

RESEARCH

Open Access



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^{1*} , Joy C. MacDermid^{1,2,3}, Tara L. Packham², George S. Athwal³ and Kenneth J. Faber³

Abstract

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 (MSK) health conditions. This study aimed to describe the reproducibility (test–retest 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 from MSK sources completed the SF-MPQ-2: at baseline ($n = 195$), and a subset did so again after 3–7 days ($n = 48$), if their response to the Global Rating of Change (GROC) scale remained unchanged. Cronbach alpha (α) and intraclass correlation coefficient ($ICC_{2,1}$), and their related 95% CI were calculated. 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 scale (0.95) and continuous subscale (0.92) scores; the remaining subscales displayed good $ICC_{2,1}$ scores (0.78–0.88). Bland–Altman analysis revealed no systematic bias between the test and retest scores (mean difference = 0.13–0.19). While the best agreement coefficients were seen on the total scale (SEM = 0.5; MDC_{90individual} = 1.2 and MDC_{90group} = 0.3), they were acceptable for the SF-MPQ-2 subscales (SEM: range 0.7–1; MDC_{90individual}: 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

Background

Shoulder disorders are among the three leading causes of musculoskeletal (MSK) pain, third only to neck pain and low back pain [1, 2]. The prevalence of shoulder disorders increases with aging [3, 4]. Shoulder disorders are associated with substantial consequences for the socioeconomic wellbeing of the patient and society; studies have

*Correspondence: sjumbo@uwo.ca

¹ Faculty of Health and Rehabilitation Sciences, Elborn College, Western University, London, ON, Canada

Full list of author information is available at the end of the article



© The Author(s) 2020. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

linked workers' absenteeism, job loss, and poor health-related quality of life (HRQoL) to symptoms associated with shoulder disorders [3, 5–8].

Pain assessment in clinical practice and research often places emphasis on monitoring pain intensity, even though pain is known to be multidimensional and experienced uniquely by individuals [9]. Patients perceive pain across six diverse dimensions: physiologic, sensory, affective, cognitive, behavioral and socio-cultural [9, 10]. The comprehensive assessment and monitoring of these dimensions should improve patient care [11]. A multidimensional pain assessment tool that provides a holistic assessment of pain has been recommended by experts [12–14] 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 et al. [15] created the SF-MPQ-2 by adding seven new items that explicitly examines neuropathic and non-neuropathic pain characteristics to the original 15-item Short McGill Pain Questionnaire (SF-MPQ). They also replaced the previous 4-point descriptive rating scale with a 10-item numerical rating scale to enhance its responsiveness [15]. Since then, multiple studies have utilized the SF-MPQ-2 as a primary outcome for pain assessment in clinical trials; its measurement properties have been examined in different populations including cancer pain [16], surgical pain [17], visceral pain [18], and neuropathic pain [19]. Among MSK conditions, studies have reported measurement evidence for patients with complex regional pain syndrome [20], back pain [21], knee osteoarthritis (OA) [22], and mixed MSK populations [23, 24]. Although the SF-MPQ-2 is becoming increasingly popular, our recent review [25–27] reported on evidence with design flaws including inadequate description of Intraclass correlation coefficient (ICC) models, insufficient justification of retest interval, and a 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 disorders.

Methods

This study was based on a cross-sectional study of internal consistency and test–retest reliability. The SF-MPQ-2 questionnaire was administered to examine reproducibility (test–retest reliability and agreement) and internal consistency at two time points: at baseline and after

3–7 days (when patients would, for the most part, be stable) [28, 29]. The participants were recruited from the Roth-McFarlane Hand and Upper Limb Centre (HULC), London, Ontario, Canada during a period of 6-months (June to November 2018). Ethics approval for a clinical database of routine outcome measures from which this data were extracted was approved by the University of Western Ontario Research Ethics Board (REB# 4986).

Patients

Adults proficient in English, above 18 years of age, that experienced pain from one or more shoulder conditions of known MSK source (for example: rotator cuff tear or tendinopathy, adhesive capsulitis, glenohumeral anterior instability, and superior labral anterior–posterior (SLAP) lesions) 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 provide consent.

Procedure

Assessors (SJ and HULC research assistants) identified eligible participants by reviewing the outpatient appointment list of patients scheduled for a clinical visit with two shoulder surgeons (KF and GA), a day prior. Potential participants were then contacted on the day of their clinical appointment and screened to ensure all criteria were satisfied; they were provided with an explanation of the objectives of the study before a questionnaire booklet containing the SF-MPQ-2 and Global Rating of Pain Scale (GROC) was administered. Each participant was verbally instructed to carefully read and circle the response 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. For the second test occasion, a subset of the participants (102 in total) that verbally confirmed being in unchanged/stable pain in the past 7-days were conveniently sampled to self-complete the SF-MPQ-2 and GROC at home within 3–7 days, if their pain remained unchanged (i.e. if they could confirm that the threshold of their perceived pain for their shoulder disorder had not changed in the past week). The GROC

scale was administered, intentionally, on both test occasions solely to serve as an objective means of comparing participants test and retest responses thus ensuring that only participants in stable/unchanged pain conditions were included in our analysis of reproducibility (test–retest reliability and agreement). Demographic information including age, hand dominance, primary cause of shoulder pain and sex were recorded.

