SYSTEMATIC REVIEW

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 Protocol

Samuel U. Jumbo, BMR.PT, MSc[1]; Joy C. MacDermid, PT, PhD[1-3]; Michael E. Kalu, BMR.PT, MSc[2]; Tara L. Packham, OT, PhD[2]; George S. Athwal, MD, FRCSC[3]; Kenneth J. Faber, MD, MHPE, FRCSC[3]

Research performed at Department of Physiotherapy, Faculty of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada

Received: 11 December 2018 *Accepted: 19 January 2020*

Abstract

Background: The Brief Pain Inventory-Short Form (BPI-SF) and Revised Short-Form McGill Pain Questionnaire Version-2 (SF-MPQ-2) are generic pain assessment tools used in research and practice for pain assessment in musculoskeletal (MSK) conditions. A comprehensive review that systematically analyses their measurement properties in MSK conditions has not been performed. This review protocol describes the steps that will be taken to locate, critically appraise, compare and summarize clinical measurement research on the BPI-SF and SF-MPQ-2 in pain-related MSK conditions.

Methods: Medline, EMBASE, CINAHL and Scopus will be searched for publications that examine the measurement properties of the Brief Pain Inventory and Revised Short-Form McGill Pain Questionnaire Version-2. Two reviewers will independently screen citations (title, abstract and full text) and extract relevant data. The extensiveness, rigor, and quality of measurement property reports will be examined with a structured measurement studies appraisal tool, and with the updated COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines. Findings will be descriptively summarized, and when possible, a meta-analysis will be performed.

Discussion: This review will summarize and compare the current level of evidence on the measurement properties of the BPI-SF and SF-MPQ-2 in a spectrum of musculoskeletal conditions. We expect clinicians/researchers dealing with MSK conditions to have synthesized evidence that informs their decision making and preferences. In addition, the review hopes to identify gaps and determine priorities for future research with or on the BPI-SF and SF-MPQ-2 in MSK conditions.

Level of evidence: Not Applicable

Keywords: Brief pain inventory, McGill pain questionnaire, Musculoskeletal conditions, Patient reported outcomes, Psychometrics properties, Systematic review

Introduction

Musculoskeletal (MSK) conditions are a major cause of long-term pain and disability (1,2). Pain resulting from MSK conditions has a significant impact on patients general quality of life and is one of the most common reasons people seek medical attention (3). Although acute and ongoing pain

Corresponding Author: Samuel U. Jumbo, Department of Physiotherapy, Faculty of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada
Email: sjumbo@uwo.ca

THE ONLINE VERSION OF THIS ARTICLE ABJS.MUMS.AC.IR

have traditionally been hard to understand, patient-reported outcome measures (PROMs) have been useful for assessing and monitoring pain experiences (1,2). Generally, PROMs can be disease-specific or generic. Disease-specific PROMs were developed to assess the impact of a specific condition on the patient while generic PROMs are applicable to multiple conditions, and capture how illness impacts several constructs including pain and quality of life (4). The Brief Pain Inventory-Short Form (BPI-SF) and the Revised Short-form McGill Pain Questionnaire Version-2 (SF-MPQ-2), are examples of multidimensional pain assessment tools that were originally developed or tested in specific disease populations but now are commonly used for pain assessment in musculoskeletal conditions (5-7). While the BPI-SF captures the severity and interference of pain with daily functioning, the SF-MPQ-2 evaluates the qualities of neuropathic and nociceptive pain (6, 8).

Generic pain assessment PROMs are assumed useful for multiple conditions but were mostly developed from assessments conducted in specific conditions, hence a need to inspect the evidence backing their use in some context (9, 10). Although there is evidence of testing the measurement properties of BPI-SF and SF-MPQ-2 in different conditions other than the populations they were originally designed for, there are no studies that has systematically evaluated the performance of either tool in MSK conditions. Previous reviews that examined the BPI-SF and SF-MPQ-2 only summarized the PROMs' measurement properties in low back pain MSK conditions (11, 12). Hence, clinicians' and researchers' decisions on the use of either the BPI-SF or SF-MPQ-2 in other MSK conditions are based on the suggestions contained in single studies. Therefore, a systematic synthesis that includes a quality appraisal of evidence on the measurement properties of BPI-SF and SF-MPQ-2 will provide helpful information for decision making on the use of the tools when studying other MSK conditions (13–16).

