Measurement Properties of Performance-Specific Pain Ratings of Patients Awaiting Total Joint Arthroplasty as a Consequence of Osteoarthritis

Ashley Halket, Paul W. Stratford, Deborah M. Kennedy, Linda J. Woodhouse, Gregory Spadoni

ABSTRACT

Purpose: To estimate the test–retest reliability of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain sub-scale and performance-specific assessments of pain, as well as the association between these measures for patients awaiting primary total hip or knee arthroplasty as a consequence of osteoarthritis.

Methods: A total of 164 patients awaiting unilateral primary hip or knee arthroplasty completed four performance measures (self-paced walk, timed up and go, stair test, six-minute walk) and the WOMAC. Scores for 22 of these patients provided test–retest reliability data. Estimates of test–retest reliability (Type 2,1 intraclass correlation coefficient [ICC] and standard error of measurement [SEM]) and the association between measures were examined.

Results: ICC values for individual performance-specific pain ratings were between 0.70 and 0.86; SEM values were between 0.97 and 1.33 pain points. ICC estimates for the four-item performance pain ratings and the WOMAC pain sub-scale were 0.82 and 0.57 respectively. The correlation between the sum of the pain scores for the four performance measures and the WOMAC pain sub-scale was 0.62.

Conclusion: Reliability estimates for the performance-specific assessments of pain using the numeric pain rating scale were consistent with values reported for patients with a spectrum of musculoskeletal conditions. The reliability estimate for the WOMAC pain sub-scale was lower than typically reported in the literature. The level of association between the WOMAC pain sub-scale and the various performance-specific pain scales suggests that the scores can be used interchangeably when applied to groups but not for individual patients.

Key Words: arthroplasty, numeric pain rating scale, osteoarthritis, performance measures, reliability, WOMAC

Halket A, Stratford PW, Kennedy DM, Woodhouse LJ, Spadoni G. Measurement properties of performance-specific pain ratings of patients awaiting total joint arthroplasty as a consequence of osteoarthritis. Physiother Can. 2008;60:255-263.

RÉSUMÉ

Objet : Évaluer la fidélité du test-retest du sous-échelle de douleur de l'index de sévérité symptomatique de l'arthrose des membres inférieurs (WOMAC—Western Ontario and McMaster Universities Osteoarthritis Index) et des évaluations spécifique à la performance, et l'association entre ces mesures pour les patients qui attendent une arthroplastie primaire totale de la hanche ou du genou suite à l'arthrose.

Méthodologie : Cent soixante-quatre patients qui attendaient une arthroplastie primaire unilatérale de la hanche ou du genou ont exécuté quatre mesures de performance (marche adaptée au rythme de chacun, Test du lever de chaise de Mathias, le test des escaliers, la marche de six minutes) et le WOMAC. Les scores pour 22 de ces patients ont fourni des données de fidélité du test-retest. Les évaluations de fidélité du test-retest (coefficient de corrélation intraclasse [CCI] de type 2,1, l'erreur standard sur la mesure [ESM]) et l'association entre les mesures ont été examinées.

Résultats : Les valeurs du CCI pour les évaluations individuelles de la douleur propre à la performance varient entre 0,70 et 0,86; les valeurs de l'ESM varient entre 0,97 et 1,33 point de douleur. L'évaluation CCI pour les classements de douleur de performance à quatre éléments et l'échelle de douleur de WOMAC sont de 0,82 et 0,57, respectivement. La corrélation entre la somme des scores de douleur pour les quatre mesures de performances et du sous-échelle de douleur de WOMAC est de 0,62.

Conclusion : Les évaluations de fidélité pour les évaluations de la douleur propre à la performance à l'aide de l'échelle d'évaluation numérique de la douleur sont conformes aux valeurs signalées pour les patients ayant un spectre de troubles musculosquelettiques. L'évaluation de fidélité du sous-échelle

Ashley Halket, BA, CAT(C), MSc: Assistant Clinical Director, Sheridan College, Oakville, Ontario. At the time of the study, Ms Halket was completing her MSc at the School of Rehabilitation Science, McMaster University, Hamilton, Ontario.

Paul W. Stratford, PT, MSc: Professor, School of Rehabilitation Science, and Associate Member, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario.