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: (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). 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-subscale yields the summary score for the subscale [15, 21]. Higher subscale or total scores suggest greater pain symptoms/experience, and more than 2 missing values renders patients' response to the questionnaire invalid [21]. 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 [15].

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 [21]. Continuous variables were descriptively summarized using means and standard deviations while percentages were used to report categorical variables. The data were then examined for normality with histograms, and the Shapiro–Wilk test. All statistical analyses were completed with Microsoft Excel Version 2013 and SPSS statistic for Windows™, Version 25.0. (Armonk, NY: IBM Corp, Released 2017).

Floor/ceiling effects

Floor/ceiling effects for the SF-MPQ-2 were assessed by identifying the number of participants with the absolute lowest (0-points = floor) and highest (10-points = ceiling) scores on the total and subscales. Floor/ceiling effects

occurring at the magnitude of 15% were considered substantial [30].

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 Patient-Reported Outcome Measure (PROM) [30–32], was assessed with Cronbach alpha (α) and associated 95% confidence intervals. An $\alpha \geq 0.7$ is a commonly accepted standard for internal consistency reliability. However, redundancy is suggested at $\alpha > 0.95$ [30, 32, 33].

Hypothesis: We expected the SF-MPQ-2 to be internally consistent with Cronbach α at 0.8 or above for its subscale scores, and 0.9 or above for its total scores as previously reported in the literature [22, 24].

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 [34]. $ICC_{2,1}$ with 95% confidence intervals (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 [35]. We chose an $ICC_{2,1}$ absolute agreement over a consistency model because it captures elements of systematic bias and is preferred for computing an 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 good $ICC_{2,1}$ scores for group level analysis at ≥ 0.80 for the total scale and ≥ 0.70 for the subscale scores as previously reported in the literature [22, 24].

Agreement properties (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 [37]. Table 1 explains the five equations used for agreement analysis. To define $SEM_{\text{agreement}}$ for the SF-MPQ-2 total and subscales scores, the pooled standard deviation calculated from participants' mean responses to

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 $MDC_{90individual}$ 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 $MDC_{90group}$ score for the entire population

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

the SF-MPQ-2 domains on both test and retest using Eq. 1 [37, 38] and the respective non-transformed $ICC_{2,1}$ for the SF-MPQ-2 domain under evaluation was keyed into Eq. 2 [37–39] (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 [39–41] and interpreted as follows: $\leq 5\%$ = very good; $> 5 - \leq 10\%$ = good; $> 10 - < 20\%$ = doubtful; and values above 20% = negative [39].

The minimal detectable change (MDC) or repeatability coefficient describes the minimum amount of change that must occur on a score to be confident that true/real change (that may or may not be clinically significant) has occurred without error after two repeated measures, within the period of the test–retest [42]. 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 ($MDC_{90individual}$) or for a group ($MDC_{90group}$) [29]. We estimated $MDC_{90individual}$ for the total and subscale scores of the SF-MPQ-2 by entering each scale’s $SEM_{agreement}$ into Eq. 3 (Table 1) assuming the data was normally distributed and free of systematic error. The $MDC_{90individual}$ confidence interval was then computed from the mean differences (*d*) of each subscale using Eq. 4 (Table 1) [29, 40, 43].

