + -

	• SF-MPQ-2 Total scale score: complete misfit with Rasch Model; several item exhibit dependence and don't exhibit differentially item function **(Turner et al.)**[52]			
	SF-MPQ-2 Neuropathic subscales: • Disorder 'Tingling' threshold corrected after collapsing to a 6-interval scale, from an 11-interval scale; passed unidimensionality test. • No item misfit; level of difficulty adequately distributed; acceptable Person fit statistics observed (x [SD] ¼ –1.17 [1.13] logits) • Although corrected, local dependence exhibited between "Burning" and "numbness" items; DIF observed on "Pain with light touch" item as severity level varies on the CSS scale. • Although corrected, person separation index below individual level of discrimination (0.78, against required 0.85) **(Packham et al. 2018.)**[39]	Specific-CRPS	Doubtful	-
Factor analysis	Confirmatory Factor Analysis; hypothesized a four-factor Solution, 3 confirmed **(Dworkin et al.)**[11] • Continuous Scale; GFI, 0.988; RMSEA, 0.054; SRMR, 0.0268 • Intermittent Scale; GFI, 0.957; RMSEA, 0.111; SRMR, 0.0459 • Neuropathic Scale; GFI, 0.889; RMSEA, 0.191; SRMR, 0.0740 • Affective Scale; GFI, 0.983; RMSEA, 0.129; SRMR, 0.0250	Specific-Back pain	Very good Very good Very good Very good	+ + - +
	• Confirmatory Factor Analysis; Compared a 1-factor, with a 4-factor solution; 4-factor solution demonstrates better fit; inter-item correlation 0.61-0.88 **(Lovejoy et al.)**[25] 1-Factor Solution: TLI = 0.82, CFI = 0.84, SRMR = 0.06, RMSEA = 0.09, AIC = 19129 (poor fit) 4-Factor Solution: TLI = 0.88, CFI =0 .89, SRMR = 0.06, RMSEA = 0.08, AIC = 18983 (best fit)	Mixed-MSK	Very good	+
	• Exploratory factor analysis; minimum loading factor 0.4; variance accounted 57.49% ; 4 factor solution supported; Heavy-pain item over load **(Adelmanesh et al.)**[1]	Mixed-MSK	Adequate	?
	• Confirmatory factor analysis; 96 Subjects; Support a 4-factor solution; chi-squared= 478, degrees of freedom = 203; GFI = 0.917; AGFI = 0.894; RMSEA= 0.05) **(Maruo et al.)**[29]	Mixed-MSK	Inadequate	+
	Pooled result: 3 "very good" studies with conflicting evidence (-/+) GRADE of evidence: High.			
Reliability Internal consistency	Cronbach Alpha: • Total scale^p: T1= 0.88 • Subscales^p (Continuous, Intermittent, Affective, Neuropathic): T1 = 0.75 - 0.81 **(Kachooei et al.)**[17]	Specific-Knee pain	Very good	+
	• Total scale: 0.93 • Subscales (Continuous, Intermittent, Affective, Neuropathic): 0.77 - 0.84 **(Dworkin et al.)**[11]	Specific-Back pain	Very good	+
	• Total scale: 0.96 • Subscales (Continuous, Intermittent, Affective, Neuropathic): 0.84 - 0.92 **(Lovejoy et al.)**[25]	Mixed-MSK	Very good	+

	• Total score^{J}: 0.907 • Subscales^{J} (Continuous, Intermittent, Affective Neuropathic): 0.857 -0.917) **(Maruo et al.)**[29]	Mixed-MSK	Very good	+
	• Total score: 0.95 • Subscales (Neuropathic): 0.83 **(Packham et al.)**[39]	Mixed-MSK	Very good	+
	• Total scale^{p}: 0.906 **(Adelmanesh et al.)**[1]	Mixed-MSK	inadequate	-
	Pooled evidence (range): SF-MPQ-2, range for total subscale= 0.88-0.96; range for Subscale score = 0.75-0.92, from 4 studies of "very good" quality **GRADE of evidence: High**			
<u>Reproducibility;</u> Test- retest	<u>Intraclass correlation coefficients (ICC)</u>: • Analyzed with Rasch Model: • Normal ICC, for each subscale 0.38 - 0.67 • Rasch Converted ICC, for each subscale 0.47- 0.63 ***(Affective subscale excluded)*** **(Turner et al.)**[52]	Specific-Knee pain	Doubtful	-
	• N = 43; Retest-interval = 3 days, • Total scale^{P} = 0.90; • Subscales^{P} (Continuous, Intermittent, Affective, Neuropathic): 0.73-0.90 **(Kachooei et al.)**[17]	Specific-Knee pain	Adequate	+
	• Total scale^{P} ICC = 0.941 **(Adelmanesh et al.)**[1]	Mixed-MSK	Doubtful	+
	• Total scale^{J}= 0.83 • Subscales^{J} (Continuous, Intermittent, Neuropathic, Affective): 0.75-0.85 **(Maruo et al.)**[29]	Mixed-MSK	Doubtful	+
	Pooled evidence: ICC range for Total score, 0.90; ICC range for subscales, 0.73-0.90, from one study of "adequate" quality **GRADE of evidence: Low**			
<u>Responsiveness;</u>	• Post knee replacement patient; 6months interval. Effect Size: Continuous scale, 1.08; Intermittent, 1.12; Neuropathic, 0.15; Affective, 0.78 Rasch converted Effect size: Continuous, 1.27; intermittent, 1.02; neuropathic, 0.09 **(Turner et al.)**[52]	Specific-Knee pain	Inadequate	-
	• Neuropathic subscale: Effect Size = 0.92 (CI, 0.53 -1.31); SRM = 0.97 (**Packham et al.)**[39]	Specific-CRPS	Adequate	?

<table>
<tr>
<td></td>
<td>

- Change in patient status between baseline and day 5, after commencement of treatment was significant for all subscales of the SF-MPQ-2 and its total score **(Dworkin et al.)**[11]

<table>
<tr><th>Subscale</th><th>Baseline
M(SD)</th><th>Day-5
M(SD)</th><th>T-Test value</th></tr>
<tr><td><b>continuous</b></td><td>5.19 (2.19)</td><td>2.79 (2.13</td><td>t(527) = 26.36, P < .01</td></tr>
<tr><td><b>Intermittent</b></td><td>5.04 (2.34)</td><td>2.45 (2.15</td><td>t(527) = 27.75, P < .01</td></tr>
<tr><td><b>affective</b></td><td>2.94 (2.26)</td><td>1.47 (1.69)</td><td>t(527) = 17.73, P < .01</td></tr>
<tr><td><b>Total</b></td><td>4.23 (2.05)</td><td>2.11 (1.77)</td><td>t(527) = 27.31, P < .01</td></tr>
</table>

</td>
<td>Specific-Back pain</td>
<td>Inadequate</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>

- Non-neuropathic patients mean difference of pre-treatment and Post-Treatment as anchored with PGIC **(Adelmanesh et al.)**[1]

<table>
<tr><th>Subscale</th><th>M(SD)</th><th>P-Value</th></tr>
<tr><td>Very much improved</td><td>35.20 (SD 11.43)</td><td>0.007</td></tr>
<tr><td>Much Improved</td><td>28.22 (SD 8.62)</td><td>0.0014</td></tr>
<tr><td>Minimally improved</td><td>21.43 (SD 5.40)</td><td>0.016</td></tr>
<tr><td>No change</td><td>9.33 (SD 6.08)</td><td>p-value not reported</td></tr>
</table>

</td>
<td>Mixed-MSK</td>
<td>Inadequate</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td colspan="4"><b>Pooled evidence: `1 adequate study of indeterminant quality
GRADE of evidence: Very low</b></td>
</tr>
<tr>
<td colspan="5"><b>Key</b>: <b><i>WOMAC,</i></b> <i>Western Ontario and McMaster Universities Arthritis Index</i>; <b>SF-36,</b><i>Short Form-36 Health Status Questionnaire</i> (<b>PF</b>=<i>Physical Function</i>, <b>BP</b>=<i>Body Pain</i>, <b>GH</b>=<i>General Health</i>, <b>VT</b>=<i>Vitality</i>, <b>RE=</b><i>Role Emotion</i>, <b>MCS=</b><i>Mental Component Summary</i>, <b>PCS=</b><i>Physical Component Summary</i>); <b>QTFC</b>, <i>Quebec Task Force Classification for Spinal Disorder</i>; <b>MPI,</b> <i>Multidimensional Pain Inventory</i>; <b>CI</b>, <i>confidence interval</i>; <b>CSS,</b> <i>Complex Regional Pain Syndrome Severity Score</i>; <b>PDI,</b> <i>Pain Disability Index</i>; <b>BDI-II,</b> <i>Beck Depression Index-Version 2</i>; <b>GAD-7,</b> <i>General Anxiety Disorder 7-item scale</i>; <b>SF-MPQ^{J}</b>, <i>Short form McGill Pain Questionnaire-Japanese Version</i>; <b>LF-MPQ^{J}</b>, <i>Long Form McGill Pain Questionnaire-Japanese version;</i> <b>PPI</b>, <i>Present Pain Intensity</i>; <b>AIC</b>, <i>Akaike Information Criterion;</i> <b>VAS,</b> <i>Visual Analogy Scale</i>; <b>HAD;</b> <i>Hospital Anxiety and</i> Depression Scale; <b>NRS,</b> <i>Numerical Rating scale</i><b>; PGIC</b>, <i>patient global impression of change;</i> <b>AGFI,</b> <i>Adjusted Goodness-of-Fit Index</i>; <b>CFI,</b> <i>comparative fit index;</i> <b>GFI,</b> <i>goodness-of-fit index;</i> <b>RMSEA,</b> <i>root mean squared error of approximation;</i> <b>SRMR,</b> <i>standardized root mean square residual;</i> <b>M</b>, <i>Mean;</i> <b>SD</b>, <i>standard deviation</i> <b>CFI</b> <i>comparative fit index;</i> <b>TLI;</b> <i>Tucker-Lewis index</i>; <b>P</b>, <i>Persian version</i><b>; J,</b> <i>Japanese version</i>

<b>Quality rating Key: (+) = sufficient; (-) = insufficient; (?) = indeterminate; NA = not applicable
Risk of Bias rating: Very good; Adequate; Doubtful; Inadequate</b></td>
</tr>
</table>

TABLE 7: COSMIN MODIFIED GRADE LEVEL OF EVIDENCE SYNTHESIS

Measurement Property	**Brief Pain Inventory-Short form**		**Revised short McGill Pain Questionnaire Version 2**	**Is one instrument better?**
	Interference	**Severity**	**Subscales and Total scores**	
1. Test-retest reliability	High (+)	Low (+)	Low (+)	Yes, BPI-SF
3. Internal consistency	High (+)	High (+)	High (+)	No
4. Responsiveness	High (+)	High (+)	Very low (?)	Yes, BPI-SF
5. Structural validity	High (+)	High (+)	High (+/-)	Yes, BPI-SF
6. Hypothesis testing (convergent validity)	High (+)	High (+)	High (+)	No
7. Hypothesis testing (known group validity)	Moderate (+)	Moderate (+)	Very low (?)	Yes, BPI-SF
8. Cross cultural validity/Measurement invariance	No evidence	No evidence	No evidence	No evidence
9. Measurement error (SEM and MDC)	No evidence	No evidence	No evidence	No evidence

APPENDIX 1

Search concepts adapted, each for the BPI-SF and SF-MPQ-2, on Medline, EMBASE, CINAHL and Scopus bibliographic databases

A. **(“Brief Pain Inventory”) AND (Psychometric OR “Measurement Properties” OR Validation OR Adaptation OR "Cross-cultural" OR Reliability OR Validity OR "Internal Consistency" OR Sensitivity OR Specificity OR Discriminative OR Responsiveness OR "Factor analysis" OR Minimal Clinically Important Difference OR "Clinically Important difference" OR Rasch)**

B. **("McGill Pain Questionnaire") AND (Psychometric OR “Measurement Properties” OR Validation OR Adaptation OR "Cross-cultural" OR Reliability OR Validity OR "Internal Consistency" OR Sensitivity OR Specificity OR Discriminative OR Responsiveness OR "Factor analysis" OR Minimal Clinically Important Difference OR "Clinically Important difference" OR Rasch)**

CHAPTER 3:

Reproducibility: reliability and agreement parameters of the Revised Short McGill Pain Questionnaire Version-2 for use in patients with musculoskeletal shoulder pain

Samuel U. Jumbo, BMR. PT. Western University, Faculty of Health and Rehabilitation Sciences, Elborn College London, Ontario, Canada. Email: **sjumbo@uwo.ca**

Joy C. MacDermid, PT, PhD. Western University, Faculty of Health and Rehabilitation Sciences, Elborn College London, Ontario, Canada; McMaster University, School of Rehabilitation Science, 1400 Main Street West, Hamilton, Ontario, Canada; Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, Ontario, Canada. Email: **jmacderm@uwo.ca**

Tara L. Packham, OT. PhD. McMaster University, School of Rehabilitation Science, 1400 Main Street West, Hamilton, Ontario, Canada. Email: **packhamt@mcmaster.ca**

George S. Athwal, MD. FRCSC. Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, Ontario, Canada. Email: **gathwal@uwo.ca**

Kenneth J. Faber, MD, MHPE, FRCSC. Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, Ontario, Canada. Email: **kjfaber@uwo.ca**

Corresponding Author: Samuel U. Jumbo, BMR. PT. Western University, Faculty of Health and Rehabilitation Sciences, Elborn College London, Ontario, Canada. **sjumbo@uwo.ca**

Funding: This work was supported by the Canadian Pain Network funded by the Canadian Institutes for Health Research. Grant Number: CIHR FRN SCA-145102.