Deborah M. Kennedy, BSc(PT), MSc: Manager of Program Development for Hip and Knee Replacement, Sunnybrook Holland Orthopaedic & Arthritic Centre, Toronto, Ontario; part-time Assistant Clinical Professor, School of Rehabilitation Science, McMaster University, Hamilton, Ontario.

Linda J. Woodhouse, PhD, MA, BSc(PT), BA: Assistant Professor, School of Rehabilitation Science, McMaster University, Hamilton, Ontario; Scientific Affiliate, Hamilton Health Sciences Corporation and Sunnybrook Health Sciences Centre, Hamilton, Ontario.

Gregory F. Spadoni, PT, MSc, FCAMT: Assistant Professor, School of Rehabilitation Science, McMaster University, Hamilton, Ontario; Director, Peak Performance Physiotherapy, Hamilton, Ontario

Address correspondence to *Ashley Halket*, 2004 Glenada Crescent, Unit 16, Oakville, ON L6H 5P5 Canada; Tel: (905) 338-8467; E-mail: ashleyhalket@hotmail.com.

DOI:10.3138/physio.60.3.255

de douleur de WOMAC est inférieure à celle typiquement signalée dans la documentation. Le niveau d'association entre le sous-échelle de douleur de WOMAC et les diverses échelles de douleur propre à la performance suggèrent que les scores sont interchangeables lorsqu'ils sont appliqués à des groupes, mais pas à des patients individuels.

Mots clés : arthroplastie, échelle numérique d'évaluation de la douleur, arthrose, mesures de performance, fidélité, WOMAC

BACKGROUND

Unilateral primary total joint replacement surgeries are performed to reduce pain and improve function in patients with end-stage osteoarthritis (OA). Thus, measures that reliably evaluate pain and function in patients with OA of the hip or knee is essential. Such measures are useful to triage patients to conservative management or surgery and to evaluate progress pre–, peri–, and post–total joint arthroplasty (TJA).[1] Self-report measures are the most common method of assessing pain and function; however, investigations have consistently demonstrated the inability of self-report measure items to differentiate between these health concepts.[1–3] In contrast, performance-related assessments have been shown to provide reasonably distinct ratings of pain and function.[4] Although information is available concerning the measurement properties of performance tasks (i.e., time and distance)[5] applied to patients with OA of the hip or knee progressing to TJA, little information is available on performance-specific pain ratings. Accordingly, it would be useful to know the extent to which performance-specific pain ratings are reliable and the extent to which they are related to non-performance-specific pain ratings such as those obtained from the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) pain sub-scale.

Historically, self-report assessments of pain and physical function with no performance component have been the preferred method of assessment in patients with OA and those progressing to arthroplasty.[1,6] For example, at the Outcome Measures in Arthritis Clinical Trials (OMERACT) III conference, pain and physical function were identified as two of the four core outcomes to be assessed in patients with OA.[1] Moreover, an intriguing aspect of the OMERACT III statement was that self-report was the preferred method for assessing physical function. This view was shared also in an authoritative review of outcome measures relevant to patients with OA.[6] Subsequent to these recommendations, a body of research has shown consistently that self-reports of physical function are strongly influenced by what patients experience when moving around, as well as by their ability to move around.[2,3,7–10] The evidence includes both theoretical considerations and observed consequences.[2,3,7–10] The theoretical evidence challenges the construct validity of the WOMAC. Specifically, a number of factor analyses applied to WOMAC items have failed to identify pain and physical function as distinct factors.[7–10] Perhaps of greater concern is that empirical investigations have shown that the WOMAC's physical function sub-scale failed to detect important deterioration in physical function of patients within 2 months after arthroplasty, even when the times to complete performance tasks had more than doubled.[11,12]

By contrast, performance-specific assessments have been shown to be more adept at distinguishing between these concepts. Specifically, a confirmatory factor analysis has supported a two-factor model with time (distance) assessments of performance tasks loading on one factor and task-specific pain ratings loading on a second factor.[4] The present study obtained assessments of pain using an 11-point numeric pain rating scale immediately following each of four performance activities.[4] The performance activities were a 40-m self-paced walk, a stair test, the timed up and go (TUG), and a 6-minute walk. Although the measurement properties of the time and distance components of these tests have been reported for patients with OA and those undergoing total hip arthroplasty (THA) or total knee arthroplasty (TKA), no information has been offered on the measurement properties of the performance-specific pain rating.[4,5]