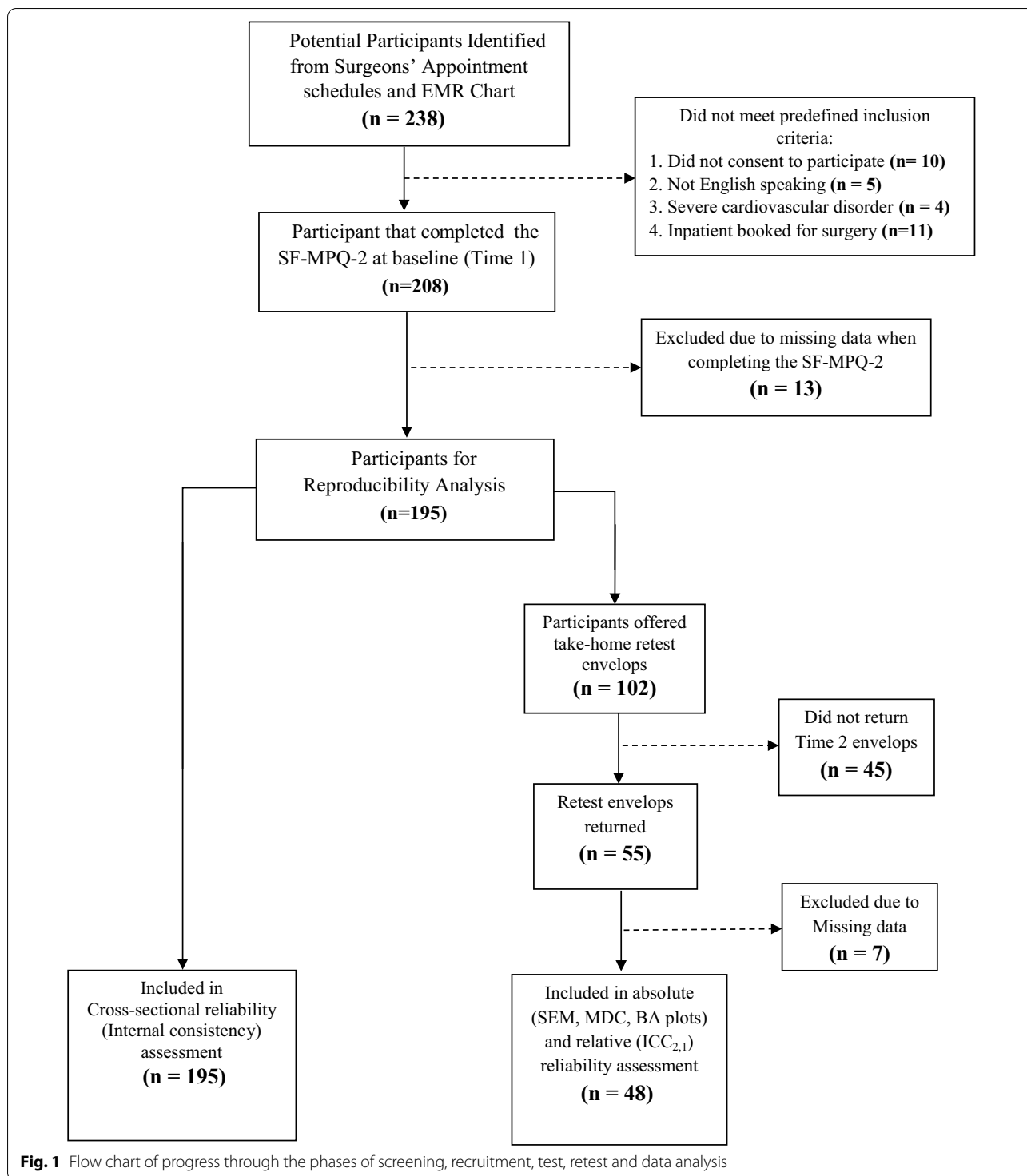
To determine the group level minimal detectable change ($MDC_{90group}$), which is useful for determining if changes have occurred in an entire population, Eq. 5 (Table 1), the formula proposed by de Vet et al. [30, 44] was employed. The proportion of the resulting MDC coefficient per SF-MPQ-2 domain to the total score of the scale was computed to yield the MDC percent score (MDC%) and interpreted as follows: $\leq 5\%$ = very good; $> 5 - \leq 10\%$ = good; $> 10 - < 20\%$ = doubtful; and values above 20% = negative [39, 40].

Bland–Altman Plots (BA Plots)

The Bland–Altman (BA) method was used to visually examine the agreement between the test and retest scores [45, 46]. Scatter plots were created to demonstrate the differences between the total and subscale scores obtained at time one and time two of the test–retest interval against their mean score for the two time points [45–48]. We then calculated the mean difference between the two measurement intervals (the ‘bias’) and the 95% limits of agreement (LoA) using: $LoA = \text{mean difference (d)} \pm 1.96 \text{ SD of the mean differences}$. The BA plots were used to visually judge the 95% limits of agreement to determine how well the scores from repeated measurements agreed: narrower LoAs suggested better agreement at the individual level [29, 47, 49]. 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 [44–48, 50]. 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 potential for systematic bias was appraised by checking if the prediction of the differences in the mean scores was statistically significant [47, 51]. Finally, outliers that presented beyond the upper and lower boundaries of the LoA were noted and explored [29, 52].

Results

Figure 1 below summarizes the flow of participants through the different phases of the study. Of the 238 eligible patients identified from the review of the surgeons’ scheduled appointment list, 195 consenting adults



satisfied the inclusion criteria and provided complete data that were considered in our analysis of cross-sectional reliability. For the analysis of test–retest reliability and agreement, of the 102 participants that agreed to participate in the second test occasion, only 48 out of

55 stable subjects provided a complete response to SF-MPQ-2 in a mean of 4 days following the index test.