Institutional Review Board: Not Applicable

Statement of Financial Disclosure and Conflict of Interest: None to report.

Acknowledgements: Samuel U Jumbo was supported by a Western University Graduate Research Scholarship. Joy C MacDermid was supported by a Canadian Institutes of Health Research (CIHR) Chair in Gender, Work and Health and the Dr. James Roth Chair in Musculoskeletal Measurement and Knowledge Translation.

Word Count: 4,124, Abstract 280

ABSTRACT

Study design: Test-retest.

Background: The Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) is a multidimensional outcome measure designed to capture, evaluate and discriminate pain from neuropathic and non-neuropathic sources. A recent systematic review found insufficient psychometric data with respect to musculoskeletal health conditions.

Objectives: To describe the reproducibility (reliability and agreement) and internal consistency of the SF-MPQ-2 for use among patients with musculoskeletal shoulder pain.

Methods: Eligible patients with shoulder pain completed the SF-MPQ-2 two times: at baseline (n=195), and after 3-7days (n= 48), if they remained in stable pain. Cronbach alpha (α) and intraclass correlation coefficient ($ICC_{2,1}$), and their related 95% CI were calculated. The Standard Error of Measurement (SEM), group and individual minimal detectable change (MDC90) and Bland-Altman (BA) plots were used to assess agreement.

Results: Cronbach α ranged from 0.83 to 0.95 suggesting very satisfactory internal consistency across the SF-MPQ-2 domains. Excellent $ICC_{2,1}$ scores were found in support of the total (0.95) and continuous scale (0.92); the remaining domains displayed good $ICC_{2,1}$ scores (0.78 -0.88). The Bland-Altman analysis revealed no systematic bias between the test and retest scores. While the best agreement coefficients were seen on the total scale (SEM = 0.5; MDC90 = 1.2 and MDC90group = 0.3), they were acceptable for the SF-MPQ-2 subscales (SEM: range, 0.7 - 1; MDC_{90}: range, 1.7 - 2.3; $MDC_{90group}$: range, 0.4 – 0.5).

Conclusion: Good reproducibility supports the SF-MPQ-2 domains for augmented or independent use in MSK-related shoulder pain assessment, with the total scale displaying the best reproducibility coefficients. Additional research on the validity and responsiveness of the SF-MPQ-2 is still required in this population.

Keywords: Reproducibility; Reliability; Agreement; McGill Pain Questionnaire; Shoulder Pain; Musculoskeletal Conditions; Patient-Reported Outcomes; Psychometric Properties

INTRODUCTION

Shoulder disorders are among the three leading causes of musculoskeletal pain.[29,35] Although present in all age groups, there is evidence of its increasing prevalence as age increases.[9,32] Shoulder disorders come with significant consequences on the socioeconomic wellbeing of the patient and the society; studies have linked workers' absenteeism, loss of job, and poor health-related quality of life (HRQoL) to symptoms associated with shoulder disorders.[9,23,28,42,56]

Pain assessment in clinical practice and research often places emphasis on monitoring pain intensity, even though we know pain is multidimensional and experienced uniquely by individuals.[37] Patients perceive pain in 6 multiple dimensions: physiologic, sensory, affective, cognitive, behavioral and socio-cultural.[3,37] The comprehensive assessment and monitoring of these dimensions should improve patient care.[24] A multidimensional pain assessment tool that provides a holistic assessment of pain has been recommended by experts[4,19,59] for use in upper extremity conditions, including shoulder disorders.

The Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) is an example of a general use multidimensional pain tool that comprehensively examines the sensory and affective dimensions of pain. Dworkin and colleagues[14] added seven new items to the former 15 items SF-MPQ to enhance the SF-MPQ-2 ability to explicitly examine both neuropathic and non-neuropathic pain characteristics. They also replaced the previous 4-point descriptive rating scale with a 10-item numerical rating scale to enhance its responsiveness.[14] Since then, multiple studies have utilized the improved SF-MPQ-2 as a primary outcome for pain assessment in clinical trials, and its measurement properties have been examined in different populations including cancer pain,[18] surgical pain,[43] visceral pain,[58] and neuropathic pain.[40] Among MSK conditions, studies have reported measurement evidence examined among patients with complex regional pain syndrome,[45] back pain,[15] knee OA,[26] and mixed MSK populations.[1,30] Although the SF-MPQ-2 is becoming increasingly popular, our recent review[25,31] reported on evidence with design flaws including inadequate description of ICC models, insufficient justification of retest interval, and lack of attention to absolute reliability parameters.

In the absence of such evidence, the primary purpose of this study was to investigate the reproducibility (test-retest reliability and agreement) and internal consistency of the Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) among persons with MSK-related shoulder problems.

METHODS

This study was based on a test–retest design. The SF-MPQ-2 questionnaire was administered to examine reproducibility (i.e. relative and absolute reliabilities and internal consistency) at two time points: at baseline and after 3-7 days (when patients would, for the most part, be stable).[12,34] The participants were recruited from the Roth|McFarlane Hand and Upper Limb Center, London, ON, Canada over a period of 6-months (June – November 2018). Ethical approval to recruit and review patients' clinical charts was waived by the Health Sciences Research Ethics Board of the University of Western Ontario, London, Ontario, Canada.

Patients

Adults proficient in English, above 18 years of age, that experienced pain from one or more shoulder conditions of known MSK source were included. Potential participants were excluded if they had: 1) an unstable cardiorespiratory condition; 2) any history of problems relating with the central nervous system e.g. hemiplegia; 3) pain resulting from neoplastic or infectious or vascular disorders or referred from internal organs; 4) any neuropathic pain symptoms resulting from thoracic outlet syndrome, carpal tunnel syndrome or any peripheral nerve entrapment, or 5) did not provided consent.

Procedure

Assessors (SJ and HULC research assistants) identified potentially eligible participants by reviewing the outpatient appointment list of patients scheduled for a clinical visit with two shoulder surgeons (KF and AG), a day prior. Potential participants were then contacted on the day of their clinical appointment to see whether they would be willing to participate. Consenting persons were screened to ensure all criteria were satisfied, then they received further explanation of the study's aims and objectives before the SF-MPQ-2 questionnaire was administered. Each participant was verbally instructed to carefully read and circle the one number that described their pain experience. In cases where participants had difficulty with selecting an answer, they were told to choose the answer that comes closest to describing their pain symptoms. If help was needed with understanding any words or phrases, or with marking their responses, the assessors assisted. The participants were instructed to complete all items in the questionnaire. Participants were permitted to withdraw from the study for any reason at any time. A subset of the participants were randomly selected to self-complete the SF-MPQ-2 at home after 3-7 days if their pain remained unchanged. Participants were given stamped envelopes (if they accepted) and instructed to return the completed questionnaire. A global rating of change scale was administered on both test-retest occasions and compared to ensure that we only reported on patients with stable pain (for test-retest). Demographic information including age, hand dominance, primary cause of shoulder pain and gender were noted in person and from their clinical record.

Outcome Measure

The Revised Short McGill Pain Questionnaire Version-2 (SF-MPQ-2) contains 22-items/pain descriptors and 4 subscales/domains that examine pain intensity and quality as follows: (i) continuous pain (throbbing, cramping, gnawing, aching, heavy, and tender pain); (ii) intermittent pain (shooting, stabbing, sharp pain, splitting pain, electric-shock, and piercing pain); (iii) neuropathic pain (hot-burning, cold-freezing, pain caused by light touch, itching, tingling or pins and needles, and numbness pain), and (iv) affective pain (tiring-exhausting, sickening, fearful, and punishing-cruel). All the items are bounded on a zero (none) to 10 (worst possible) numerical rating scale. The mean of the 22-items yields the SF-MPQ-2 total score, while the mean of the items that comprise each of four-subscales yields the summary score for the subscale.[14,15] Higher subscale or total scores suggest greater pain symptoms/experience, and more than 2 missing values renders patients' response to the questionnaire invalid.[15] The SF-MPQ-2 uses a recall period of 7-days, instructing the person to base their rating on their symptoms in the past week.[13]

Statistical analyses

The SF-MPQ-2 total and subscale scores were considered as interval variables. Data quality and screening, including the percentage of missing data, outliers, and presence of floor/ceiling effects was performed. Respondents with two or more missing items were excluded, in line with the developers' instructions.[15] Continuous variables were descriptively summarized using means and standard deviations while percentages were used to report categorical variables. The data was then examined for normality graphically with histograms, and statistically with the Shapiro-Wilk test. All statistical analysis were completed with Microsoft Excel Version 2013 and SPSS statistic for windows, Version 25.0. (Armonk, NY: IBM Corp, Released 2017).

Floor/ceiling effects

The SF-MPQ-2 was assessed for floor/ceiling effect by identifying the number of participants with the absolute lowest (0-points = floor) and highest scores (10-points = ceiling) on its total and subscales. Floor/ceiling effects occurring at the magnitude of 15% were considered substantial.[52]

Hypothesis: We expected substantial floor effects on the neuropathic and affective subscales of the SF-MPQ-2 because they evaluate pain dimensions that are relatively uncommon in orthopaedic shoulder disorders.

Cross sectional reliability (Internal consistency)

Internal consistency, the degree of item inter-relatedness/equivalence in a PROM,[11,51,52] was assessed with Cronbach alpha (α) and associated 95% confidence intervals. A commonly accepted requirement for internal consistency reliability is that it should be at above 0.7. However, redundancy was established at $\alpha > 0.95$.[50–52]

Hypothesis: We expected the SF-MPQ-2 to be internally consistent with Cronbach alpha (α) at 0.8 or above for its subscale scores, and 0.9 or above for its total scores.

Relative reliability (Test-retest reliability)

The intraclass correlation coefficient ($ICC_{2,1}$) was used to assess the retest reliability of the SF-MPQ-2 total and subscales.[48] $ICC_{2,\,1}$ with 95% confidence interval (CI) were computed using the two-way mixed and absolute agreement model, that assumes the patients were randomly selected but the occasions were fixed choices.[47] We chose $ICC_{2,1}$ absolute agreement over consistency model because it captures elements of systematic bias and is preferred for computing absolute reliability indicator. $ICC_{2,1}$ values for the SF-MPQ-2 total and subscale scores were considered Negative ≤ 0.49, Doubtful 0.50–0.69, Good 0.70–0.89, and Excellent 0.90–1.00.[36]

Hypothesis: We expected adequate $ICC_{2,1}$ scores for group level analysis at ≥ 0.80 (total score), and ≥ 0.70 (subscale score) as previously reported in the literature.[1,26]

Absolute reliability (Standard Error of Measurement [SEM] and Minimal Detectable Change [MDC])

Standard Error of Measurement (SEM) is defined as the standard deviation of errors of measurement associated with particular test takers scores.[22] **Table 1** explains the five equations used for agreement analysis. To define $SEM_{agreement}$ for the SF-MPQ-2 total and subscales scores, the pooled standard deviation calculated from participants mean responses to the SF-MPQ-2 domains on both test and retest using **equation 1**[22,57] and the respective non-transformed $\text{ICC}_{2,1}$ for the SF-MPQ-2 domain under evaluation was keyed into **equation 2**[22,44,57] **(Table 1)**. Further, the proportion of the resulting SEM per domain to the total score of the scale was calculated to yield the SEM percentage or SEM%, as previously used[5,44,49] and interpreted as follows: ≤5% = very good; >5% to ≤10% = good; >10% to <20% = doubtful; and values above 20% = negative[44]

The minimal detectable change (MDC) or repeatability coefficient describes the minimum amount of change that must be seen on a tool scores to be confident that true/real change has occurred without error after two repeated measure, within the period of the test-retest.[21] For this study, a 90% confidence interval was estimated for the Minimal Detectable Change (MDC_{90}). Like the SEM, it is also expressed in the unit of the measure and may be computed at an individual level ($\text{MDC}_{90\text{individual}}$) or for a group ($\text{MDC}_{90\text{group}}$)[12]. We estimated $\text{MDC}_{90\text{individual}}$ for the total and subscale scores of the SF-MPQ-2 by entering each scales $\text{SEM}_{\text{agreement}}$ into **equation 3 (Table 1)** assuming the data was normally distributed and free of systematic error. The $\text{MDC}_{90\text{individual}}$ confidence interval was then computed from the mean differences **(d)** of each subscale using **equation 4 (Table 1)**[5,10,12] To determine the group level minimal detectable change ($\text{MDC}_{90\text{group}}$), which is useful for determining if changes have occurred in an entire population, **equation 5 (Table 1)** the formula proposed by de Vet et al.[52,55] was employed. Furthermore, as was estimated for SEM, the proportion of the resulting MDC coefficient per SF-MPQ-2 domain to the total score of the scale was computed, as previously done,[5,44] to yield the MDC percent score (MDC%) and interpreted was follows: ≤5% = very good; >5% to ≤10% = good; >10% to <20% = doubtful; and values above 20% = negative.[44]