This systematic review protocol is important because it serves as guide for conducting the review. The purpose of this comprehensive review protocol is to explain the steps that will be taken to systematically locate, summarize, critically appraise and compare 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.

Materials and Methods

Design

The protocol for this review has been registered with Prospero (CRD 42018095862). The review will be reported using the Preferred Reporting Items for Systematic Review and Meta-Analyses Protocol statement and checklist (17). In brief, the review will be conducted in four steps: (i) multiple bibliographic databases will be searched to identify relevant citations; (ii) title, abstract and full-text screenings will be conducted based on pre-determined eligibility requirements; (iii) all relevant data will be extracted; (iv) a structured measurement studies quality assessment tool and the COSMIN guidelines will be used to evaluate, synthesize and compare the quality of measurement research available on the BPI-SF and SF-MPQ-2 in MSK conditions before a meta-analysis will be pursued.

Data Source and Search Strategy

The following bibliographic databases will be searched to identify relevant citations: Medline—OVID (1946 to 2018); EMBASE—OVID (1947 to 2018); CINAHL—Cumulative Index to Nursing and Allied Health Literature (1937 to 2018); Scopus (1995 to 2018). Grey literature will not be searched. We have developed and piloted our search strategies in consultation with a health-research Librarian (see Appendix 1 for the detailed search expressions). The search strategies combine terms representing the name of both tools i.e. Brief Pain Inventory and McGill Pain Questionnaire, with terms representing the concepts of measurement properties including validity, reliability, and responsiveness. The terms for measurement properties were refined from the search-filter validated by Terwee et al. (18); the search strategy has been successfully piloted in the CINAHL database (see Table 1 for the piloted search results). All our searches will be adapted to fit the uniqueness of the four bibliographic databases. The search for BPI-SF articles will not be restricted by time or language; however, the search for SF-MPQ-2 will be restricted to its first publication date (i.e. January 1st, 2009) to avoid retrieving undesired articles on its parent versions (SF-MPQ and MPQ) (6).

Eligibility requirements

Inclusion Criteria

We will include studies that evaluated any of the measurement properties of the BPI-SF and SF-MPQ-2 in sample populations of adults (16 years of age or greater) where at least 70% of participants had MSK conditions. There will be no language restriction.

Exclusion Criteria

We will exclude: (a) all conference presentations (with no link to the full article), editorials including short clinical communications and commentaries, clinical trial protocols, and individual case reports; (b) studies that reported pain from other populations other than MSK populations. Examples are studies with unclear terms describing participants' pain, such as: 'non-malignant pain', 'chronic pain' and 'non-cancer pain', without relating such pain to be of MSK origin fully or partly; (c) studies that reported pain from individuals living with disability (e.g. congenital and developmental abnormalities); (d) studies that reported pain from individuals suffering from neoplasm, infections, surgical procedures not due to MSK conditions (e.g. coronary heart surgery, laparotomy), neurological or neuropathic conditions (not described as fibromyalgia or complex regional pain syndrome, or lumbar/cervical/thoracic radiculopathies coexisting in MSK conditions like neck pain, or back pain), and HIV AIDs pain.

Table 1. Result of pilot search obtained from CINAHL on BPI-SF and SF-MPQ-2

Search ID#	Search Terms	Search Options	Actions (results)
S1	(MM "McGill Pain Questionnaire") OR "mcgill pain questionnaire"	Boolean/Phrase	(1,575)
S2	(MH "Brief Pain Inventory") OR "brief pain inventory"	Boolean/Phrase	(1,119)
S3	(MM "Psychometrics") OR "psychometric properties"	Boolean/Phrase	(7,576)
S4	"measurement properties" OR (MM "Outcome Assessment") OR (MH "Measurement Error+")	Boolean/Phrase	(9,048)
S5	(MH "Instrument Validation") OR (MH "Validation Studies") OR "validation"	Boolean/Phrase	(65,782)
S6	"adaptation"	Boolean/Phrase	(30,104)
S7	"cross cultural"	Boolean/Phrase	(3,410)
S8	(MH "Reliability and Validity+") OR "reliability and validity" OR (MH "Predictive Validity") OR (MH "Internal Validity") OR (MH "Discriminant Validity") OR (MH "Criterion-Related Validity+") OR (MH "Consensual Validity") OR (MH "Concurrent Validity")	Boolean/Phrase	(149,777)
S9	(MM "Internal Consistency") OR "internal consistency"	Boolean/Phrase	(18,784)
S10	(MH "Sensitivity and Specificity") OR "sensitivity and specificity" OR (MH "ROC Curve")	Boolean/Phrase	(47,025)
S11	(MH "Discriminant Analysis") OR (MH "Discriminant Validity")	Boolean/Phrase	(4,634)
S12	(MH "Factor Analysis") OR "factor analysis"	Boolean/Phrase	(25,470)
S13	"clinically important difference"	Boolean/Phrase	(556)
S14	"minimal clinically important difference"	Boolean/Phrase	(323)
S15	(MH "Rasch Analysis") OR "rasch analysis"	Boolean/Phrase	(1,348)
S16	S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9 OR S10 OR S11 OR S12 OR S13 OR S14 OR S15	Boolean/Phrase	(241,727)
S17	S1 AND S16	Limiters - Published Date: 20090101-20181231	(133)
S18	S2 AND S16	Boolean/Phrase	(313)