Table 2 summarizes the characteristic and demographic distribution of the baseline population. The study population was equally comprised of males and females,

Table 2 Patient baseline characteristics (N = 195)

Variables	N/%
Age in years (mean + SD)	(62 + 17) 195/100%
<i>Shoulder disorders</i>	
Glenohumeral joint arthroplasty	39/20%
Humeral and others fractures (i.e. clavicular, costal, scapular)	23/12%
Rotator cuff pathologies	48/25%
Dislocation	12/6%
Osteoarthritis	18/9%
Impingement/bursitis	15/8%
Other (MSK pain-related/non-specific)	40/21%
<i>Affected shoulder</i>	
Right	111/56%
Left	71/36%
Both	13/6%
<i>Sex</i>	
Males	103/53%
Females	92/47%

N number of patients, SD standard deviation

with a mean age of 62 years, with different shoulder disorders of various MSK pathologies including rotator cuff injuries, humeral fracture and glenohumeral joint arthroplasty.

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 small difference in scores (see Table 3 for results). Parametric statistics were used in our analysis because the

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 = 15%	0/48 = 0%	4/48 = 8%	1/48 = 2%
SF-MPQ-2 Intermittent	11/48 = 23%	0/48 = 0%	15/48 = 31%	0/48 = 0%
SF-MPQ-2 Affective	19/48 = 40%	1/48 = 2%	20/48 = 42%	0/48 = 0%
SF-MPQ-2 Neuropathic	14/48 = 29%	0/48 = 0%	11/48 = 23%	0/48 = 0%
SF-MPQ-2 Total	3/48 = 6%	0/48 = 0%	4/48 = 8%	0/48 = 0%

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

sample size was greater than 30 participants (based on the central limit theorem). Despite that, we still examined for differences in reproducibility coefficients obtained using the transformed and non-transformed ICC scores.

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 a decline in status can be captured, and vice versa for floor effects [21]. 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 effect 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.

Internal consistency (cross-sectional reliability)

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

Agreement properties (absolute test-retest reliability)

Table 5 summarizes the agreement parameters supporting the SF-MPQ-2 domains. The total scale SEM_{agreement} was very low (0.51points) and approximately 5% of the total score of the scale, which is 'very good' according to

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 Agreement parameters (absolute reliability) of the SF-MPQ-2 total and subscale scores (N = 48)

Variables	SEM _{agreement}	SEM (%)	MDC _{90individual} (95% CI)	MDC (%)	MDC _{90group}
SF-MPQ-2 Continuous	0.8	8	1.8 (− 1.6 to 2.0)	18	0.4
SF-MPQ-2 Neuropathic	0.8	8	1.8 (− 1.7 to 1.9)	18	0.4
SF-MPQ-2 Intermittent	1.0	10	2.3 (− 2.1 to 2.4)	23	0.5
SF-MPQ-2 Affective	0.7	7	1.7 (− 1.5 to 1.8)	17	0.4
SF-MPQ-2 Total	0.5	5	1.2 (− 1.0 to 1.4)	12	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 SEM_{agreement} or MDC_{90individual} of domain represented on the SF-MPQ-2 to the total score of the scale (i.e. 10 points)

our criteria. Individual subscale SEM_{agreement} ranged from 0.73 to 0.99 (approximately ≤ 10% of the total score), which is also ‘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 (MDC) at a 90% confidence level. Of all the SF-MPQ-2 domains, the total scale had the lowest MDC score at 1.20 points (i.e. 12%) while the intermittent subscale had the most substantial MDC scores at 2.29 points (i.e. 23%). For Group MDC₉₀, estimates were acceptable and expectedly lower than those obtained for MDC_{90individual}; the results fluctuated within 0.28 (total) to 0.54 (intermittent) points across the SF-MPQ-2 domains (Table 5).

Relative test–retest reliability

The test–retest reliability of the SF-MPQ-2 domains was rated “Good” to “Excellent” (Table 6). Our results for ICC_{2,1} were based on an 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. The neuropathic, affective and intermittent subscales displayed good ICC_{2,1} coefficients (Table 6) in support of relative reliability.