Bland-Altman Plots (BA Plots):

The Bland-Altman method was used to visually examine the agreement between the test and retest scores.[6,7] Scatter plots for the total and subscales scores were each plotted for the difference between scores obtained at time one and time two of the test-retest interval against their mean score for the two time points.[6–8,41] We then calculated the mean difference between the two measurement intervals (the 'bias') and the 95% limits of agreement (LOA) using: LOA = mean difference (d) ± 1.96 SD of the mean differences. The BA plots were used to visually judge the 95% limits of agreement to determine how well score from repeated measurements agreed: narrower LOAs suggested better agreement at the individual level.[12,17,41] Agreement at the group level was determined by how close the bias (mean difference) was to zero. Also, the distribution of scatter points on the BA plots were visually scrutinized for evidence of variability or heteroscedasticity, where the absence of a linear relationship between test-retest mean differences and their mean scores, per subscale, suggest the absence of systematic bias.[6–8,41,54,55] Furthermore, linear regression models were used to explore the presence of systematic bias. For each domain of the SF-MPQ-2, mean scores and differences in mean scores were modelled as the independent and dependent variables, respectively. The presence of systematic bias was confirmed by a significant prediction of the differences by the means scores.[41,53] Finally, outliers that presented beyond the upper and lower boundaries of the LOA were noted and explored.[12,16]

RESULTS

Figure 1 summarizes the flow of participants through the different phases of the study. Of the 238 potential patients identified from the review of scheduled appointment list, 195 consenting adults that satisfied the inclusion criteria, provided complete data that was considered in our analysis of cross-sectional reliability. For the analysis of relative and absolute reliability, 48 out of 55 stable patients returning completed copies of the SF-MPQ-2 did not have missing data; the mean duration for retest response was 4 days. **Table 2** summarizes the characteristic and demographic distribution of the baseline population. Persons completing this study could be described as an older population, (mean age = 62 years), representing nearly equal proportions of males and females, and presenting with different shoulder problems of various MSK pathologies including rotator cuff injuries, humeral fracture and arthroplasty, and shoulder pain.

Both the graphical and statistical tests of normality revealed the dataset was skewed/abnormal. To address the assumption of normality for further analysis, a square root calculation was used to transform the data. A closer look at the reliability coefficients obtained using the transformed and untransformed data revealed only a very trivial difference in scores. Because our sample size was large enough, and beyond 30 participants (based on the central limit theorem), parametric statistics were adopted in our analysis. Despite that, we still checked for differences in reproducibility coefficients obtained using the transformed and non-transformed ICC scores **(see Table 3 for results).**

Floor and ceiling effects

The presence of floor/ceiling effect may suggest an outcome measure is not responsive to detecting improvement (ceiling effect) even though decline in status can be captured, and vice versa – for floor effects.[15] The number of patients who obtained the absolute maximum (Ten, 10) and minimal (zero, 0) scores on the SF-MPQ-2 total and subscales are summarized in **Table 3**. The greatest level of floor effects was observed on the affective subscale at both periods of the test-retest. Substantial floor effects were also noted on the neuropathic and intermittent subscales. None of the SF-MPQ-2 indices had remarkable ceiling effects.

Reliability

Internal consistency (cross-sectional reliability)

Table 4 summarizes the result obtained for cross sectional reliability. The SF-MPQ-2 displayed excellent internal consistency with robust alpha coefficients presenting within a range that suggest the absence of redundancy: alpha coefficients for the total subscale peaked at 0.95 as posited, while that for the subscales fluctuated around 0.83 to 0.86 points. Inter-item correlations were satisfactory, ranging from 0.23-0.53 across the scales.

Relative Test-retest reliability

Good to excellent results were seen in support of test-retest reliability of the SF-MPQ-2 domains (**Table 5).** Our results for ICC$_{2,1}$ was based on analysis conducted with the non-transformed data, as they did not differ from that obtained with transformed data. ICC$_{2,1}$ scores were highest on the continuous and total subscales and rated excellent according to our criteria. Also, the neuropathic, affective and intermittent subscales displayed good ICC$_{2,1}$ coefficient **(Table 4)** in support of relative reliability.

Absolute test-retest reliability (agreement parameters)

Table 5 summarize the absolute reliability coefficients supporting the SF-MPQ-2 domains. The total scale SEM$_{\text{agreement}}$ was very low (0.51points) and approximately 5% of the total score of the scale, which is 'very good' according to our criteria. Individual subscale SEM$_{\text{agreement}}$ ranged from 0.73 -0.99 (approximately ≤ 10 % of the total score), which is 'good' according to our criteria. At the individual level, acceptable scores within 1.19 – 2.29 points were seen in support of minimal detectable change at 90% confidence level. The best and worst scores were noted on the total scale (1.19 point, i.e. 11.9% of the total score) and the intermittent subscale (2.29 point, i.e. 22.9% of the total score), respectively. For Group MDC$_{90}$, estimates were acceptable and expectedly lower than those obtained for MDC$_{90\text{individual}}$; the results fluctuated within 0.28 (total) to 0.54 (intermittent) points across the SF-MPQ-2 domains (**Table 5).**

Bland-Altman Analysis/Plots

The results of our Bland–Altman analysis are presented in **Table 4**. Also, Bland-Altman plots superimposed with the LoA and mean difference (bias) scores for each domain of the SF-MPQ-2 are graphically illustrated **(Figure 2 to 6)**. All the SF-MPQ-2 domains displayed acceptable LoA at 95% confidence level with the highest distance ranging 5 points (intermittent subscale). The total scale score displayed the narrowest LoA (range = 3 points), with the remaining subscales within satisfactory limits. Mean difference scores (bias) were very acceptable for all the SF-MPQ-2 domains (0.15 – 0.19 points).

Visual inspection of scatter points on the BA plots for each domain of the SF-MPQ-2 revealed that the magnitude of mean difference against the mean scores were uniformly distributed from the point of zero and most scatter points were within the 95% Limit of Agreement but for few outliers. This supports the absence of systematic bias and suggest a good level of agreement among test-retest scores. Furthermore, for each of the SF-MPQ-2 domains, there was no evidence of the mean difference scores predicting the mean average after our regression model analysis. This gives more weight to the absence of systematic bias and confirms good level of agreement between the test-retest scores **(Table 5).**

The few outliers noted were explored. First, we determined if they were erroneous responses in entry by rechecking hard copies but, indeed, they were 'interesting' outliers [2] and labelled according to their #RS on each BA plot. The greatest number of interesting outliers presented on the intermittent (n=6, 12.5%) and neuropathic (n=4, 10%) subscales. The least number of outliers were seen on the affective subscale (n=2, 4.1%). In general, however, the presence of these outliers did not indicate the presence or absence of bias.[2]

DISCUSSION

This study provides strong support of the reproducibility of the SF-MPQ-2 for use in multidimensional pain assessment of people with musculoskeletal shoulder pain. We found good to excellent reproducibility coefficients in support of internal consistency, relative reliability and absolute reliability. The limits of agreement for the subscales and total scores were very satisfactory.

Although some floor effects can be expected on the neuropathic, intermittent and affective subscales, we attribute this to the lower prevalence of these problems in our populations and the high discriminative property of the SF-MPQ-2 subscales. Conceptually, the SF-MPQ-2 was expanded to provide a single tool that can classify pain from both neuropathic and nociceptive sources.[14,15] As outcome measures can be evaluative or discriminative, combining both purposes within an outcome measure is likely to result in these type of issues. For instance, participants with pain emerging from neuropathic sources are more inclined to respond adequately to the neuropathic subscale with no floor effect, as has been observed with the use of the SF-MPQ-2 among CRPS patients.[45] This implies that floor effects on the SF-MPQ-2 domains may not always represent redundancy but may suggest that an item does not describe the patient's pain experience.[25]

Cross sectional reliability was established for the SF-MPQ-2 total and subscale scores with satisfactory coefficients supporting internal consistency in line with previous estimates among mixed-MSK[30] (total, 0.93; subscale, 0.84-0.92), CRPS[45] (total, 0.95; neuropathic subscale, 0.83), knee OA[26] (total, 0.88; subscale 0.75-0.81) and acute back pain[15] (total, 0.93; subscale, 0.77-0.84) patient populations. Inter-item correlations were also adequate. The adequate Cronbach's alpha obtained signifies the absence of redundancy in the domains of SF-MPQ-2 thus confirming their unidimensionality[51] to capture the different pain characteristics they assess.

In the present study, $ICC_{2,1}$ coefficients were good to excellent for all the SF-MPQ-2 domain scores (total, 0.93; subscales, 0.78 - 0.91), suggesting they can discriminate patients adequately at the individual level (total and continuous scale), and at the group level (all the SF-MPQ-2 domains).[12,27] These results are comparable or better than previous findings reporting estimates among knee OA[26] (total, 0.90; subscale, 0.73-0.90) and mixed MSK patients[1,33] (total, 0.90-0.941; subscale, 0.73-0.90). Although acceptable, the low performance of the neuropathic subscale (0.78), with an ICC score that overlapped the ‘moderate’ confidence interval threshold suggest less variability on this subscale which makes it more difficult to achieve a high $\mathrm{ICC}_{2,1}$ score.

Absolute reliability estimates allow clinicians to assess true change in a patient in comparison to change that might be expected from measurement error.[52,55] Currently, no previous data have examined absolute reliability indices for the SF-MPQ-2 scores in any population. This makes direct interpretation and comparison difficult; however, our use of Ostelo et al.[44] definition of SEM and MDC by percentages allows comparison across the domains of the SF-MPQ-2, and with its former version (SF-MPQ). The SEM for the total score (≤ 5% of total scale score) was ‘very good’ and comparable to that reported for the former version (SF-MPQ) among OA patients (≤ 3.64%)[20] but better than those seen among mixed MSK patients assessed with the Norwegian version of the SF-MPQ (≤ 10%).[49] Although not as favorable as estimates noted on the total subscale, the affective and intermittent/continuous (or sensory subscale on the previous version) subscales displayed ‘good’ SEM (≤ 10%) that was similar to that seen among OA patient (≤10%)[20] or better than those seen among mixed MSK (≤14%)[49] with the previous SF-MPQ version. Basically, SEM estimates for all the SF-MPQ-2 subscales were satisfactory and suggest an adequate evaluative capacity that can yield scores less prone to error when utilized by researchers/clinicians for MSK shoulder pain assessment over time.

The MDC scores represents the minimal change in scores after repeated administration that clinicians/researchers can interpret is not due to error for an individual or group in a population.[21] The $\mathrm{MDC}_{90\mathrm{indivdiual}}$ scores obtained for the SF-MPQ-2 domains implies that change at a magnitude equal or greater than 1.8 (neuropathic), 1.7 (affective), 1.8 (continuous), 2.3 (intermittent), 1.2 (total) points represents genuine improvement beyond error with 90 percent confidence. The MDC scores for the total scale (≤ 11.9% of the total score of the scale) were comparable to previous studies with the former version (SF-MPQ) among OA patients (≤ 11.5%) and better than the results seen among mixed MSK patients (≤

26.4% of total score). The MDC$_{90\text{group}}$ means that change of atleast 0.4 (affective), 0.5 (intermittent), 0.3 (total), 0.4 (neuropathic), 0.4 (continuous) points must be noted for a group to be 90-percent confident that it is change beyond random or systematic error. In general, MDC scores are useful when interventions are administered: to be sure the intervention is effective, it must demonstrate change beyond the MDC score reported for the scale. Also, MDC$_{90\text{group}}$ indices can be used for sample size estimation in a randomized controlled trial, as they determine the number of participants that will be needed to detect a change on the measure beyond error for a group, if the Minimal Clinical Important Difference score for the population is unknown.

The Bland-Altman plots revealed very satisfactory LoA in support of the SF-MPQ-2 subscales. Although the interpretation of how far apart two measurements can be before they are no longer considered interchangeable depends on the contextual application,[41] the LoA between test-retest of the SF-MPQ-2 domains were reasonably smaller than those seen in previous studies with its former version (SF MPQ)[20,49] and suggest minimal variation between the occasion of test-retest.[54] Furthermore, no bias was found in measurement between the test-retest, since the inter-occasion mean difference was minimal. This suggests that learning or test accommodation are not issues with using the SF-MPQ-2; moreover, our compliance to recommended time intervals (3-7 days)[12,34,38] may have favored the agreement outcomes. The intermittent subscale had the greatest number of outliers of all the BA plots (12.5%) and may be from the highly volatile nature of the pain descriptors comprising the scale.

The SF-MPQ-2 total scores displayed the best reproducibility parameters in support of its relative, absolute and level of agreement parameters. This could be from the number of items contained in the scale. For instance, better ICC scores can be expected when variability is low. And among other factors, variability decreases when a greater number of descriptors comprise a scale, in comparison to those with fewer descriptors.[12] As all 22 items of the SF-MPQ-2 contribute to the summary total scale scores, it is possible this favors reproducibility.