S17 (n = 133) = Total citations identified for screening on SF-MPQ-2
S18 (n = 313) = Total citations identified for screening on BPI-SF

Study selection and screening

We will export the retrieved citations from each of the bibliographic database to Covidence systematic review management software (© 2017 Covidence) for de-duplication and independent study selection. Two review authors will independently conduct the title and full article screening based on the eligibility requirements. A hand-search of the reference list of articles emerging from the full-text review will be conducted. Disagreements between the screening authors will be arbitrated by a third reviewer to determine publication eligibility. Furthermore, authors of studies will be reached through a maximum of three e-mail attempts to clarify issues on the selection of a study when a decision is difficult to reach by the three review authors, perhaps due to the unclear description of study participants MSK conditions.

Data extraction

A structured data-extraction form will be developed from the guide available in the second review author's previous work [JCM] (15). Two review authors will work independently to extract data from the included studies with a third review author available to resolve disagreements.

Data items

We will extract data on several measurement properties including 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 (Area Under the Curve (AUC), change correlation indices, standardized response mean [SRM], effect-size [ES]), interpretability properties (clinically important difference [CID], minimal clinical important difference [MCID]), measurement error indices, measurement invariance/ cross-cultural validities. Data will be summarized according to the subscale of the tools evaluated, and the MSK populations they represent, as follows:

a) "Mixed" studies that satisfy the requirements of ≥70-percent sample size proportion representing MSK conditions, but of different mechanism or pathophysiology or described by different body regions.

b) "Specific" studies conducted among homogenous MSK samples described by the body region affected or

the pathophysiology/mechanism.

There is a tendency for some measurement properties to be evaluated in subgroups of a parent 'Mixed' sample. In such cases, we will extract subgroup reports as stand-alone studies and report them as representing the relevant MSK condition identified in the study. Our grouping is flexible and depends on the characteristics of the MSK sample encountered in the included studies.

Other information to be extracted includes: (a) the characteristics of the studies, (e.g. country, language, study design, study setting, sample size, the SF-MPQ-2 and BPI-SF version/subscale/item used), and (b) participants characteristics including age and sex. To avoid missing important details, as previously mentioned, extraction will be conducted in pairs, and the result will be compared. We will track articles using the standardized data extraction form.

Review Team's Hypotheses

As recommended by the COSMIN initiative, the following hypotheses have been developed to guide the quality assessment of measurement properties for this review (13, 14):

a) We expect correlations between the two questionnaires and other pain/health-related outcome tools to be 0.3 and above in magnitude. Correlations will be classified as low-to-moderate at 0.3-0.69, and *"high"* at 0.7 and above (15, 16).

b) We expect reports on AUC to be ≥ 0.7 to represent discrimination beyond chance alone.

c) We expect the correlation of the two questionnaires change scores and other pain/health-related outcome measures to be ≥ 0.3 rho in magnitude, to be considered significant.

We are, however, unable to define hypotheses to assess responsiveness based on the standardized response mean (SRM) or effect size (ES) because these are context-specific indices that depend on several factors including the interventions used in the studies. We expect authors to provide clear hypotheses that define the magnitude of expected change in their studies when these indices for responsiveness are reported.

Critical Appraisal of Included Studies

The critical appraisal will be conducted in two phases to check the adequacy and breadth of authors reports and to assess the risk of bias accompanying reported measurement properties. Both phases of appraisal are complementary and allow the clinician/researcher to gauge the quality of evidence supporting both tools in MSK conditions. The first phase addresses issues related to the sufficiency of measurement evidence reports. The second phase of appraisal examines the risk of bias inherent with reports because a study at high risk to bias presents undependable evidence which requires further investigation, regardless of the sufficiency of reports.