Bland–Altman (BA) analysis/plots

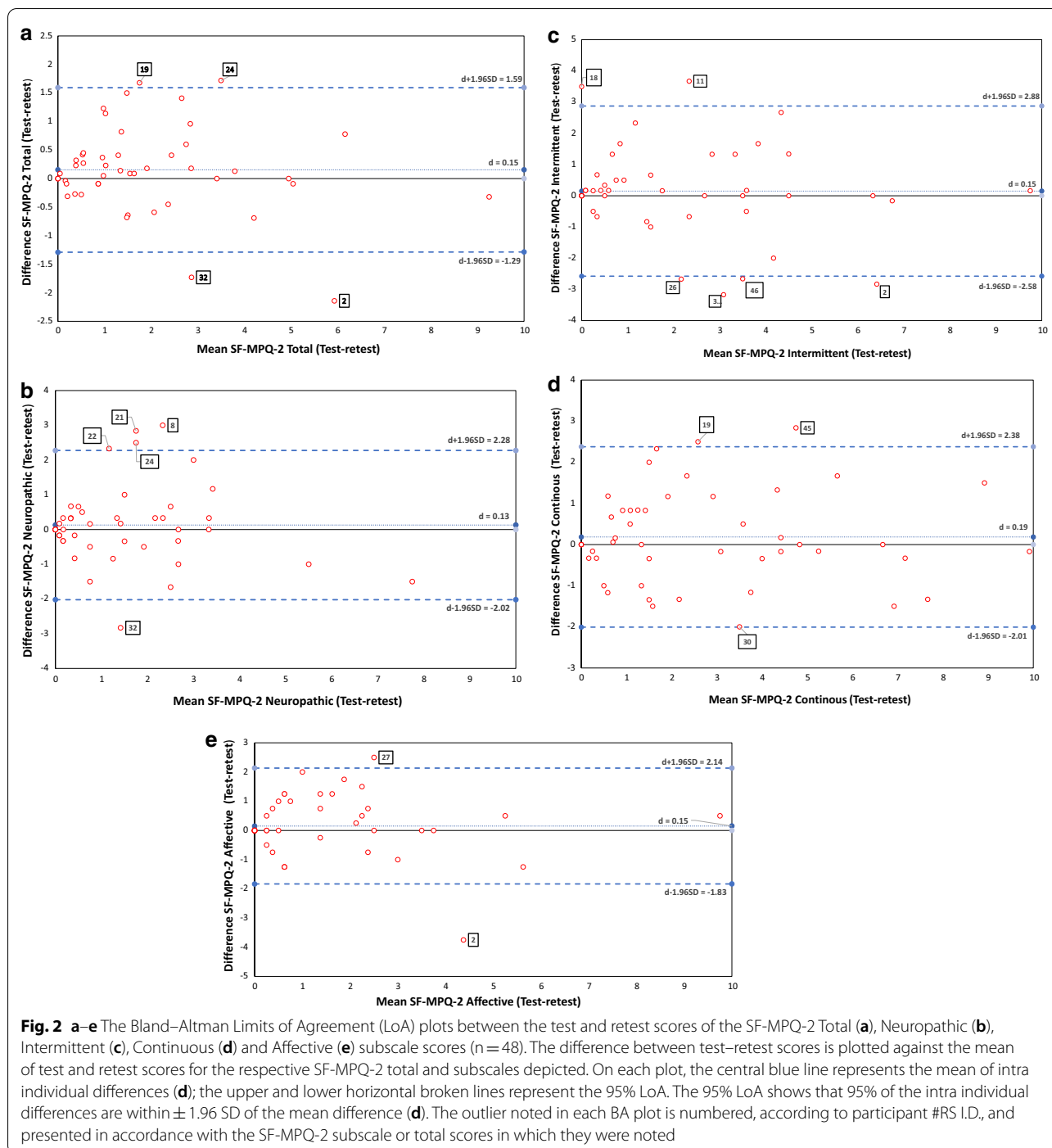
The results of our Bland–Altman analysis are presented in Table 6. The Bland–Altman plots superimposed with the LoA and mean difference (bias) scores for each domain of the SF-MPQ-2 are graphically illustrated (Fig. 2a–e). All of the SF-MPQ-2 domains displayed acceptable LoA at a 95% confidence level 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).

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 to 0.51	− 2.01, 2.38	^a 0.90 (0.83–0.94)
SF-MPQ-2 Intermittent	2.1 (2.3)	2.0 (2.4)	0.15 (1.39)	− 0.24 to 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 to 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 to 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 to 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)



Visual inspection of scatter points on the BA plots for each domain of the SF-MPQ-2 revealed that the magnitude of the mean differences against the mean scores were uniformly distributed from the zero point and most scatter points were within the 95% LoA with the exception of a 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. These findings suggest that systematic bias is unlikely and confirms good level of agreement between the test–retest scores (Table 6).

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 [53] and labelled according to their #RS on each BA plot. The greatest number of interesting outliers presented on the intermittent ($n=6$, 12%) and neuropathic ($n=4$, 10%) subscales. The least number of outliers were seen on the affective subscale ($n=2$, 4%). In general, however, the presence of these outliers did not indicate the presence or absence of bias [53].

Discussion

This study provides reproducibility evidence that supports the use of the SF-MPQ-2 in multidimensional pain assessment of people with MSK shoulder pain. The SF-MPQ-2 displayed good to excellent coefficients in support of its relative reliability and absolute reliability properties. The limits of agreement for the subscales and total scores were very satisfactory.