STUDY LIMITATIONS AND AREAS FOR FUTURE RESEARCH

While the present study findings provide preliminary evidence supporting the reproducibility of the SF-MPQ-2 for use in shoulder problems, it has several limitations. First, the study sample size (48 participants) was just under 50 participants as recommended by the COSMIN.[39,46] Second, the patient population were from a single tertiary referral practice, hence our findings may not be the same in a less differentiated cohort; it may also impact on generalizability. Third, since participants completed the retest (Time 2) at home, we were unable to clarify instructions. However, independent completion is a requirement for routine administration. Further, the high level of agreement between scores of the tests and the absence of systematic bias suggest this was not a problem. Fourth, sample mean age was 62 (± 17.3) years, which may not adequately reflect the reliability of younger populations. Finally, we did not determine minimal clinically important difference.

CONCLUSION

We conclude that the SF-MPQ-2 provides good to excellent test-retest reliability for multidimensional pain assessment among patients with musculoskeletal shoulder pain conditions.

REFERENCES

1. Adelmanesh F, Jalali A, Attarian H, et al. Reliability, Validity, and Sensitivity Measures of Expanded and Revised Version of the Short-Form McGill Pain Questionnaire (SF-MPQ-2) in Iranian Patients with Neuropathic and Non-Neuropathic Pain. *Pain Med*. 2012;13(12):1631-1638. doi:10.1111/j.1526-4637.2012.01517.x.

2. Aguinis H, Gottfredson RK, Joo H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organ Res Methods*. 2013;16(2):270-301. doi:10.1177/1094428112470848.

3. Ahles TA, Blanchard EB, Ruckdeschel JC. The multidimensional nature of cancer-related pain. *Pain*. 1983;17(3):277-288. doi:10.1016/0304-3959(83)90100-8.

4. Badalamente M, Coffelt L, Elfar J, et al. Measurement Scales in Clinical Research of the Upper Extremity, Part 2: Outcome Measures in Studies of the Hand/Wrist and Shoulder/Elbow. *J Hand Surg Am*. 2013;38(2):407-412. doi:10.1016/j.jhsa.2012.11.029.

5. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. *Qual Life Res*. 2001;10(7):571-578. doi:10.1023/A:1013138911638.

6. Bland JM., Altman DG. Statiscal Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *Lancet*. 1986;327(8476):307-310. doi:10.1016/S0140-6736(86)90837-8.

7. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135-160. doi:10.1191/096228099673819272.

8. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ*. 1992;304(6840):1491-1494. doi:10.1136/bmj.304.6840.1491.

9. Chard MD, Hazleman R, Hazleman BL, King RH, Reiss BB. Shoulder disorders in the elderly: a community survey. *Arthritis Rheum*. 1991;34(6):766-769.

doi:10.1002/art.1780340619.

10. Chesworth BM, Hamilton CB, Walton DM, et al. Reliability and Validity of Two Versions of the Upper Extremity Functional Index. *Physiother Canada*. 2014;66(3):243-253. doi:10.3138/ptc.2013-45.

11. Cronbach LJ. Test “reliability”: Its meaning and determination. *Psychometrika*. 1947;12(1):1-16. doi:10.1007/BF02289289.

12. Dewan N, MacDermid JC, MacIntyre N, Grewal R. Reproducibility: Reliability and agreement of short version of Western Ontario Rotator Cuff Index (Short-WORC) in patients with rotator cuff disorders. *J Hand Ther*. 2016;29(3):281-291. doi:10.1016/J.JHT.2015.11.007.

13. Dworkin RH, Turk DC, Revicki DA, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). *Pain*. 2009;144(1):35-42. doi:10.1016/j.pain.2009.02.007.

14. Dworkin RH, Turk DC, Revicki DA, et al. Development and initial validation of an expanded and revised version of the Short-form McGill Pain Questionnaire (SF-MPQ-2). *Pain*. 2009;144(1):35-42. doi:10.1016/j.pain.2009.02.007.

15. Dworkin RH, Turk DC, Trudeau JJ, et al. Validation of the Short-Form McGill Pain Questionnaire-2 (SF-MPQ-2) in Acute Low Back Pain. *J Pain*. 2015;16(4):357-366. doi:10.1016/j.jpain.2015.01.012.

16. Flansbjer U-B, Holmbäck AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med*. 2005;37(2):75-82. doi:10.1080/16501970410017215.

17. Fox B, Henwood T, Neville C, Keogh J. Relative and absolute reliability of functional performance measures for adults with dementia living in residential aged care. *Int Psychogeriatrics*. 2014;26(10):1659-1667. doi:10.1017/S1041610214001124.

18. Gauthier LR, Young A, Dworkin RH, et al. Validation of the Short-Form McGill Pain Questionnaire-2 in Younger and Older People With Cancer Pain. *J Pain*. 2014;15(7):756-770. doi:10.1016/j.jpain.2014.04.004.

19. Goldhahn J, Beaton D, Ladd A, Macdermid J, Hoang-Kim A. Recommendation for measuring clinical outcome in distal radius fractures: a core set of domains for standardized reporting in clinical practice and research. *Arch Orthop Trauma Surg*. 2014;134(2):197-205. doi:10.1007/s00402-013-1767-9.

20. Grafton K V, Foster NE, Wright CC. Test-Retest Reliability of the Short-Form McGill Pain Questionnaire. *Clin J Pain*. 2005;21(1):73-82. doi:10.1097/00002508-200501000-00009.

21. Haley SM, Fragala-Pinkham MA. Interpreting Change Scores of Tests and Measures Used in Physical Therapy. *Phys Ther*. 2006;86(5):735-743. doi:10.1093/ptj/86.5.735.

22. Harvill LM. An NCME Instructional Module on. Standard Error of Measurement. *Educ Meas Issues Pract*. 1991;10(2):33-41. doi:10.1111/j.1745-3992.1991.tb00195.x.

23. van der Heijden GJMG. Shoulder disorders: a state-of-the-art review. *Baillieres Best Pract Res Clin Rheumatol*. 1999;13(2):287-309. doi:10.1053/berh.1999.0021.

24. Ho K, Spence J, Murphy MF. Review of pain-measurement tools. *Ann Emerg Med*. 1996;27(4):427-432. doi:10.1016/S0196-0644(96)70223-8.

25. Jumbo S, MacDermid JC, Michael K, Packham TL, Athwal GS, Faber K. Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review. 2019.

26. Kachooei AR, Ebrahimzadeh MH, Erfani-Sayyar R, Salehi M, Salimi E, Razi S. Short Form-McGill Pain Questionnaire-2 (SF-MPQ-2): A Cross-Cultural Adaptation and Validation Study of the Persian Version in Patients with Knee Osteoarthritis. *Arch bone Jt Surg*. 2015;3(1):45-50. doi:10.22038/ABJS.2015.3827.

27. Kottner J, Audige L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Int J Nurs Stud*. 2011;48(6):661-671. doi:10.1016/j.ijnurstu.2011.01.016.

28. Largacha M, Parsons IM, Campbell B, Titelman RM, Smith KL, Matsen F. Deficits in shoulder function and general health associated with sixteen common shoulder diagnoses: A study of 2674 patients. *J Shoulder Elb Surg*. 2006;15(1):30-39. doi:10.1016/j.jse.2005.04.006.

29. Linsell L, Dawson J, Zondervan K, et al. Prevalence and incidence of adults consulting for shoulder conditions in UK primary care; patterns of diagnosis and referral. *Rheumatology*. 2006;45(2):215-221. doi:10.1093/rheumatology/kei139.

30. Lovejoy TI, Turk DC, Morasco BJ. Evaluation of the Psychometric Properties of the Revised Short-Form McGill Pain Questionnaire. *J Pain*. 2012;13(12):1250-1257. doi:10.1016/j.jpain.2012.09.011.

31. MacDermid J, Jumbo S, Kalu M, Packham T, Athwal G, Faber K. Measurement Properties of the Brief Pain Iinventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire Version-2 (SF-MPQ-2) in Pain-related Musculoskeletal Conditions: A Systematic Review. In: *Abstracts Accepted for Publication*. Vol 78. BMJ Publishing Group Ltd and European League Against Rheumatism; 2019:2128.1-2128. doi:10.1136/annrheumdis-2019-eular.3525.

32. Macdermid JC, Khadilkar L, Birmingham TB, Athwal GS. Validity of the QuickDASH in Patients With Shoulder-Related Disorders Undergoing Surgery. *J Orthop Sport Phys Ther*. 2015;45(1):25-36. doi:10.2519/jospt.2015.5033.

33. Maruo T, Nakae A, Maeda L, et al. Validity, Reliability, and Assessment Sensitivity of the Japanese Version of the Short-Form McGill Pain Questionnaire 2 in Japanese Patients with Neuropathic and Non-Neuropathic Pain. *Pain Med*. 2014;15(11):1930-1937. doi:10.1111/pme.12468.

34. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. *J Clin Epidemiol*. 2003;56(8):730-735. doi:10.1016/S0895-4356(03)00084-2.

35. McCormick A, Fleming D, Charlton J. Morbidity statistics from general practice: fourth national study 1991-92. *Ser MB5 no 3*. 1995. https://ci.nii.ac.jp/naid/10016136454/. Accessed August 25, 2018.

36. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30-46. doi:10.1037//1082-989X.1.1.30.

37. McGuire DB. Comprehensive and multidimensional assessment and measurement of pain. *J Pain Symptom Manage*. 1992;7(5):312-319. doi:10.1016/0885-3924(92)90064-O.

38. Mehta SP, Mhatre B, MacDermid JC, Mehta A. Cross-cultural Adaptation and Psychometric Testing of the Hindi Version of the Patient-rated Wrist Evaluation. *J Hand Ther*. 2012;25(1):65-78. doi:10.1016/j.jht.2011.08.001.

39. Mokkink LB, de Vet HCW, Prinsen CAC, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res*. 2018;27(5):1171-1179. doi:10.1007/s11136-017-1765-4.

40. Morgan KJ, Anghelescu DL. A Review of Adult and Pediatric Neuropathic Pain Assessment Tools. *Clin J Pain*. 2017;33(9):844-852. doi:10.1097/AJP.0000000000000476.

41. Myles PS, Cui J. I. Using the Bland–Altman method to measure agreement with repeated measures. *Br J Anaesth*. 2007;99(3):309-311. doi:10.1093/bja/aem214.

42. Nygren A, Berglund A, von Koch M. Neck-and-shoulder pain, an increasing problem. Strategies for using insurance material to follow trends. *Scand J Rehabil Med Suppl*. 1995;32:107-112. http://www.ncbi.nlm.nih.gov/pubmed/7784832. Accessed August 25, 2018.

43. Ortner C, Turk D, Theodore B, Siaulys M, Bollag L, Landau R. The short-formmcgill pain questionnaire-revised to evaluate persistent pain and surgery-related symptoms in healthy women undergoing a planned cesarean delivery. *Reg Anesth Pain Med*. 2014;39(6):478-486. doi:10.1097/AAP.0000000000000158.

44. Ostelo RWJG, de Vet HC., Knol DL, van den Brandt PA. 24-item Roland-Morris Disability Questionnaire was preferred out of six functional status questionnaires for post-lumbar disc surgery. *J Clin Epidemiol*. 2004;57(3):268-276. doi:10.1016/j.jclinepi.2003.09.005.

45. Packham TL, Bean D, Johnson MH, et al. Measurement Properties of the SF-MPQ-2 Neuropathic Qualities Subscale in Persons with CRPS: Validity, Responsiveness, and Rasch Analysis. *Pain Med*. October 2018. doi:10.1093/pm/pny202.

46. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27(5):1147-1157. doi:10.1007/s11136-018-1798-3.

47. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil*. 1998;12(3):187-199. doi:10.1191/026921598672178340.

48. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428. doi:10.1037/0033-2909.86.2.420.

49. Strand LI, Ljunggren AE, Bogen B, Ask T, Johnsen TB. The Short-Form McGill Pain Questionnaire as an outcome measure: Test-retest reliability and responsiveness to change. *Eur J Pain*. 2008;12(7):917-925. doi:10.1016/j.ejpain.2007.12.013.

50. Streiner D, Norman G, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. 5th ed. New York, USA: Oxford University Press; 2015.

51. Tavakol M, Dennick R. Making sense of Cronbach’s alpha. *Int J Med Educ*. 2011;2:53-55. doi:10.5116/ijme.4dfb.8dfd.

52. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42. doi:10.1016/j.jclinepi.2006.03.012.

53. Tsang P, Walton D, Grewal R, MacDermid J. Validation of the QuickDASH and DASH in Patients With Distal Radius Fractures Through Agreement Analysis. *Arch Phys Med Rehabil*. 2017;98(6):1217-1222.e1. doi:10.1016/j.apmr.2016.11.023.

54. Uddin Z, MacDermid JC, Ham HH. Test–retest reliability and validity of normative cut-offs of the two devices measuring touch threshold: Weinstein Enhanced Sensory Test and Pressure-Specified Sensory Device. *Hand Ther*. 2014;19(1):3-10. doi:10.1177/1758998313515191.

55. de Vet HC, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. *Int J Technol Assess Health Care*. 2001;17(4):479-487. https://www.cambridge.org/core/journals/international-journal-of-technology-assessment-in-health-care/article/reproducibility-and-responsiveness-of-evaluative-outcome-measures/7392A84979AB714F22150F79504E6860. Accessed March 10, 2019.

56. Virta L, Joranger P, Brox JI, Eriksson R. Costs of shoulder pain and resource use in primary health care: a cost-of-illness study in Sweden. *BMC Musculoskelet Disord*. 2012;13(1):17. doi:10.1186/1471-2474-13-17.