Phase 1: Quality of Measurement Property Reports

Two review authors will appraise the breadth and quality of reports supporting evidence contained in each article with the structured clinical measurement appraisal tool (see Table 2 and 3 for the tool's detailed description) (19). The tool has previously been shown to be highly reliable in evaluating the quality of clinical measurement studies, including musculoskeletal outcome measures (16, 20). The tool evaluates the extent to which a study complies with the following criteria: 1) thorough literature review to define the research question; 2) specific inclusion/exclusion criteria; 3) specific hypotheses; 4) appropriate scope of psychometric properties; 5) sample size; 6) follow-up; 7) the authors referenced specific procedures for administration, scoring, and interpretation of procedures; 8) measurement techniques were standardized; 9) data were presented for each hypothesis; 10) appropriate

Use this form to rate the quality of a clinical measurement study. To decide which score to provide for each item on your quality checklist, pick the descriptor that sounds most like what was reported in the study you are evaluating. Items rank descriptors are provided in the guide. (Forms and guides to extract study data for evidence synthesis are available from developer at macderj@mcmaster.ca)

Table 2. Quality Appraisal for Clinical Measurement Research Reports Evaluation Form

Evaluation criteria	Score		
Study question	2	1	0
1. Was the relevant background work cited to define what is currently known about the measurement properties of measures under study, and the potential contributions of the current research question to informing that knowledge base?			
Study Design			
2. Were appropriate inclusion/exclusion criteria defined?			
3. Were specific clinical measurement questions/hypotheses identified?			
4. Was an appropriate scope of measurement properties considered?			
5. Was an appropriate sample size used?			
6. Was appropriate retention/follow-up obtained? (for studies involving retesting; otherwise n/a)			

Table 2. Continued

Measurements
7. Were specific descriptions provided of the measure under study and the method(s) used to administer it?
8. Were standardized procedures used to administer all study measures in a manner that minimized potential sources of error/ bias (including the study measure and its comparators)?
Analyses
9. Were analyses conducted for each specific hypothesis or purpose?
10. Were appropriate statistical tests performed to obtain point estimates of the measurement properties?
11. Were appropriate ancillary analyses done to quantify the confidence in the estimates of the clinical measurement property (Precision/Confidence intervals; benchmark comparisons/ROC curves, alternate forms of analysis like SEM/MID, etc.)?
Recommendations
12. Were clear, specific and accurate conclusions made about the clinical measurement properties; that were associated with appropriate clinical measurement recommendations and supported by the study objectives, analysis and results?
Subtotals (of columns 1 and 2)
Total score (sum of subtotals/24*100); if for a specific paper or topic an item is deemed inappropriate then you can sum of items/2*number of items *100

To decide which score to provide for each item on your quality checklist, read the following descriptors. Pick the descriptor that sounds *most* like the study you were evaluating with respect to a given item. If there is no documentation about any specific aspect of an item; then you must evaluate assuming that it was not done. Given the diversity in clinical measurement properties and design options, the evaluator has to make judgments using the criteria below and extend the principles to specific aspects that may not be covered in these brief exemplars. In many cases, the study will not look exactly like the descriptor so there will be some interpretation as to which level of optimal methods for clinical measurement studies have been achieved. In such cases, the evaluator can use the general approach that if this study research design and conduct is consistent with best practice (score=2); is acceptable but suboptimal (score=1); is not done/documented, substantially inadequate or inappropriate (score=0).

Table 3. Quality Appraisal of a Clinical Measurement Study Interpretation Guide

		Descriptors
Study question		
Score		
1	2	The authors: - performed a thorough literature review indicating what is currently known, and not known, about the clinical measurement properties of the instruments or tests under study - presented a critical, and unbiased view of what is known about the current measurement properties - indicated how the current research question fills a gap in the current knowledge base - established a research question based on the above.
	1	All of the above criteria were not fulfilled, but a sound rationale was provided for the research question.
	0	A foundation for the current research question was not clear; and the rationale was not founded on previous literature.
Study design		
2	2	Specific inclusion/exclusion criteria for the study were defined, that described the patients enrolled. The subjects were described in terms of health condition/demographics, key relevant outcome mediators and the recruitment context (setting).
	1	Some information on participants and place is provided (not all of above). For example, age/sex/diagnosis and the name or type of the practice is listed; but no additional information.
	0	No information on type of clinical settings or study participants is provided (other than number/mean age).