The substantial floor effect observed on the neuropathic, intermittent and affective subscales can be attributed to the robust discriminative properties of the SF-MPQ-2 subscales and to the lower prevalence of these problems in our study population. Conceptually, the SF-MPQ-2 was expanded to provide a single tool that can classify pain between neuropathic and non-neuropathic sources [15, 21]. As outcome measures can be evaluative or discriminative, combining both purposes within an outcome measure is likely to result in these types of statistical issues. For instance, participants with pain emerging from neuropathic sources will be more inclined to respond adequately to the neuropathic subscale, thereby reducing the likelihood of floor effects. This has been observed with the use of the SF-MPQ-2 among complex regional pain syndrome (CPRS) patients [20]. This implies that floor effects on the SF-MPQ-2 domains may not always represent redundancy, but rather, 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 that are similar to previous estimates among mixed-MSK [23] (total, 0.93; subscale, 0.84–0.92), CRPS [20] (total, 0.95; neuropathic subscale, 0.83), knee OA [22] (total, 0.88; subscale 0.75–0.81) and acute back pain [21] (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 [32] 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 that they can adequately discriminate among patients at the individual level (total and continuous scale) and at the group level (all of the SF-MPQ-2 domains) [29, 54]. These results are comparable or better than previous findings reporting estimates among knee OA [22] (total scale, 0.90; subscales, 0.73–0.90) and mixed MSK patients [24, 55] (total scale, 0.90–0.94; subscales, 0.73–0.90). Although acceptable, the lower performance of the neuropathic subscale (0.78), with an ICC score that overlapped the ‘moderate’ confidence interval threshold (0.64–0.87), suggests greater variability on this subscale, which makes it more difficult to achieve a high 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 [30, 44]. 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 the Ostelo et al. [39] 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 ($\leq 5\%$ of total scale score) was ‘very good’ and comparable to that reported for the former version (SF-MPQ) among OA patients ($\leq 3.64\%$) [56], but better than those seen among mixed MSK patients assessed with the Norwegian version of the SF-MPQ ($\leq 10\%$) [41]. Although not as favorable as estimates noted on the total scale, the affective and intermittent/continuous subscales had ‘good’ SEM coefficients ($< 10\%$), which were comparable to findings reported with the sensory subscale of the former SF-MPQ version among OA patients ($< 10\%$) [56], and superior to that reported in a mixed MSK population ($< 14\%$) [41]. 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 represent the minimal change in scores after repeated administration that clinicians/researchers can interpret as not due to chance variation for an individual or group in a population [42]. The MDC_{90individual} 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 chance with 90% confidence. The MDC scores for the total scale ($\leq 12\%$ of the total score of the scale) were comparable to previous studies with the former version (SF-MPQ) among OA patients ($\leq 11.5\%$) and better than the results seen among mixed MSK patients ($\leq 26.4\%$ of total score). For

the $MDC_{90\text{group}}$ scores, the results obtained for the SF-MPQ-2 domains imply that a change of at least 0.4 (affective), 0.5 (intermittent), 0.3 (total), 0.4 (neuropathic), 0.4 (continuous) points must be observed in a group to be 90% confident that this was change beyond random or systematic error. In general, minimal detectable change 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 in the measure beyond error for a group, if the Minimal Clinically Important Difference (MCID) score for the population is unknown.

The Bland–Altman plots revealed satisfactory limits of agreement in support of the SF-MPQ-2 subscales. However, the interpretation of how far apart two measurements can be before they are no longer considered interchangeable depends on the contextual application [47]. The limits of agreement between the test–retest of the SF-MPQ-2 domains were reasonably smaller than those seen in previous studies of its former version (SF-MPQ) [41, 56], suggesting there is less variation between the test and the retest of the SF-MPQ-2 [50]. Furthermore, no bias was found in the measurements between the test–retest, as 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) [28, 29, 57] may have favored the agreement outcomes. The intermittent subscale had the greatest number of outliers of all the Bland–Altman plots (12%) and may be due to the 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. Variability decreases when a greater number of descriptors comprise a scale, in comparison to those with fewer descriptors [29]. As all 22 items of the SF-MPQ-2 contribute to the summary total scale scores, it is possible this favors reproducibility.

Study limitations

While the present study findings provide preliminary evidence supporting the reproducibility of the SF-MPQ-2 for use in patients with shoulder disorders, it has several limitations. First, the study sample size (48 participants) was just under 50 participants which has been suggested as a benchmark by the COSMIN [58, 59]. However, in

conflict with the COSMIN recommendation, our sample size calculation suggested at least 46 patients were required (see Appendix 1), which indicates our study was adequately powered. Second, the patient population were from a single tertiary referral practice and our findings may not be generalizable to a different context. 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) years, which may not adequately reflect the reliability of younger populations although shoulder pathology prevalence increases with age. Finally, we did not determine minimal clinically important difference.

Conclusion

We conclude that the SF-MPQ-2 is satisfactorily internally consistent and provides good to excellent reproducibility coefficients (test–retest reliability and agreement) for multidimensional pain assessment among patients with musculoskeletal shoulder pain conditions. The total scale displays the best reproducibility coefficients. Additional research on the validity and responsiveness of the SF-MPQ-2 is still required in this population.