Although many studies have explored the measurement properties of the 11-point numeric pain rating scale, we are unaware of investigations that have applied this measure in the context of performance-specific pain ratings completed by patients with OA awaiting THA or TKA. This is important because measurement properties pertain not to a measure but, rather, to the measure's scores. The rationale for this statement is explained by Messick:

> The emphasis is on scores and measurements as opposed to tests or instruments because the properties that signify adequate assessment are properties of scores, not tests. Tests do not have reliabilities or validities, only test responses do. This is an important point because test responses are a function not only of the items, tasks, or stimulus conditions but of the persons responding and the context of measurement.[13]

Accordingly, the goals of this study were (1) to estimate the test–retest reliability of performance-specific assessments of pain in patients awaiting primary unilateral THA or TKA as a consequence of OA; (2) to

estimate the test–retest reliability of the WOMAC pain sub-scale on the same patient sample; (3) to estimate the difference in reliability between the performance-specific estimates of composite pain scores and WOMAC pain sub-scale scores; and (4) to examine the association between composite performance-specific pain scores and WOMAC pain scores.

METHODS

The study was conducted at the Holland Orthopaedic and Arthritic Centre of Sunnybrook Health Sciences Centre in Toronto, Ontario. Ethics approval was obtained from the Research Ethics Boards of both Sunnybrook and McMaster University. All participants provided written informed consent prior to participating in the study.

Participants

The sample consisted of individuals scheduled for either THA or TKA at the Holland Centre who met the following eligibility criteria. First, patients were considered for inclusion if they were scheduled for primary THA or TKA as a result of osteoarthritis. Second, patients had to be capable of completing the four physical performance tests and have sufficient English skills to give written informed consent and to complete the self-report measures of pain. Individuals with neurological, cardiac, psychiatric, or other medical conditions that would compromise their ability to complete the functional tasks were excluded from the study.

Design

The present study represents a secondary analysis of data gathered as part of a larger longitudinal study that followed patients pre- and post-TJA.[14] Data for the current component of the study were collected prospectively and preoperatively: individuals were initially tested at the point of consultation with the orthopaedic surgeon and again preoperatively. The test–retest reliability subsample consisted of patients whose preoperative assessments occurred within 90 days of one another. We chose 90 days for two reasons: first, this is a common interval between physician assessments; and, second, there is evidence to suggest that many patients will not undergo a true change over an extended interval preoperatively.[5]

During each testing session, participants were asked to complete the entire WOMAC LK (Likert version) 3.1 and all four physical performance measures in the same order: self-paced walk test (SPW), stair test (ST), timed up and go test (TUG), 6-minute walk test (6MW). A 2-minute rest interval followed the SPW and ST, and a 5-minute rest interval preceded the 6MW. Immediately following each physical performance measure, participants were asked to rate their pain on a numeric pain rating scale (NPRS).

Measures

Western Ontario and McMaster Universities Osteoarthritis Index

The WOMAC LK 3.1 consists of 24 items divided into 3 sub-scales measuring pain, joint stiffness, and physical function. Each question is scored on a 5-point scale (0–4). The pain sub-scale of the WOMAC consists of 5 items that ask about pain (1) while walking on flat ground; (2) while going up or down stairs; (3) at night while in bed; (4) while sitting or lying; and (5) while standing upright. The WOMAC is a self-report measure that asks respondents to recall and rate the amount of pain recently experienced during these activities. Total pain scores can vary between 0 and 20; higher scores represent greater levels of pain.[15] Literature-based estimates of test–retest reliability for the WOMAC pain sub-scale vary between 0.77 and 0.86.[9,16,17] Estimates of the standard error of measurement (SEM) for the WOMAC pain sub-scale are in the range of 1.69 to 2.16.[18]

Numeric Pain Rating Scale

Patients provided performance-specific pain ratings on an 11-point (0–10) NPRS immediately after completing each performance test. The anchors for this scale were ''no pain'' (0) and ''pain as bad as it can be'' (10), and patients were instructed to circle the numeric value that best represented their current pain level.[19] The NPRS has been used to assess pain in a variety of orthopaedic conditions; however, its measurement properties have not been reported for performance-specific assessments of pain in this patient population. In patients with a spectrum of orthopaedic problems, estimates of test–retest reliability are in the order of 0.64 to 0.86, and typical SEM values for vary from 1.04 to 1.3.[20]

For each of the following performance tests, patients were allowed to use assistive devices. A practice session did not precede the tests.