Table 3. Continued

3	2	Specific hypotheses or research questions are provided. The stated study purpose provides specific research questions or hypotheses that indicate which specific measurement properties will be evaluated. This should include the specific type of reliability (intra/inter-rater or test-retest) being tested or the type of validity (construct/criterion/content; longitudinal/concurrent; convergent/divergent) being tested. A prior hypothesis should describe the level of reliability expected; and for validity, expected relationships (strength of associations) or constructs.
	1	The types of reliability and validity being tested were apparent in the methods/title, but clear and specific research questions or hypotheses were not specified.
	0	Specific types of reliability or validity under evaluation were not clearly defined nor were specific hypotheses on reliability and validity stated. ("*The purpose of this study was to investigate the reliability and validity of*..." can be rated as zero if no further detail on the types of reliability and validity or the nature of specific hypotheses is stated).
4	2	An appropriate scope of clinical measurement properties would be indicated by 1. A detailed focus on reliability that included multiple forms of reliability (at least two of – intra-rater, inter-rater, test retest); as well as both relative and absolute reliability (e.g., ICCs and SEM/MID or limits of agreement) 2. A detailed focus on validity that included multiple forms of validity (content (judgmental); structured (e.g., expert review/survey, qualitative interviews, ICF linking) or structural (e.g., factor analyses or Rasch), construct (known group differences; convergent/divergent associations), criterion (concurrent/predictive), responsiveness; predictive, evaluative or discriminative properties were established 3. Three or more indicators of reliability and validity were examined concurrently and provide a rich view on measurement properties.
	1	Two or more clinical measurement properties were evaluated, however, scope was narrow and did not meet above criteria. (e.g., internal consistency and one other indicator of validity or reliability).
	0	The scope of clinical measurement properties was very narrow as indicated by a narrow evaluation of only one form of reliability or validity.
5	2	Authors performed a sample size calculation and obtained their recruitment targets. Post-doc power analyses and/or confidence intervals confirm that the sample size was sufficient to define relatively precise estimates of reliability or validity.
	1	The authors provide an acceptable rationale for the number of subjects included in the study, but did not present specific sample size calculations or post-doc power analyses (or had a sample >100 but no justification).
	0	Size of the sample was not rationalized or is clearly underpowered.
6	2	90% or more of the patients enrolled for study were re-evaluated.
	1	70% or more of the enrolled patients were re-evaluated.
	0	Less than 70% of the patients enrolled in the study were re-evaluated
Measurements		
7	2	Documentation is provided for how the studied test is performed. This includes adequate description of the measure/test and how it is administered or scored. The authors may provide or reference a published manual/article that outlines specific procedures for administration, scoring (including scoring algorithms, handling of missing data) and interpretation that included any necessary information about positioning/active participation of the client, any special equipment required, calibration of equipment if necessary, training required, cost, examiner procedures/actions. If no manual is available, then the text describes key details of procedures in sufficient detail so they could be replicated.
	1	The test(s) and its administration procedures are referenced; but there is inadequate description of the test procedures.
	0	Minimal description of test procedures without appropriate references.