Abbreviations

SF-MPQ-2: Revised Short McGill Pain Questionnaire Version-2; SF-MPQ: Short-Form McGill Pain Questionnaire; ICC: Intraclass correlation; CI: Confidence interval; SEM: Standard error of measurement; MDC: Minimal detectable change; LoA: Limits of agreement; BA plot/analysis: Bland–Altman analysis/plots; COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments; CPRS: Complex Regional Pain Syndrome.

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.

Authors' contributions

SJ conceived and led the study design, data collection, data analysis and drafted the manuscript. JCM and TP contributed to the design and development of the broader overall project and revised the manuscript for intellectual content. GA and KF assisted with data collection and revised the manuscript for intellectual content. All authors read and approved the final manuscript.

Funding

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

Availability of data and materials

This study data set is not assessable to the general public. However, the analysis of the dataset is assessable on reasonable request to the corresponding author, Samuel U. Jumbo.

Ethics approval and consent to participate

Ethics approval for a clinical database of routine outcome measures from which this data was extracted was approved by the University of Western Ontario Research Ethics Board (REB# 4986).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Faculty of Health and Rehabilitation Sciences, Elborn College, Western University, London, ON, Canada. ² School of Rehabilitation Science, McMaster University, 1400 Main Street West, Hamilton, ON, Canada. ³ Roth McFarlane Hand and Upper Limb Centre, St. Joseph's Hospital, London, ON N6A 4L6, Canada.

Appendix 1: Formula used for sample size calculation

$$n = 2 + \frac{0.5k(Z_{\alpha} + Z_{\beta})^2}{\delta^2(k - 1)}$$

where:

- k = number of occasions = 2
- Z_{α} represents the Z-value associated with the α value of interest. Therefore, Z-value (1-tailed) for 0.05 is equivalent to 1.645
- Z_{β} represents the Z-value associated with a Type II error. At power 80%, β equals 0.20, and Z-value (1-tailed) equals 0.842.
- δ = R value of Z transformed expected hypothesis minus R value of Z transformed null hypothesis, i.e. $\delta = Z_{\text{Rexpected}} - Z_{\text{Rnull}}$
Where
 $Z_{\text{Rexpected}}$ = reliability value expected from analysis
 $\text{Rexpected} = 0.9$
 Z_{Rnull} = lower confidence limit for the desired confidence interval width
- $Z_{\text{Rexpected}} = 0.5 \text{ natural log } \frac{1+(k-1)\text{Rexpected}}{1-\text{Rexpected}} = 1.47$
- $Z_{\text{Rnull}} = 0.5 \text{ natural log } \frac{1+(2-1)\text{Rlowerlimit}}{1-\text{Rlowerlimit}} = 1.09$
- $\text{Rlowerlimit} = 0.09 - 0.10 = 0.80$ (at 0.10 Confidence Interval width)
- $Z_{\text{Rnull}} = 0.5 \text{ natural log } \frac{1+(2-1)0.8}{1-0.8} = 1.09$
- $\delta = 1.47 - 1.09 = 0.38$
- $n = 2 + \frac{0.5k(Z_{\alpha} + Z_{\beta})^2}{\delta^2(k - 1)} = 2 + \frac{0.5(2)(1.645 + 0.842)^2}{(0.38)^2(2 - 1)}$
- n = 46 patients

A sample size of 46 participants will be required for the reliability analysis.

Received: 29 October 2019 Accepted: 3 November 2020
Published online: 11 November 2020