57. Walton D, MacDermid J, Nielson W, Teasell R, Chiasson M, Brown L. Reliability, Standard Error, and Minimum Detectable Change of Clinical Pressure Pain Threshold Testing in People With and Without Acute Neck Pain. *J Orthop Sport Phys Ther*. 2011;41(9):644-650. doi:10.2519/jospt.2011.3666.

58. Wang J-L, Zhang W-J, Gao M, Zhang S, Tian D-H, Chen J. A cross-cultural adaptation and validation of the short-form McGill Pain Questionnaire-2: Chinese

version in patients with chronic visceral pain. *J Pain Res*. 2017;Volume 10:121-128. doi:10.2147/JPR.S116997.

59. Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. Comparative responsiveness of generic and specific quality-of-life instruments. *J Clin Epidemiol*. 2003;56(1):52-60. doi:10.1016/S0895-4356(02)00537-1.

Figure 1: Flow chart of progress through the phases of screening, recruitment, test, retest and data analysis.

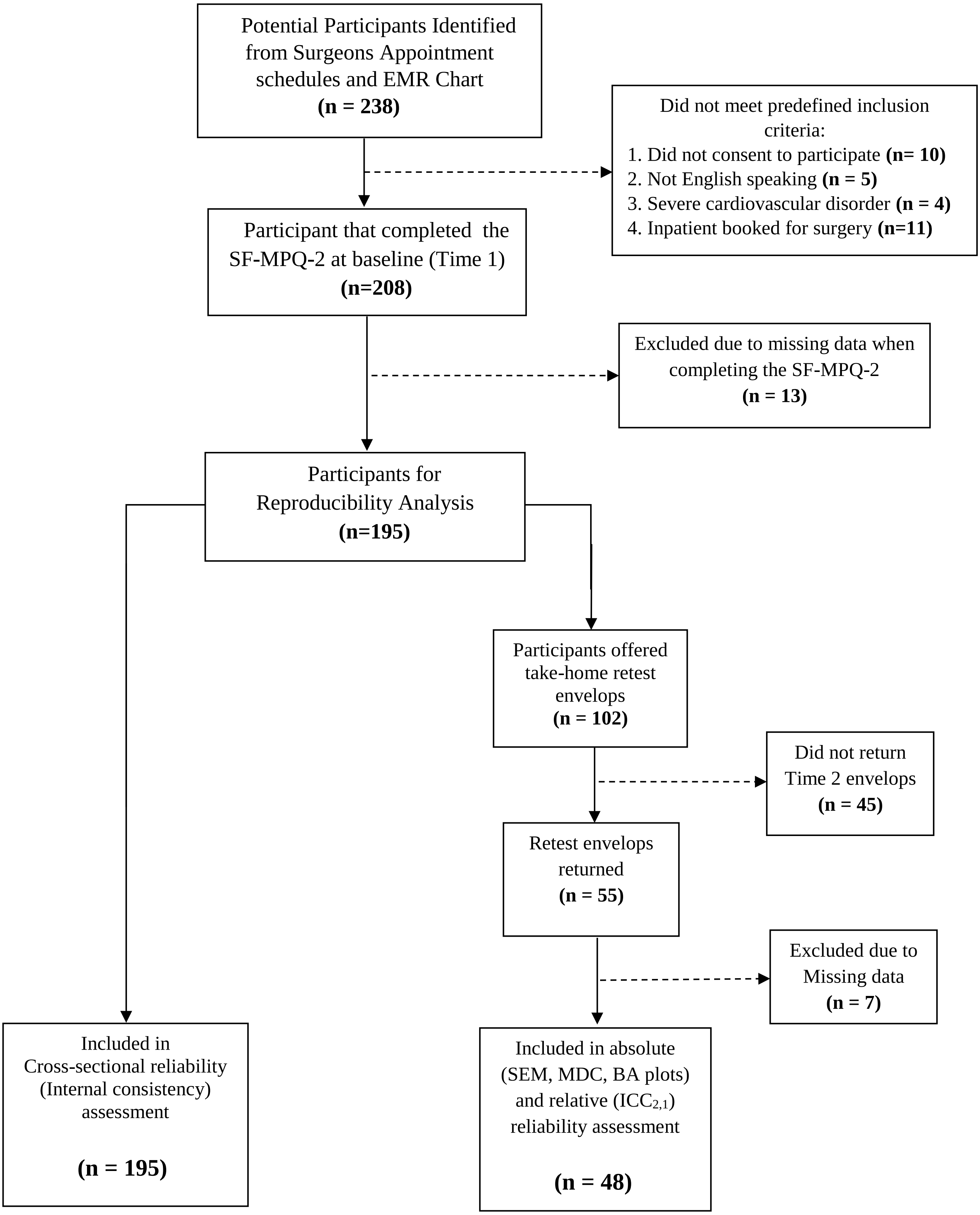

Potential Participants Identified from Surgeons Appointment schedules and EMR Chart (n = 238)
Did not meet predefined inclusion criteria:
1. Did not consent to participate (n= 10)
2. Not English speaking (n = 5)
3. Severe cardiovascular disorder (n = 4)
4. Inpatient booked for surgery (n=11)
Participant that completed the SF-MPQ-2 at baseline (Time 1) (n=208)
Excluded due to missing data when completing the SF-MPQ-2 (n = 13)
Participants for Reproducibility Analysis (n=195)
Participants offered take-home retest envelops (n = 102)
Did not return Time 2 envelops (n = 45)
Retest envelops returned (n = 55)
Excluded due to Missing data (n = 7)
Included in Cross-sectional reliability (Internal consistency) assessment (n = 195)
Included in absolute (SEM, MDC, BA plots) and relative (ICC$_{2,1}$) reliability assessment (n = 48)

TABLE 1: Summary of Equations Used in Agreement Analysis

EQUATION	FORMULA	PURPOSE
1	$SD_{pooled} = (SD_{test} + SD_{retest}) / 2$	For estimating pooled standard deviation (SD_{pooled}) from the test and retest scores. The SD_{pooled} is among the indices required for $SEM_{agreement}$ estimation.
2	$SEM_{agreement} = \text{Standard Deviation}_{pooled} \times \sqrt{1 - ICC_{2,1}}$	For estimating $SEM_{agreement}$, which is important for the $MDC90_{individual}$ estimation.
3	$MDC_{90individual} = 1.64 \times \sqrt{2} \times SEM_{agreement}$	For determining the point estimate of $MDC_{90individual}$, which is required for estimating the confidence interval range and the $MDC_{90group}$ scores per subscale of the SF-MPQ-2
4	95% CI for $MDC_{90individual} = d \pm MDC_{90individual}$	For computing the 90% confidence interval range for the $MDC_{90individual}$ score obtained for each subscale of SF-MPQ-2
5	$MDC_{90group} = MDC_{90individual} / \sqrt{n} \times 1.64$	For estimating the $MDC_{90\ group}$ score for the entire population.

Key: ***$SEM_{agreement}$*,** Standard Error of Measurement (agreement); ***SD_{test}*,** Standard Deviation of test scores; ***SD_{retest}*,** Standard deviation of retest scores; ***SD_{pooled}*,** pooled Standard Deviation; ***n*,** sample size; ***CI*,** confidence interval; ***$MDC_{90individual}$*,** Individual level Minimal Detectable Change at 90% CI; ***$MDC_{90group}$*,** Group level Minimal Detectable Change at 90% CI; ***d*,** mean difference; ***$ICC_{2,1}$*,** Intraclass correlation coefficient.

TABLE 2: Patient Baseline Characteristic (N = 195)

Variables	**N / %**
Age in years (mean ± SD)	(62 ± 17.3) 195/100%
Shoulder problem	
Arthroplasty	39 / 20%
Fracture humeral & others	23 / 12%
Rotator cuff pathologies	48 / 25%
Pain	40 / 21%
Dislocation	12 / 6%
OA	18 / 9%
Impingement/bursitis	15 / 8%
Affected Shoulder	
Right	111 / 56%
Left	71 / 36%
Both	13 / 6%
Sex	
Males	103 / 53 %
Females	92 / 47%

N, number of patients; ***SD,*** Standard deviation

TABLE 3: Floor and ceiling effects for test-retest scores of the SF MPQ-2 total and subscale scores (N= 48)

Variables	**Test**		**Retest**	
	Floor	**Ceiling**	**Floor**	**Ceiling**
SF-MPQ-2 **Continuous**	7/48 = 14.6%	0/48 = 0%	4/48 = 8.3%	1/48 = 2.1%
SF-MPQ-2 **Intermittent**	11/48 = 22.9%	0/48 = 0%	15/48 = 31.3%	0/48 = 0%
SF-MPQ-2 **Affective**	19/48 = 39.6%	1/48 = 2.1%	20/48 = 41.7%	0/48 = 0%
SF-MPQ-2 **Neuropathic**	14/48 = 29.2%	0/48 = 0%	11/48 = 22.9%	0/48 = 0%
SF-MPQ-2 **Total**	3/48 = 6.3%	0/48 = 0%	4/48 = 8.3%	0/48 = 0%

SF-MPQ-2, Revised Short McGill Pain Questionnaire Version-2; ***%,*** proportion in percentages

TABLE 4: Cross-sectional Reliability of the SF-MPQ-2 total and subscale scores (N=195)

Variables	**Internal consistency (N=195)**	
	Cronbach alpha (95% CI)	**Inter-item correlation**
SF-MPQ-2 **Continuous**	0.87 (0.84 – 0.90)	0.43 – 0.67
SF-MPQ-2 **Intermittent**	0.87 (0.84 – 0.90)	0.42 – 0.77
SF-MPQ-2 **Neuropathic**	0.85 (0.81 – 0.88)	0.32 – 0.81
SF-MPQ-2 **Affective**	0.83 (0.79 – 0.87)	0.44 – 0.78
SF-MPQ-2 **Total**	0.95 (0.94 – 0.96)	0.21 – 0.78

SF-MPQ-2, Revised Short McGill Pain Questionnaire Version-2; ***CI,*** Confidence Interval

TABLE 5: Absolute reliability (agreement parameters) of the SF-MPQ-2 total and subscale scores (N = 48)

Variables	$\text{SEM}_{\text{agreement}}$	SEM (%)	$\text{MDC}_{90\ \text{individual}}$ (95% CI)	MDC (%)	$\text{MDC}_{90\ \text{group}}$
SF-MPQ-2 **Continuous**	0.8	7.8	1.8 (-1.6 – 2.0)	18.1	0.4
SF-MPQ-2 **Neuropathic**	0.8	7.8	1.8 (-1.7 – 1.9)	18.0	0.4
SF-MPQ-2 **Intermittent**	1.0	9.9	2.3 (-2.1 – 2.4)	22.9	0.5
SF-MPQ-2 **Affective**	0.7	7.3	1.7 (-1.5 – 1.8)	16.8	0.4
SF-MPQ-2 **Total**	0.5	5.1	1.2 (-1.0 – 1.4)	11.9	0.3

SF-MPQ-2, Revised Short McGill Pain Questionnaire Version-2; ***CI,*** Confidence Interval; ***SEM,*** Standard Error Measurement; **MDC,** *Minimal Detectable Change.*