Self-Paced Walk Test

This 40 m walk test took place indoors over a 20 m course that was completed twice. The turns were excluded and were not timed. Patients were instructed to ''walk as quickly as you can without overexerting yourself.'' Test–retest reliability and SEM for the timed component of the test have been estimated at 0.91 (Type 2,1 intraclass correlation coefficient [ICC]) and 1.73 seconds respectively in a sample of patients similar to those taking part in this study.[5]

Stair Test

Patients were asked to ascend and descend nine steps in a comfortable manner. Each step was 20 cm in height. Patients were informed that the pace should be comfortable and safe. Test–retest reliability and SEM for the timed component of the test have been estimated

at 0.90 (Type 2,1 ICC) and 2.35 seconds respectively in a sample of patients similar to those taking part in this study.[5]

Timed Up and Go (TUG) Test

For this timed test patients were instructed to rise from a seated position in a standard armchair and walk at a comfortable pace to a tape mark 3 m away. Patients then returned to the chair and resumed a seated position. Test–retest reliability and SEM for the timed component of the test have been estimated at 0.75 (Type 2,1 ICC) and 1.07 seconds respectively in a sample of patients similar to those taking part in this study.[5]

6-Minute Walk Test

Patients were instructed to cover as much distance as possible in 6 minutes and were given the opportunity to stop and rest if necessary. Because encouragement has been shown to influence the distance walked, standardized verbal encouragement (''You're doing well, keep up the good work'') was provided once every minute by the physiotherapist throughout the test.[21] This test was conducted indoors on a 46 m rectangular walking track. The surface was uncarpeted and marked at 1 m intervals. Test–retest reliability and SEM for the timed component of the test have been estimated at 0.94 (Type 2,1 ICC) and 26.29 m respectively in a sample of patients similar to those taking part in this study.[5]

Composite Pain Measures

It has long been recognized that increasing the number of items in an instrument is an effective strategy to reduce measurement error.[19,22] For this reason, we constructed four composite performance pain scores by summing the pain rating values from (1) all four performance measures; (2) SPW, ST, and TUG; (3) SPW and TUG; and (4) SPW, TUG, and 6MW. The rationale for the second and third composite measures was that a previous investigation has shown that a substantial number of patients are unable to complete the ST and 6MW tests several weeks after arthroplasty surgery.[14] The rationale for the fourth composite measure was that a previous study applying confirmatory factor analysis demonstrated that a model excluding the ST was better able to distinguish between the concepts of pain and function than one that included the ST.[4]

Analysis

Descriptive statistics, including means and SDs, were calculated for the pain scores, age, and body mass index (BMI); frequencies were determined by joint and gender for both the total sample ($N = 164$) and the reliability subsample ($n = 22$).

Using Minitab 14 statistical software (Minitab Inc., State College, PA), we applied the Anderson-Darling test for normality of the data, generated the ANOVA table and variance estimates for the reliability analyses, and obtained bootstrap estimates for the difference in reliability coefficients. We calculated relative (Type 2,1 ICC) and absolute (SEM) reliability coefficients and 95% CI for individual and composite pain rating scores. We also calculated the minimal detectable change at the 90% confidence level (MDC_{90}) as follows: $SEM \times 1.65 \times \sqrt{2}$.

We compared the reliability estimates for the composite performance pain rating scores to that of the WOMAC pain sub-scale. We applied a bootstrap approach because the variance of the difference in reliability coefficients obtained on the same patient sample cannot be calculated directly. Specifically, the bootstrapping consisted of generating 1,000 pair-wise reliability estimates for each composite pain scale and the WOMAC pain sub-scale by sampling with replacement from the 22 patients in the reliability sample. For each comparison, the difference in pair-wise bootstrap estimates of the reliability coefficients were obtained and ranked from lowest to highest. The 95% CI for the difference in reliability coefficients was represented by the 25th- and 975th-ranked values. We also used the bootstrap data to estimate the correlation between the reliability coefficients of WOMAC and the composite pain ratings. This information is required to estimate the sample size for formal hypothesis-testing studies of a potential difference in reliability coefficients.

Using SPSS 15 statistical software (SPSS Inc., Chicago, IL), we explored the relationships between composite and WOMAC pain scales by first plotting the data, then calculating Pearson's correlation coefficients, 95% confidence bands, and 95% prediction bands.

The sample size for this study was based on convenience—all patients from the larger study sample who had two preoperative assessments within 90 days of one another—and not on a formal sample-size calculation.