References

1. McCormick A, Fleming D, Charlton J. Morbidity statistics from general practice: fourth national study 1991–92. Office of Population Censuses and Surveys, London. 1995. <https://trove.nla.gov.au/version/25109531>. Assessed 25 Aug 2019
2. Linsell L, Dawson J, Zondervan K, Rose P, Randall T, Fitzpatrick R, et al. Prevalence and incidence of adults consulting for shoulder conditions in UK primary care; patterns of diagnosis and referral. *Rheumatology*. 2006;45(2):215–21.
3. Chard MD, Hazleman R, Hazleman BL, King RH, Reiss BB. Shoulder disorders in the elderly: a community survey. *Arthritis Rheum*. 1991. <https://doi.org/10.1002/art.1780340619>.
4. 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. <https://doi.org/10.2519/jospt.2015.5033>.
5. 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;32:107–12.
6. van der Heijden GJMG. Shoulder disorders: a state-of-the-art review. *Baillieres Best Pract Res Clin Rheumatol*. 1999;13(2):287–309.
7. 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.
8. 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):30–9.
9. McGuire DB. Comprehensive and multidimensional assessment and measurement of pain. *J Pain Symptom Manag*. 1992;7(5):312–9.
10. Ahles TA, Blanchard EB, Ruckdeschel JC. The multidimensional nature of cancer-related pain. *Pain*. 1983;17(3):277–88.
11. Ho K, Spence J, Murphy MF. Review of pain-measurement tools. *Ann Emerg Med*. 1996;27(4):427–32.
12. Badalamente M, Coffelt L, Elfar J, Gaston G, Hammert W, Huang 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):407–12.
13. 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.
14. 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.
15. 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.
16. Gauthier LR, Young A, Dworkin RH, Rodin G, Zimmermann C, Warr D, et al. Validation of the Short-Form McGill Pain Questionnaire-2 in younger and older people with cancer pain. *J Pain*. 2014;15(7):756–70.
17. Ortner C, Turk D, Theodore B, Siaulys M, Bollag L, Landau R. The short-form mcgill 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):478–86.
18. 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;10:121–8.
19. Morgan KJ, Angheliescu DL. A review of adult and pediatric neuropathic pain assessment tools. *Clin J Pain*. 2017;33(9):844–52.
20. Packham TL, Bean D, Johnson MH, MacDermid JC, Grieve S, McCabe CS, et al. Measurement properties of the SF-MPQ-2 neuropathic qualities subscale in persons with CRPS: validity, responsiveness, and Rasch analysis. *Pain Med*. 2018;20(4):799–809.
21. Dworkin RH, Turk DC, Trudeau JJ, Benson C, Biondi DM, Katz NP, et al. Validation of the Short-Form McGill Pain Questionnaire-2 (SF-MPQ-2) in acute low back pain. *J Pain*. 2015;16(4):357–66.
22. 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.
23. Lovejoy TI, Turk DC, Morasco BJ. Evaluation of the psychometric properties of the Revised Short-Form McGill Pain Questionnaire. *J Pain*. 2012;13(12):1250–7.
 24. Adelmanesh F, Jalali A, Attarian H, Farahani B, Ketabchi SM, Arvantaj A, 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):1631–8.
 25. Jumbo S: Psychometric properties of the Brief Pain Inventory-Short Form and Revised Short McGill Pain Questionnaire Version-2 in musculoskeletal conditions. <https://ir.lib.uwo.ca/etd/6490> (2019). Assessed 4 Sept 2019.
 26. MacDermid J, Jumbo S, Kalu M, Packham T, Athwal G, Faber K. AB1328. 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. *Ann Rheum Dis*. 2019;78:2128.
 27. 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 McGill Pain Questionnaire-Version-2 (SF-MPQ-2) in pain-related musculoskeletal conditions: a systematic review protocol. *Arch Bone Jt Surg*. 2020;8(2):131–41.
 28. 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):730–5.
 29. 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):281–91.
 30. 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*. 2007;60(1):34–42.
 31. Cronbach LJ. Test "reliability": its meaning and determination. *Psychometrika*. 1947;12(1):1–16.
 32. Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ*. 2011;2:53–5.
 33. Streiner D, Norman G, Cairney J. Health measurement scales: a practical guide to their development and use. 5th ed. Oxford: Oxford University Press; 2015.
 34. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–8.
 35. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil*. 1998. <https://doi.org/10.1191/026921598672178340>.
 36. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30–46.
 37. Harvill LM. An NCME instructional module on standard error of measurement. *Educ Meas Issues Pract*. 1991. <https://doi.org/10.1111/j.1745-3992.1991.tb00195.x>.
 38. 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. <https://doi.org/10.2519/jospt.2011.3666>.
 39. 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):268–76.
 40. 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):571–8.
 41. 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):917–25.
 42. Haley SM, Fragala-Pinkham MA. Interpreting change scores of tests and measures used in physical therapy. *Phys Ther*. 2006;86(5):735–43.
 43. Chesworth BM, Hamilton CB, Walton DM, Benoit M, Blake TA, Bredy H, et al. Reliability and validity of two versions of the upper extremity functional index. *Physiother Canada*. 2014. <https://doi.org/10.3138/ptc.2013-45>.
 44. 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):479–87.
 45. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327(8476):307–10.
 46. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999. <https://doi.org/10.1177/096228029900800204>.
 47. Myles PS, Cui JJ. Using the Bland–Altman method to measure agreement with repeated measures. *Br J Anaesth*. 2007;99(3):309–11.
 48. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ*. 1992;304(6840):1491–4.
 49. 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 Psychogeriatr*. 2014;26(10):1659–67.
 50. 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. <https://doi.org/10.1177/1758998313515191>.
 51. 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):1217–22.
 52. Flansbjerg 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.
 53. Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. *Organ Res Methods*. 2013. <https://doi.org/10.1177/1094428112470848>.
 54. 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*. 2011;48(6):661–71.
 55. Maruo T, Nakae A, Maeda L, Kenrin S, Takahashi K, Morris S, 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. <https://doi.org/10.1111/pme.12468>.
 56. Grafton KV, Foster NE, Wright CC. Test-Retest Reliability of the Short-Form McGill Pain Questionnaire. *Clin J Pain*. 2005;21(1):73–82.
 57. 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.
 58. 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;27(5):1147–57.
 59. 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;27(5):1171–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.