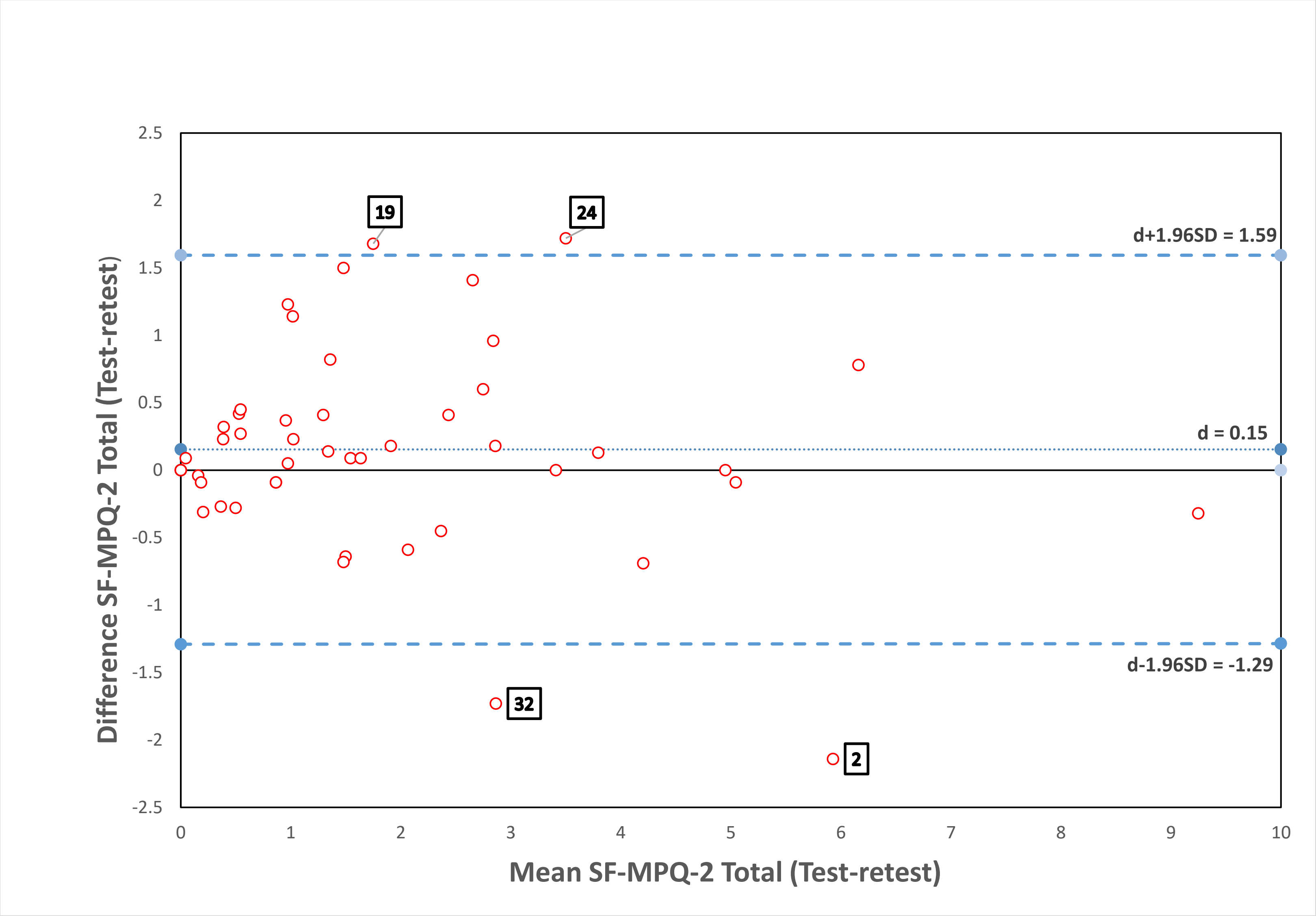
SEM (%) and MDC (%) is expressed as the proportion of the obtained $\text{SEM}_{\text{agreement}}$ or $\text{MDC}_{90\text{individual}}$ of domain represented on the SF-MPQ-2 to the total score of the scale (i.e. 10 points).

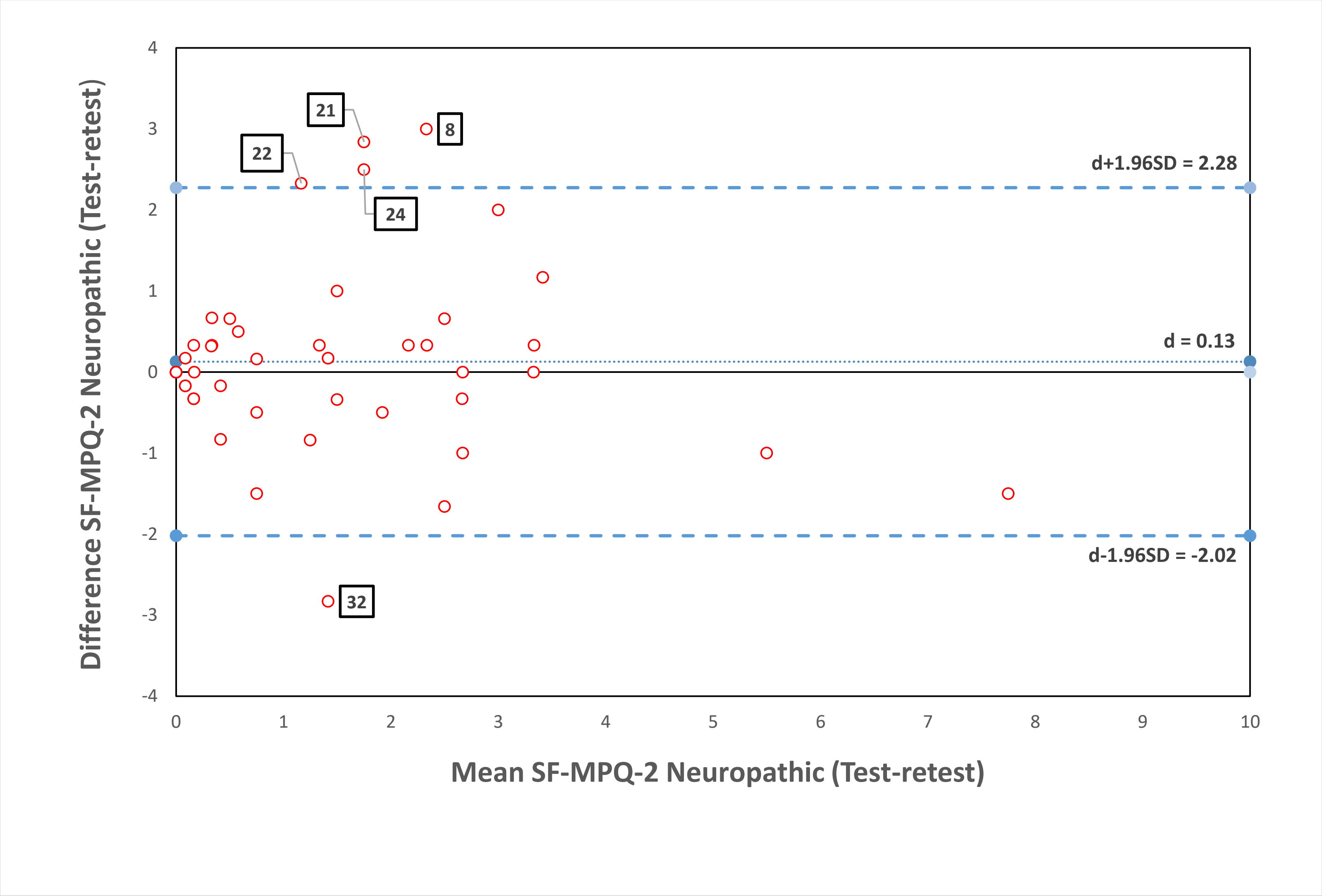
TABLE 6: Relative reliability of the SF-MPQ-2 total and subscale Scores (N = 48)

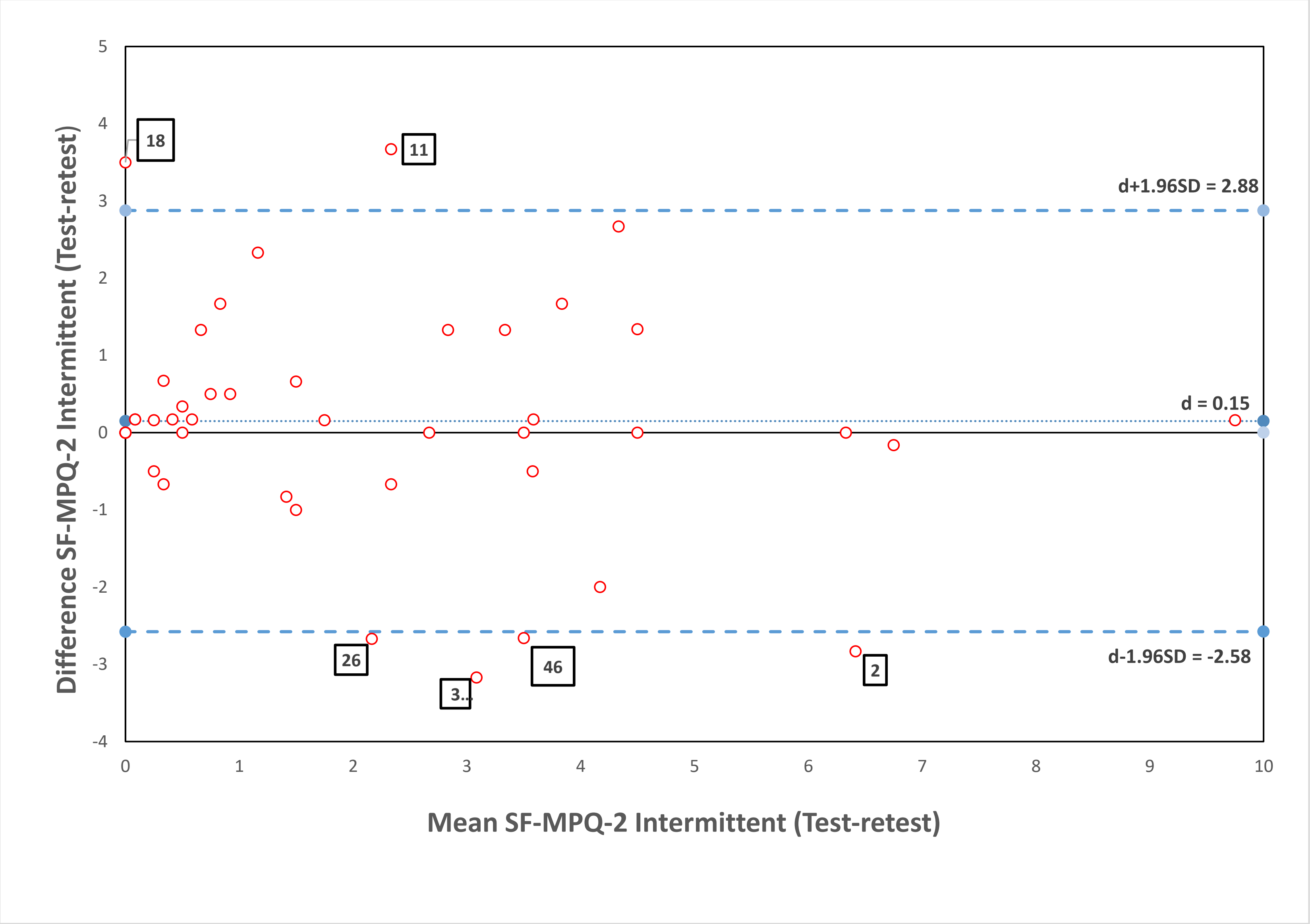
Variables	**Test-Retest Reliability**						
						Single measure ICC$_{2,1}$ (95% CI)	
	Test Mean (SD)	**Test Mean (SD)**	**d (SD)**	**95% CI of d**	**95% LOA**	**Transformed data**	**Non-transformed data**
SF-MPQ-2 **Continuous**	2.8 (2.6)	2.7 (2.6)	0.19 (1.12)	-0.14 – 0.51	-2.01, 2.38	[a]0.90 (0.83 – 0.94)	[a]0.91 (0.84 – 0.95)
SF-MPQ-2 **Intermittent**	2.1 (2.3)	2.0 (2.4)	0.15 (1.39)	-0.24 – 0.54	-2.58, 2.88	[a]0.82 (0.71 – 0.90)	[a]0.82 (0.71 – 0.90)
SF-MPQ-2 **Neuropathic**	1.5 (1.6)	1.3 (1.7)	0.13 (1.10)	-0.19 – 0.45	-2.02, 2.28	[a]0.78 (0.64 – 0.87)	[a]0.78 (0.64 – 0.87)
SF-MPQ-2 **Affective**	1.5 (1.9)	1.3 (2.0)	0.15 (1.01)	-0.14 – 0.45	-1.83, 2.14	[a]0.85 (0.75 – 0.92)	[a]0.87 (0.78 – 0.92)
SF-MPQ-2 **Total**	2.0 (1.9)	1.9 (2.0)	0.15 (0.73)	-0.06 – 0.37	-1.29, 1.59	[a]0.92 (0.86 – 0.96)	[a]0.93 (0.87 – 0.96)