Table 3. Continued

8	2	This item addresses the overall study procedures for administering all study measures (study measure and its comparators) in an unbiased way. Test procedures should not introduce systematic errors in the estimation of the clinical measurement properties. This includes standardized procedures for who completed or administered the measures. For self-report, this includes order of presentation, who completed at what time interval; handling of missing items. If relevant, then the paper should include how cultural literacy issues were handled (e.g., exclusion, assisted or surrogate completion). For impairment measures, procedures would include calibration of any equipment; use of consistent measurement tools and scoring, a priori exclusion of any participants likely to give invalid results/unable to complete testing (not exclusion of after enrollment); use of standardized instructions and test procedures. This can include order of administration of test and quality checking of scores. For reliability testing, the appropriate retest interval will depend on the nature of the condition; but for acute conditions it may require retesting within 48 hours; whereas chronic/stable conditions are commonly retested within 4-14 days. For estimation of clinical change, retest intervals should be ones during which a meaningful clinical change would have occurred (and from an intervention with known effectiveness). The evaluator decides overall whether this has sufficiently been addressed by the methods described.
	1	No obvious sources of bias in the study test protocol or how tests were performed/administered is apparent; but there were suboptimal procedures or an inadequate description of the measurement protocol to be insured control of bias or that procedures were standardized.
	0	No description of the overall procedures for administering study tests; OR an obvious source of bias in data collection methods.
Analyses		
9	2	Authors clearly defined which specific analyses were conducted for each of the stated specific hypotheses/questions of the study. This may be accomplished through organization of the results under specific subheadings or by demarcating which analyses addressed specific clinical measurement properties. Data was presented for each hypothesis/research question posed.
	1	Data was presented that addressed each of the measurement questions posed, but authors did not link specific analyses to specific research questions or hypotheses.
	0	Data was not presented for every hypothesis or clinical measurement property outlined in the purposes or methods.
10	2	Tests selected - Appropriate statistical tests were conducted to calculate a point estimate for clinical measurement properties. Examples are provided below; but are not exhaustive. 1. Reliability (Relative=ICCs for quantitative, Kappa for nominal data); absolute (SEM or plot of score differences vs. average score showing mean and 2 SD limit – as per Altman and Bland) 2. Clinical relevance - minimal detectable change, clinically important difference 3. Validity a. Validity associations - Pearson correlations for normally distributed data, Spearman rank correlations for ordinal data; or other correlations, if appropriate b. Validity tests of significant difference - an appropriate global test like analysis of variance was used where indicated, with post-hoc tests that adjusted for multiple testing c. Validity of items scaling/responses - Rasch analysis or item response 4. Responsiveness- standardized response means or effect sizes or other recognized responsiveness indices were used.
	1	Appropriate statistical tests were used in some instances; but suboptimal choices were made in other analyses.
	0	Inappropriate use of statistical tests - incorrect tests for type of data; or a lack of analysis
11	2	The study goes beyond a single statistical point estimate of a clinical measurement property and providing supporting statistical analyses that increases confidence in the findings in terms of precision of the (key) indicator; or provide an alternate form of analysis of the clinical measurement property. The evaluator decides if these analyses are appropriate and informative. For example, with reliability, at least 2 of the following would constitute appropriate and informative analysis beyond a point estimate of a reliability coefficient: 1. confidence intervals around the point estimate; 2. Comparison to appropriate, referenced benchmarks or standards; or 3. SEM or MDC. For correlations, tests of significance or confidence intervals were presented and indicators of the criterion benchmarks were provided. For studies involving cross-cultural validation, the analyses should compare multiple clinical measurement properties previously established for the measure and explain the extent to which the translated version is in accordance with these previously reported properties on the source measure.
	1	Either precision definition (confidence intervals) or appropriate benchmark comparison were used - NOT both. OR Some analyses were associated with indicators of precision or alternate form of analysis -but not all key indicators.
	0	Inappropriate use of benchmarks or confidence intervals; or indicators of precision or alternate form are absent

Table 3. Continued

Recommendations		
12	2	Authors made specific conclusions and clinical measurement recommendations that were clearly related to each hypotheses/ question posed in the study and that were supported by the data presented. Ideal recommendations would state the estimated status of the clinical measurement property, the confidence in the estimate and the context for which those apply. To achieve a 2, the conclusion must be specific; and conclusions cannot overstate the clinical measurement properties observed the study; nor ignore suboptimal measurement properties found.
	1	Authors made conclusions and clinical measurement recommendations that were basically true (supported by study data); but vague. That is, they do not specify the extent, confidence or context of the findings. (The measure is "reliable and valid ") OR authors made specific clinical measurement recommendations; but for only some of the study hypotheses.
	0	Authors did not make conclusions about clinical measurement; OR made recommendations that were in contradiction to the actual data presented

statistics-point estimates; 11) appropriate statistical precision-estimates; and 12) valid conclusions and recommendations. Each of the 12 criteria are graded using the following scheme: 0-point, if judged, 'not done/documented' OR 'substantially inadequate' OR 'inappropriate'; 1-point, if judged, 'acceptable but suboptimal'; and 2-points, if judged, 'consistent with best practice'. Any criteria deemed inapplicable to an article, is scored 'NA'. The overall quality rating of an article is calculated by dividing the summed score of the evaluated criteria by the number of evaluated criteria in the article. The highest and lowest obtainable total quality score for an article, if all 12 criteria is assessed, is 24-points and 0-points respectively. An article's total quality score can be converted to a percentage, which allows equilibrium across articles quality ratings; particularly, when some of the 12 criteria are deemed 'not applicable' to an article. In this review, the quality percentage score of articles between 0%–30% will be marked as Poor, 31%–50% as Fair, 51%–70% as Good, 71%–90% as Very Good, and > 90% as Excellent. The second review author (JCM), who is also the author of the structural critical appraisal tool, will be contacted to resolve any disagreements in that phase of the critical appraisal.