RESULTS

The total sample consisted of 164 patients (79 female, 85 male) with a mean (SD) age of 63.5 years (10.3) and a mean BMI of 30.3 kg/m^2 (5.2). Of these patients, 87 (46 female) were awaiting TKA and 77 (33 female) were awaiting THA. Within this sample, only 22 patients presented for a second preoperative assessment within 90 days. These 22 patients formed the sample for the test–retest reliability study, consisting of 10 women and 12 men with a mean age of 62.0 (10.0) years and a mean BMI of 29.5 (5.4). Of these patients, 9 (4 female) were awaiting TKA and 13 (6 female) were awaiting THA. The mean number of days between testing sessions was 62.5 (19.4). The Anderson-Darling test results were

Table 1 Mean (SD) Pain Scores for the Individual and Composite Measures

Measure	*Entire Sample (N = 164)*	*Reliability Sample (n = 22)*		*Reliability Sample* (n = 20)*	
		Time 1	*Time 2*	*Time 1*	*Time 2*
Self-paced walk (SPW)	3.5 (2.4)	3.6 (2.5)	3.2 (2.4)	3.6 (2.5)	3.4 (2.4)
Stair test (ST)	3.9 (2.7)	3.4 (2.3)	3.5 (2.5)	3.4 (2.4)	3.7 (2.5)
Timed up and go (TUG)	3.2 (2.6)	3.5 (2.5)	3.7 (2.7)	3.5 (2.6)	3.6 (2.8)
6 minutes walk (6MW)	5.2 (2.6)	5.1 (2.0)	5.5 (2.4)	5.1 (1.9)	5.4 (2.5)
Composite SPW, ST, TUG, 6MW	15.8 (9.3)	15.6 (8.7)	15.9 (8.6)	15.5 (8.8)	16.0 (8.8)
Composite SPW, ST, TUG	10.6 (7.1)	10.4 (7.0)	10.4 (6.7)	10.4 (7.2)	10.6 (6.9)
Composite SPW, TUG	6.7 (4.8)	7.0 (4.9)	6.9 (4.8)	7.0 (5.0)	6.9 (5.0)
Composite SPW, TUG, 6MW	11.9 (7.0)	12.2 (6.6)	12.4 (6.8)	12.1 (6.7)	12.3 (7.0)
WOMAC pain sub-scale	8.9 (3.5)	8.6 (2.6)	8.5 (3.7)	8.5 (2.6)	8.2 (3.6)

*These results represent the findings after removing two patients we believed had truly changed.

Table 2 Intraclass Correlation Coefficients (Type 2,1) and Standard Errors of Measurement by Pain Measure

Measure	*n = 22*			*n = 20*		
	$ICC^{*}_{2,1}$ (95% CI)	*SEM** (95% CI)*	*MDC^{***}_{90}*	*$ICC_{2,1}$ (95% CI)*	*SEM (95% CI)*	*MDC_{90}*
Self-paced walk (SPW)	0.78 (0.54, 0.90)	1.14 (0.88, 1.63)	2.65	0.88 (0.73, 0.95)	0.85 (0.64, 1.24)	1.98
Stair test (ST)	0.70 (0.41, 0.87)	1.33 (1.03, 1.90)	3.09	0.77 (0.51, 0.90)	1.19 (0.90, 1.74)	2.77
Timed up and go (TUG)	0.86 (0.70, 0.94)	0.97 (0.74, 1.38)	2.26	0.92 (0.83, 0.97)	0.72 (0.55, 1.05)	1.67
6-minute walk (6MW)	0.70 (0.40, 0.86)	1.22 (0.94, 1.75)	2.84	0.80 (0.57, 0.92)	1.00 (0.76, 1.46)	2.33
Composite SPW, ST, TUG, 6MW	0.82 (0.62, 0.92)	3.67 (2.82, 5.24)	8.54	0.92 (0.82, 0.97)	2.42 (1.84, 3.54)	5.63
Composite SPW, ST, TUG	0.80 (0.60, 0.92)	3.04 (2.34, 4.34)	7.07	0.89 (0.76, 0.96)	2.31 (1.76, 3.37)	5.37
Composite SPW, TUG	0.84 (0.66, 0.93)	1.96 (1.50, 2.79)	4.57	0.92 (0.82, 0.94)	1.41 (1.07, 2.06)	3.20
Composite SPW, TUG, 6MW	0.84 (0.66, 0.93)	2.70 (2.07, 3.85)	6.28	0.94 (0.86, 0.98)	1.66 (1.27, 2.43)	3.86
WOMAC Pain Subscale	0.57 (0.21, 0.80)	2.08 (1.60, 2.98)	4.84	0.66 (0.32, 0.85)	1.85 (1.41, 2.70)	4.30

*intraclass correlation coefficient
**standard error of measurement
***minimal detectable change

consistent with a normal distribution for all pain measures. A summary of the pain scores is reported in Table 1.