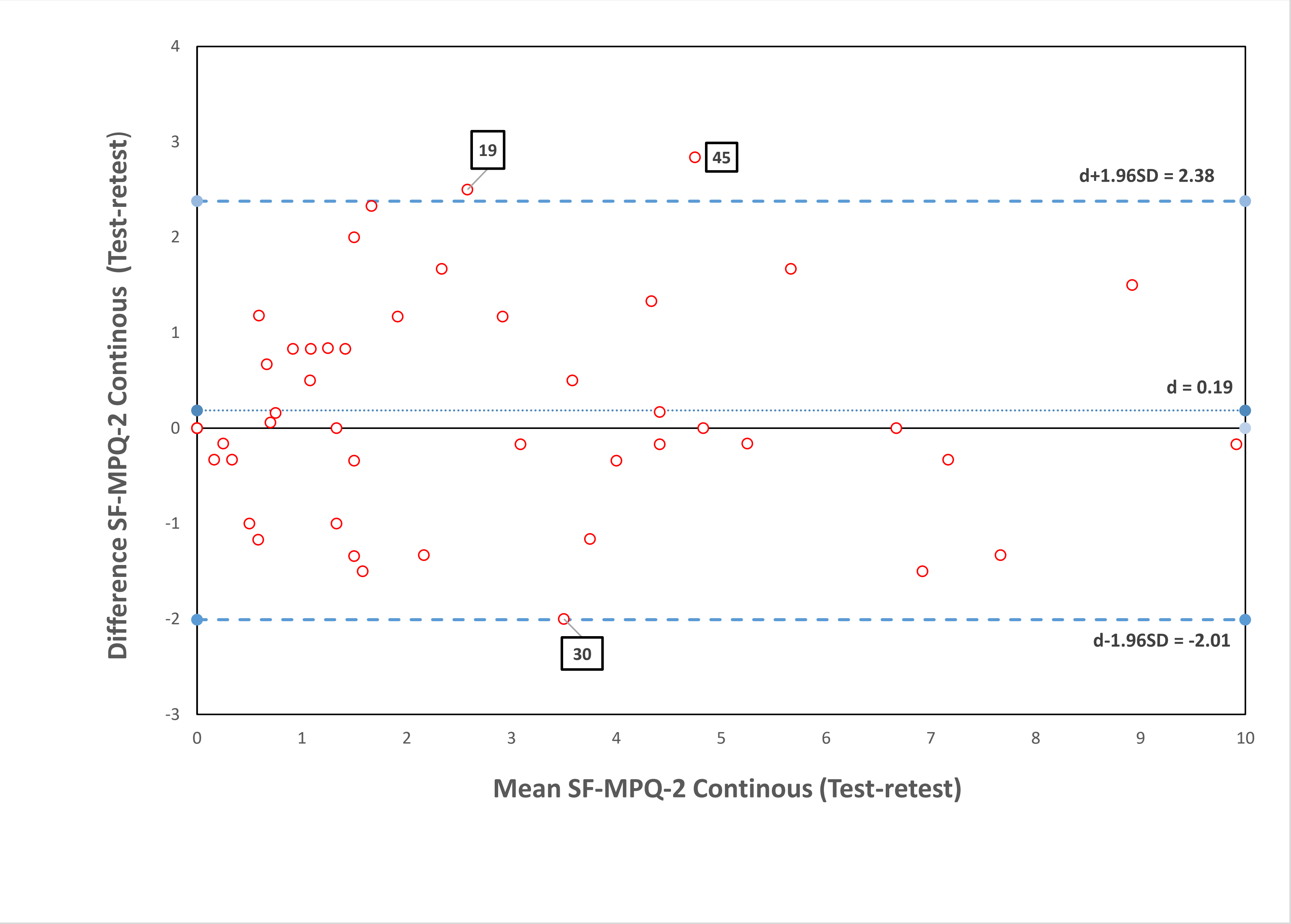
SF-MPQ-2, Revised Short-form McGill Pain Questionnaire Version-2; ***d***, Mean difference (test-retest); ***SD***, Standard deviation; ***CI***, Confidence interval; ***LOA***, Limits of Agreement; ***ICC***, Intraclass Correlation Coefficient.

[a] All correlation coefficient (r) were statistically significant at $p < 0.001$ (2-tailed).


19
24
32
2
d+1.96SD = 1.59
d = 0.15
d-1.96SD = -1.29
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
0
1
2
3
4
5
6
7
8
9
10
Difference SF-MPQ-2 Total (Test-retest)
Mean SF-MPQ-2 Total (Test-retest)


4
3
2
1
0
-1
-2
-3
-4
0
1
2
3
4
5
6
7
8
9
10
Difference SF-MPQ-2 Neuropathic (Test-retest)
Mean SF-MPQ-2 Neuropathic (Test-retest)
21
8
22
24
32
d+1.96SD = 2.28
d = 0.13
d-1.96SD = -2.02


Difference SF-MPQ-2 Intermittent (Test-retest)
Mean SF-MPQ-2 Intermittent (Test-retest)
d+1.96SD = 2.88
d = 0.15
d-1.96SD = -2.58
18
11
26
3..
46
2
5
4
3
2
1
0
-1
-2
-3
-4
0
1
2
3
4
5
6
7
8
9
10


4
3
2
1
0
-1
-2
-3
0
1
2
3
4
5
6
7
8
9
10
19
45
30
d+1.96SD = 2.38
d = 0.19
d-1.96SD = -2.01
Difference SF-MPQ-2 Continous (Test-retest)
Mean SF-MPQ-2 Continous (Test-retest)

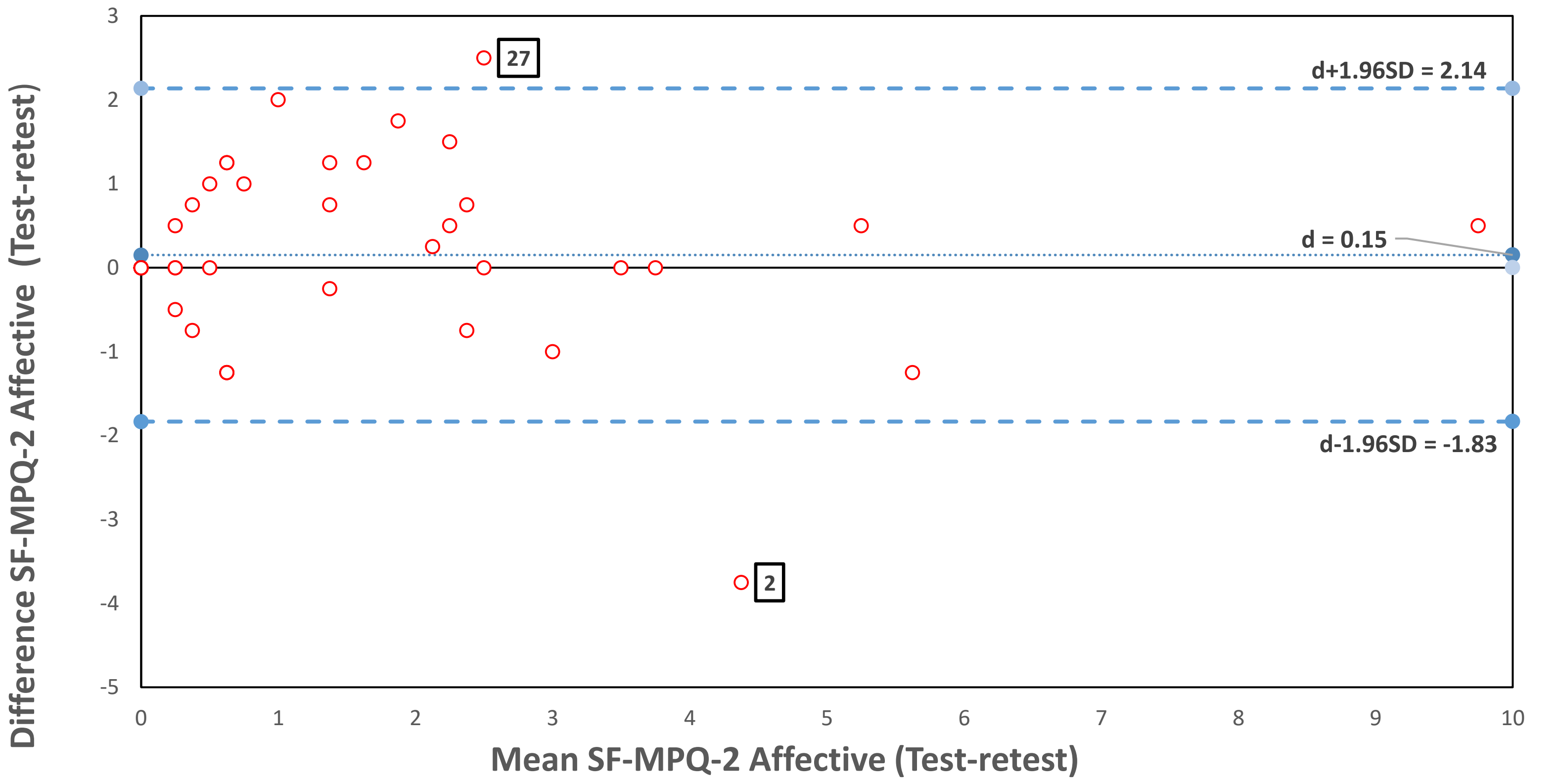


Figure 2 to 6 represent the Bland-Altman Limits of Agreement (LOA) plots between the test and retest scores of the SF-MPQ-2 Total (Fig 2), Neuropathic (Fig 3), Intermittent (Fig 4), Continuous (Fig 5) and Affective (Fig 6) subscale scores (n = 48). The difference between test-retest scores is plotted against the mean of test and retest scores for the respective SF-MPQ-2 total and subscales depicted. On each plot, the central blue line represents the mean of intra individual differences (d); the upper and lower horizontal broken lines represent the 95% LOA. The 95% LOA shows that 95% of the intra individual differences are supposed to within d ± 1.96 SD of mean difference (d). The outlier noted in each BA plot is numbered, according to participant #RS I.D, and present in accordance to the SF-MPQ-2 subscale or total they were noted.

CHAPTER 4:

DISCUSSION

Summary

In order to use a patient-reported outcome measure (PROM) in research or clinical practice, it is important to understand its measurement performance, cost and utility of the measure (1–4). The Brief Pain Inventory-Short Form (5,6) and Revised Short McGill Pain Questionnaire Version-2 (7) are general-use multidimensional tools recommended for use either independently or alongside other measures for comprehensive pain assessment in musculoskeletal conditions. This thesis examined existing measurement evidence supporting their use in pain-related musculoskeletal conditions. The first thesis manuscript (chapter 2) was a systematic review that addressed the quality and content of psychometric evidence supporting the BPI-SF and SF-MPQ-2 in MSK conditions. The review identified gaps in the literature (8,9) which informed our second thesis manuscript (chapter 3) study aim of determining the reproducibility of the SF-MPQ-2. A sample of adults with musculoskeletal shoulder pain were then recruited to complete the SF-MPQ-2 in two occasions for us to be able to examine reliability and agreement properties (chapter 3).

The systematic review study (chapter-2) examined the available measurement evidence reported for the BPI-SF and SF-MPQ-2 in mixed and specific MSK conditions. The search identified 25-articles addressing both tools properties in MSK conditions, however, more than half (17-articles) focused of the BPI-SF, perhaps, from its long-time presence in the literature (10). Because both tools are general-use PROMs (often applicable in any context for pain assessment), studies reporting psychometrics in mixed and specific MSK populations were included if MSK conditions represented ≥70% of the sample to enhance the generalizability of our findings. Despite our inclusion decisions, we were unable to locate studies examining psychometric properties for both tools in homogenous upper extremity conditions. The findings of our evidence synthesis, based on the COSMIN modified GRADE (11), suggest high-quality evidence supports both tools internal consistency and criterion-convergent validities in MSK populations. However, the BPI-SF displayed better quality evidence in support of its responsiveness, test-retest reliability, known group validity and structural validities over that of the SF-MPQ-2.

Our review identified three important gaps in the literature. First, studies investigating content validity, cross-cultural equivalence and MCID/CID were lacking for both tools. Second, evidence backing responsiveness and known group validity were mostly flawed (based on the COSMIN guidelines). The authors of the included studies did not provide hypotheses with specific directions and magnitudes of expected change. Third, the reliability assessment for both tools focused mainly on estimating intraclass correlation coefficients and Cronbach alpha with no effort towards defining agreement parameters. The first manuscript made important recommendations for future research and the second manuscript of this thesis addressed some of the gaps in the literature.

The second research manuscript investigated the reproducibility of the SF-MPQ-2 for use among patients with musculoskeletal shoulder pain. This second research manuscript addressed three important gaps in the literature: a) the current dearth in comprehensive evidence regarding the reproducibility properties of the SF-MQ-2 in MSK conditions; b) the absence of any measurement evidence backing the SF-MPQ-2 in upper extremity MSK conditions; c) established the reproducibility of the SF-MPQ-2 for use among patients with musculoskeletal shoulder pain. As a strength, a representative sample of patients with musculoskeletal shoulder pain was captured and a satisfactory retest interval (3-7days), as recommended in the literature, was used. Furthermore, the stability of patients' responses was supported with the concomitant administration of the Global Rating of Change scale. We used a wide range of statistical approaches to establish the reliability and agreement properties of the SF-MPQ-2 while adhering to existing guidelines (11–13). The second research study established acceptable internal consistency, relative reliability (ICC 2,1) and agreement parameters (SEM and MDC) for the SF-MPQ-2 use in musculoskeletal shoulder pain, and the Bland-Altman method (14,15) confirmed no evidence of systematic bias between retest occasions.

Strengths and Limitations of the Two Manuscripts

The main strength of the first manuscript was the rigorous steps taken to reach conclusions in the review. Two quality assessment processes [COSMIN (1,3,11,13) and MacDermid's methods (16)] were completed to reach conclusions on both tools performance in MSK conditions. Furthermore, we presented evidence distinctively for mixed and specific MSK population to ensure potential tool users have contextual information on how both tools performance in peculiar MSK conditions.

Two strengths of the second research study were our reliance on established guidelines (3,12) and the robustness of our reproducibility analysis. We adhered to established guideline instructions in choosing an appropriate retest interval (3-7 days) and the used the Global Rating of Change to determine if patients were in stable pain threshold (17). Also, our statistical analysis were detailed: we assessed both relative, cross-sectional, and absolute reliability properties and used the Bland-Altman method to determine reproducibility. We did this to ensure confidence in our findings, which influences potential users' choice for the SF-MPQ-2, hence encouraging its clinical applicability.

However, this thesis has some limitations. First, we compared both tools even though they examine slightly different dimensions of pain (BPI-SF = interference; SF-MPQ-2 = Quality). Second, making conclusions when using the COSMIN Modified GRADE does not equate to a study that directly compares both tools. Third, even though our review identified several gaps in the literature, we could only address several issues such as defining the reproducibility of the SF-MPQ-2 in MSK population. Therefore, future studies should focus on determining the SF-MPQ-2 validity, responsiveness and structural stability using Rasch modelling in upper extremity MSK conditions. Finally, our study participants came from a regional specialty clinic and the generalizability of our findings is not known.