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

The risk of bias and the quality of measurement properties accompanying individual study's report on measurement properties will be assessed using the COSMIN risk of bias and quality criteria checklists (14, 21-23). Regardless of individual authors definition of measurement properties, we will define all measurement properties according to the COSMIN consensus-based taxonomy of measurement properties (24). First, the articles will be assessed for risk of bias using the updated COSMIN risk of bias checklist, which has 10 boxes that contains several items/questions scored on a 4-point rating scale as 'very good', 'adequate', 'doubtful' and 'inadequate'(21). The lowest rating for any item/question on a study's measurement property determines its overall rating for methodological risk of bias. We will exclude two boxes (PROM development and content validity) of the ten boxes in the COSMIN risk of bias checklist because the BPI-SF and SF-MPQ-2 were not originally developed for validation in MSK conditions (5, 6, 13).

Second, the quality of the articles will be assessed using the COSMIN quality criteria rating. All the extracted data on measurement properties for each of the tools in the included articles will be rated against the quality standard for measurement properties described in the COSMIN manual (also available at: https://www.cosmin.nl/) (14, 22, 23). Each measurement property (e.g. internal consistency, test-retest reliability) will be rated as: "Sufficient" (+), if within the benchmark quality criteria; "Indeterminate" (?), if there was lacking report to contrast with the benchmark quality criteria, and; "Insufficient" (-), if the reported measurement property was below the benchmark quality criteria.

Planned Method of Analysis

The pooled mean values on statistical measurement properties evaluated in homogenous MSK populations using similar methodologies will be calculated. Results of correlations, Cronbach alphas (α), intraclass correlation coefficients (ICC), standardized response mean (SRM), standard error of measurement (SEM), effect sizes (ES), will be summarized using meta-analysis. Forest plots will be used to estimate the level of heterogeneity. A sensitivity analysis will be performed on the measurement reports to compare the findings from studies with low risk of bias ('very good' to 'adequate' quality rating) with those with high risk of bias ('doubtful' to 'inadequate' quality ratings). We will then weigh the meta-analysis according to the study sample sizes.

Studies with sample population or methodology

heterogeneity will be assessed using narrative synthesis; tables will be used to summarize their findings. Our descriptions will include the MSK conditions evaluated with the tools (Mixed or Specific), the methodologies employed to assess both tools' measurement properties, the findings on the measurement properties and the quality of the studies. In our synthesis, more attention will be given to the consistency of reports contained in studies rated sufficient (+), with low risk of bias ('very good' or 'adequate' quality rating). First, we will pool/summarize the results extracted for the measurement properties per tool with emphasis placed on studies with low risk of bias ('very good' and/or 'adequate' quality ratings). The estimated intraclass correlation coefficient (ICC) for the BPI-SF and SF-MPQ-2 will be summarized using a weighted average. Cronbach's alpha for internal consistency will be summarized using the observed range of occurrences while proportions of correlations supporting criterion-convergence within the 'low-to-moderate' range (rho = 0.3-0.69), and 'high' range (rho ≥ 0.7) will be noted. For structural validity, responsiveness and known group validity, narrative synthesis will be used to summarize findings on both tools' measurement properties.

Evidence Synthesis

We will apply the COSMIN Modified GRADE approach, as described in the updated COSMIN manual to determine the level of evidence supporting the two questionnaires measurement properties (13). The COSMIN Modified GRADE approach considers the risk of bias, inconsistency, imprecision and indirectness associated with the pooled findings on each tool's measurement properties. We will apply the COSMIN modified GRADE approach under the assumption that the MSK conditions studied holistically represent other MSK conditions, irrespective of their type (mixed or specific) (13, 21). Two reviews authors will independently conduct the quality assessment and evidence synthesis and then meet to synthesize their findings. If misunderstandings arise from their meeting, a third review author will be contacted.