Implications

This thesis has direct implications for research and clinical practice. First, our review will serve as a useful resource for potential users of the tools including guideline developers, researchers and clinicians to understand the quality, content and scope of measurement evidence backing the use of both tools in peculiar and mixed MSK population studies. Second, although we have shown that the BPI-SF has better psychometric properties than the SF-MPQ-2, we suggest that clinicians and researchers should consider the BPI-SF for use, if the qualities/characteristics of pain are not the primary aim of patients' assessment. Third, establishing the reproducibility of the SF-MPQ-2 for use among shoulder pain patients means researchers and clinicians can be confident that the SF-MPQ-2 yields dependable scores, and will be a useful tool for multidimensional pain assessment in shoulder pain conditions.

Future Direction

Our review identified substantial gaps in the literature ranging from methodological flaws, absence of evidence in upper extremity MSK populations, and the lack of assessment of some measurement properties. Going forward, studies of the psychometric properties of the BPI-SF and SF-MPQ-2 in MSK conditions should include:

a) Standard procedures for further establishing reliability that includes defining agreement parameters in mixed and specific MSK conditions, including upper extremity MSK conditions.

b) A comprehensive assessment of content validity in MSK conditions, bearing in mind that each tool captures slightly different concepts of pain (SF-MPQ-2= pain quality; BPI-SF = pain interference). The content analysis could include formal cognitive debriefing and ICF linking processes.

c) Assessment of responsiveness, minimal detectable differences, and clinically important differences for both tools in MSK conditions is important. In addition, a clear hypothesis with direction and expected magnitude should be provided and established anchor/external criterion should be utilized.

d) Determining the known group validity for both tools in MSK conditions is necessary to establish the usefulness of the tools. Future studies should employ appropriate statistical approaches including the use of ROC analysis, and only anchors/external criterions with established psychometric properties should be utilized in such assessment of known group validity.

e) Finally, a direct comparison of both tools in various contextual environments should be conducted.

In conclusion, this thesis adds to the existing pool of literature regarding the psychometric and agreement parameters of the Brief Pain Inventory-Short Form and Revised Short McGill Pain Questionnaire Version 2 in Musculoskeletal Conditions.

REFERENCES

1. Prinsen CAC, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a “Core Outcome Set” – a practical guideline. Trials. 2016 Dec;17(1):449.

2. Mokkink LB, Terwee CB, Stratford PW, Alonso J, Patrick DL, Riphagen I, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res. 2009 Apr;18(3):313–33.

3. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol [Internet]. 2007 Jan 1 [cited 2018 Aug 26];60(1):34–42. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S0895435606001740#bib41

4. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess (Rockv). 1998;2(14).

5. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singapore. 1994 Mar;23(2):129–38.

6. Cleeland CS. The Brief Pain Inventory User Guide [Internet]. 2008. Available from: https://www.mdanderson.org/documents/Departments-and-Divisions/Symptom-Research/BPI_UserGuide.pdf

7. Dworkin RH, Turk DC, Revicki DA, Harding G, Coyne KS, Peirce-Sandner S, et al. Development and initial validation of an expanded and revised version of the Short-form

McGill Pain Questionnaire (SF-MPQ-2). Pain [Internet]. 2009 Jul 1 [cited 2018 Aug 25];144(1):35–42. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S0304395909001250

8. Jumbo S, MacDermid JC, Michael K, Packham TL, Athwal GS, Faber K. Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review. Western University; 2019.

9. MacDermid J, Jumbo S, Kalu M, Packham T, Athwal G, Faber K. Measurement Properties of the Brief Pain Iinventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire Version-2 (SF-MPQ-2) in Pain-related Musculoskeletal Conditions: A Systematic Review. In: Abstracts Accepted for Publication [Internet]. BMJ Publishing Group Ltd and European League Against Rheumatism; 2019 [cited 2019 Jul 7]. p. 2128.1-2128. Available from: http://ard.bmj.com/lookup/doi/10.1136/annrheumdis-2019-eular.3525

10. Daut RL, Cleeland CS, Flanery RC. Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. Pain. 1983 Oct;17(2):197–210.

11. Mokkink LB, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, de Vet HCW, et al. COSMIN methodology for systematic reviews of Patient - Reported Outcome Measures (PROMs). 2018.

12. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud [Internet]. 2011 Jun 1 [cited 2019 Mar 7];48(6):661–71. Available from:

https://www.sciencedirect.com/science/article/pii/S0020748911000368

13. 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. Qual Life Res. 2018 May;27(5):1171–9.

14. Bland JM., Altman DG. Statiscal Methods for Assessing Agreement Between Two Methods of Clinical Measurement. Lancet [Internet]. 1986 Feb 8 [cited 2019 Mar 10];327(8476):307–10. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S0140673686908378

15. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res [Internet]. 1999 Jun 1 [cited 2019 Mar 10];8(2):135–60. Available from: http://journals.sagepub.com/doi/10.1177/096228029900800204

16. MacDermid JC, Law M, Michlovitz S. Outcome measurement in evidence-based rehabilitation. In: Law M, MacDermid JC, editors. Evidence-based rehabilitation: A guide to practice. 3rd ed. Thorofare NJ, USA: Slack Incorporated; 2014. p. 65–104.

17. Dewan N, MacDermid JC, MacIntyre N, Grewal R. Reproducibility: Reliability and agreement of short version of Western Ontario Rotator Cuff Index (Short-WORC) in patients with rotator cuff disorders. J Hand Ther [Internet]. 2016 Jul 1 [cited 2018 Aug 26];29(3):281–91. Available from: https://www-sciencedirect-com.proxy1.lib.uwo.ca/science/article/pii/S089411301500188X?via%3Dihub#bib47

CURRICULUM VITAE

Samuel Ugochukwu Jumbo BMR.PT, MSc. (C)

Profile

Discipline Trained In: Physiotherapy

Research Disciplines: Physiotherapy, Musculoskeletal Health, Geriatrics-Gerontology

Areas of Research: Rehabilitation, Aging Process, MSK-related upper extremity dysfunctions among older adults, Rehabilitation Care and Services for older adults, multidimensional pain assessment, outcome measurement

Fields of Application: Education, Biomedical Aspects of Human Health, Health System Management, Pathogenesis and Treatment of Diseases

Degrees

2017/9 - 2019/8	Master in Rehabilitation Science, Western University, London Ontario Degree Status: Completed Supervisor: Prof. Joy C. MacDermid
2005/3 - 2011/11	Bachelor in Medical Rehabilitation (Physiotherapy), University of Maiduguri, Borno State, Nigeria Degree Status: Completed Supervisor: Prof. Adetoyeje Y. Oyeyemi

Publications

1. Anieto EM, Nwankwo A, **Jumbo SU** & Kalu ME (2018) The effect of aerobic exercise on lipid profile of patients with HIV infection undergoing the highly active antiretroviral therapy (HAART): a protocol for a systematic review with meta-analysis. Journal of Exercise Rehabilitation.14(4):559-565. Doi: 10.12965/jer.1836258.129

2. Emoefe D, Adandom, I, **Jumbo S**, Nwankwo H, Obi CP & Kalu ME (2018) The Burden Experience for Formal & Informal Caregivers of Older Adults with hip fractures in Nigeria. SAGE Open Nursing. 4:1-10. DOI: 10.1177/2377960818785155

Non peer-reviewed publications

1. **Jumbo SU,** MacDermid JC, Kalu ME, Packham TL, Athwal GS, & Faber KJ. Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review protocol. PROSPERO 2018 CRD42018095862

2. Akinrolie O, Barclay B, Strachan S, Gupta A, Unyime J, **Jumbo S**, Askin N, Rabbani R, Zarychanski R, Abou-Setta A. Effectiveness of motivational interviewing on physical activity in older adults. PROSPERO 2019 CRD42019131174

Manuscripts submitted for peer review

1. Adandom I, **Jumbo S**, Emoefe D, Nwankwo H, Akinola B, & Kalu ME (January 2019) Healthcare professionals' experiences in identifying and managing psychosocial and cognitive factors during hip/knee fracture rehabilitation for older adults in Nigeria. An interpretive description study. *Submitted*. International Journal of Therapy and Rehabilitation

2. Kalu ME, Nwankwo H, Anieto E, Austin E, Adandom I, Emoefe D, **Jumbo, S**, Akinrole, O, Obi P, Nwankwo C, Ekezie U, Mohammad S, Ajulo M, Opara M, & Abaraogu U (March 2019) Physiotherapists' role during hospital-to-home transition for older adults with hip fracture and mobility limitation. A qualitative research protocol. *Submitted.* OBM Geriatrics

3. **Jumbo SU**, MacDermid JC, Kalu ME, Packham TL, Athwal GS, & Faber KJ (December 2018) Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review protocol. *Submitted*. Archive of bone and joint surgery

Presentations

1. **Jumbo SU,** MacDermid JC, Kalu ME, & Packham T, Athwal G, and Kenneth F. (2018) "Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review"- Presented at the European Congress of Rheumatology 2019 (12 - 15th June 2019) Madrid Spain. Presented at the World Confederation for Physical Therapy Annual Conference, Geneva Switzerland (10-13th May 2019). Presented at Ontario Physiotherapy Association conference Toronto, Canada (5-7th April 2019). Presented at the Bone & Joint Institute Conference @ Western University, London Ontario, Canada (11th -12th, May 2018)

2. Akinrolie O, Barclay B, Strachan S, Gupta A, Unyime J, **Jumbo S**, Askin N, Rabbani R, Zarychanski R, Abou-Setta A. "Effectiveness of motivational interviewing on physical activity among older adults: A systematic review and meta-analysis." Presented at the Centre for Ageing 36th Annual Spring Research Symposium/Workshop. 6-7th May 2019

3. **Jumbo SU**, Onyeke C, Kalu ME, Odumodu, IJ, Paul, C, MacDermid JC (2019): "The care experiences of veterans with chronic low back pain (CLBP) receiving physiotherapy care in a Nigerian Military Hospital"- a poster presentation at the Health and Rehabilitation Science Graduate Student Conference @ Western University, London, Ontario. February 6-7th 2019

4. Israel A**, Jumbo SU**, & Kalu ME (2018) **"**Researching the inclusion of psychosocial and cognitive factors during the management of hip/knee fractures among the older adults in Nigeria; A qualitative inquiry"- a poster presented at the 1st Annual conference of the Association of Clinical & Academic Physiotherapist of Nigeria, Coal City Enugu, Nigeria @ University of Nigeria, Enugu, Nigeria (15-10 Oct. 2018)

5. Akinrolie Y, **Jumbo SU**, Kalu ME (2018) "Knowledge about risk factors and practice about fall prevention in older adults among physiotherapists in Nigeria" presented at the 48th Annual Scientific & Education meeting of the Canadian Association of Gerontology Conference holding at Simeon Fraser University, (18-21st Oct. 2018)

6. Emoefe D, Adandom I, **Jumbo SU**, Obi CP, Nwankwo H & Kalu ME (2018) "Burden experience of formal & informal caregivers of older adults with hip fracture in southern Nigeria" an oral presentation at the Emerging Researchers in Ageing Conference @ The British Society of Gerontology Conference held @ the university of Manchester, Manchester UK (3-6th July 2018)

7. Kalu ME**, Jumbo SU**, & Osifeso T (2018) **"**Emerging evidence on Ageing research in Africa"- a symposium presentation at the Canadian Association of African Studies Annual Conference @ Queen's University, Kingston ON, Canada (3rd -5th May 2018)

8. **Jumbo SU,** Emoefe, D, Adandom, I, Kalu ME, & MacDermid JC (2018) Burden experience of formal & informal caregivers of older adults with hip fracture in southern Nigeria - a poster presentation at the Health and Rehabilitation Graduate Student Conference – Western University @ London, January 2018; and at London Health Research Trainee Conference. (May 10, 2018)

Published Conference Abstracts

1. Kalu ME, Emoefe D, Obi P, Ojembe B, **Jumbo SU** & Dal Bello Hass, V. (2018). Hospital-to-home discharge process of older adults with hip fractures back to their homes: Healthcare provider and informal caregiver experiences in Nigeria. Canadian Association of Gerontology, 47th Annual.
In Press

2. Kalu ME, **Jumbo SU** & Akinrolie Y. (2018). Knowledge about risk factors and practice about fall prevention in older adults among physiotherapists in Nigeria. Canadian Association of Gerontology, 47th Annual Conference
In Press

3. **Jumbo SU**, MacDermid JC, Kalu ME, & Packham T, Athwal G, and Kenneth F. (2018) "Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review"- World Confederation for Physical Therapy Annual Conference, 2019 Geneva Switzerland (10-13th May 2019).
https://www.abstractstosubmit.com/wcpt2019/archive/#/viewer/abstract/937

4. MacDermid J, **Jumbo SU**, Kalu M, Packham T, Athwal G, Faber K. Measurement Properties of the Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-Form McGill Pain Questionnaire Version-2 (SF-MPQ-2) in Pain-related Musculoskeletal Conditions: A Systematic Review. In: Abstracts Accepted for Publication. Vol 78. BMJ Publishing Group Ltd and European League Against Rheumatism; 2019:2128.1-2128. doi:10.1136/annrheumdis-2019-eular.3525.