Discussion

Musculoskeletal conditions encompasses a broad range of problems affecting the muscles, bones, soft tissue, and joints. Although 291 identified MSK conditions have been defined, and each present with a unique pain experience, generic pain assessment tools are sometimes assumed useful for pain assessment in these conditions (2). The BPI-SF and SF-MPI-2 are examples of commonly used multidimensional PROMs for pain evaluation in different types of MSK conditions in the clinical and research settings. However, a review that pools and translates measurement research findings on both PROMs psychometric properties has not been conducted. This gap in the literature leaves clinicians/researchers with no choice than to make selection decisions based on their personal observation of face validity/appearance, ready availability/access to the tools, word of mouth/recommendation of colleagues, their observed use in other practice setting, and the suggestions contained in single studies. A systematic synthesis of a group of single studies would provide information on the measurement properties of BPI-SF and SF-MPQ-2 in the broad-scope of MSK conditions; and this would provide a more reliable and evidence-based rationale for selecting and using these tools in the clinic and research setting.

In summary, our review protocol focuses on understanding the available evidence of the measurement properties of the BPI-SF and SF-MPQ-2 in painful MSK conditions. The review will employ two comprehensive critical appraisal processes and yield synthesized evidence that will guide the choice of clinicians and researchers evaluating MSK conditions with the BPI-SF and SF-MPQ-2 in MSK conditions. Furthermore, our findings hope to identify the strengths and/or weaknesses of both tools and we hope to offer recommendations for future research either with or on the tools.

Samuel U. Jumbo BMR.PT MSc[1]
Joy C. MacDermid PT PhD[1-3]
Michael E. Kalu BMR.PT MSc[2]
Tara L. Packham OT PhD[2]
George S. Athwal MD FRCSC[3]
Kenneth J. Faber MD MHPE FRCSC[3]
1 Department of Physiotherapy, Faculty of Health and Rehabilitation Sciences, Western University, London, Ontario, Canada
2 School of Rehabilitation Science, McMasters University, Hamilton, Ontario, Canada
3 Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, Ontario, Canada

References

1. Hoy D, March L, Brooks P, Blyth F, Woolf A, Bain C, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis. 2014; 73(6):968-74.
2. Global Burden of Disease Study 2013 Collaborators. 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.
3. Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, et al. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain. 2008; 9(2):105-21.
4. Black N. Patient reported outcome measures could help transform healthcare. BMJ. 2013; 346(1):f167.
5. Cleeland CS, Ryan KM. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singapore. 1994; 23(2):129-38.
6. 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. 2009; 144(1):35-42.
7. 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.
8. Cleeland CS. The brief pain inventory user guide. Houston, TX: The University of Texas MD Anderson Cancer Center; 2009.
9. Chen TH, Li L, Kochen MM. A systematic review: how to choose appropriate health-related quality of life (HRQOL) measures in routine general practice? J Zhejiang Univ Sci. 2005; 6(9):936-40.
10. Robinson-Papp J, George MC, Dorfman D, Simpson DM. Barriers to chronic pain measurement: a qualitative study of patient perspectives. Pain Med. 2015; 16(7):1256-64.
11. Chapman JR, Norvell DC, Hermsmeyer JT, Bransford RJ, DeVine J, McGirt MJ, et al. Evaluating common outcomes for measuring treatment success for chronic low back pain. Spine (Phila Pa 1976). 2011; 36(21 Suppl):S54-68.
12. Ramasamy A, Martin ML, Blum SI, Liedgens H, Argoff C, Freynhagen R, et al. Assessment of patient-reported outcome instruments to assess chronic low back pain. Pain Med. 2017; 18(6):1098-110.
13. Mokkink LB, Prinsen C, Patrick DL, Alonso J, Bouter LM, de Vet HC, et al. COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). User Manual. 2018; 78(1):6-63.
14. Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018; 27(5):1147-57.
15. MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, et al. Measurement properties of the neck disability index: a systematic review. J Orthop Sports Phys Ther. 2009; 39(5):400-17.
16. 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; 34(13):1059-69.
17. 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.
18. Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009; 18(8):1115-23.
19. Law MC, MacDermid J. Evidence-based rehabilitation: a guide to practice. 2nd ed. New Jersey: Slack Incooperated; 2008.
20. 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):289-98.
21. Mokkink LB, de Vet HC, Prinsen CA, 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; 27(5):1171-9.
22. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007; 60(1):34-42.
23. Prinsen CA, 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; 17(1):449.
24. 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; 63(7):737-45.

APPENDIX 1.

Piloted search concepts, each for the BPI-SF and SF-MPQ-2, to be adapted on the following bibliographic databases: Medline CINAHL EMBASE and Scopus.

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)