Table 2 presents a summary of the test–retest reliability analyses for the individual tests and composite pain scores. The estimated differences in reliability coefficients between the composite performance ratings of pain and the WOMAC pain rating were as follows: (1) composite (SPW, ST, TUG, 6MW) 0.25 (95% CI: −0.03, 0.58); (2) composite (SPW, ST, TUG) 0.23 (95% CI: −0.04, 0.55); (3) composite (SPW, TUG) 0.27 (95% CI: 0.01, 0.59); (4) composite (SPW, TUG, 6MW) 0.27 (95% CI: 0.01, 0.60). In each case, the difference in the point estimates of the reliability coefficients favours the composite performance rating; however, only the confidence intervals on composites 3 and 4 exclude the value of zero. The bootstrap estimates of the correlation between reliability coefficients are as follows: (1) WOMAC pain sub-scale versus composite (SPW, ST, TUG, 6MW) 0.26; (2) WOMAC pain sub-scale versus composite (SPW, ST, TUG) 0.32; (3) WOMAC pain sub-scale versus composite (SPW, TUG) 0.33; and (4) WOMAC pain sub-scale versus composite (SPW, TUG, 6MW) 0.31.

In addition to presenting the results for the entire reliability sample ($n = 22$), we also report the results of analyses omitting two of these patients ($n = 20$). The rationale for sharing the latter results is that the omitted patients displayed substantial change on both the WOMAC pain sub-scale and the performance pain scales, which we interpreted clinically as true change rather than measurement error. Table 2 also reports estimates of MDC_{90}. The interpretation of MDC_{90} is that 90% of truly stable patients will display random fluctuations equal to or less than this value.

The correlation coefficients between the WOMAC and the composite pain scores are as follows: sum of all four pain scores 0.62 (95% CI: 0.51, 0.70); sum of SPW, ST, and TUG pain scores 0.59 (95% CI: 0.48, 0.68); sum of SPW and TUG pain scores 0.57 (95% CI: 0.46, 0.67); sum of SPW, TUG, and 6MW pain scores 0.61 (95% CI: 0.50, 0.70). Figures 1–4 display scatter plots of the data, regression lines, 95% confidence bands, and 95% prediction bands. The confidence bands, represented by the narrow curved lines around the regression line, indicate where the mean WOMAC pain score is likely to lie for a given composite pain rating value. The prediction bands, represented by the wider curved lines around the regression line, indicate where an individual's WOMAC pain score is likely to lie for a reported composite pain score.

DISCUSSION

In the evolution of instrument validation, investigations often proceed from parameter-estimation to

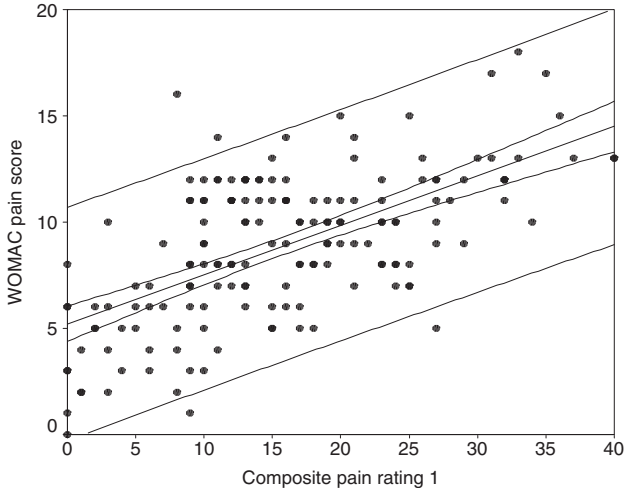

Figure 1 Scatter plot with regression line and 95% confidence (narrow curved lines) and prediction bands (wide curved lines) for WOMAC pain sub-scale and composite of all four functional pain measures

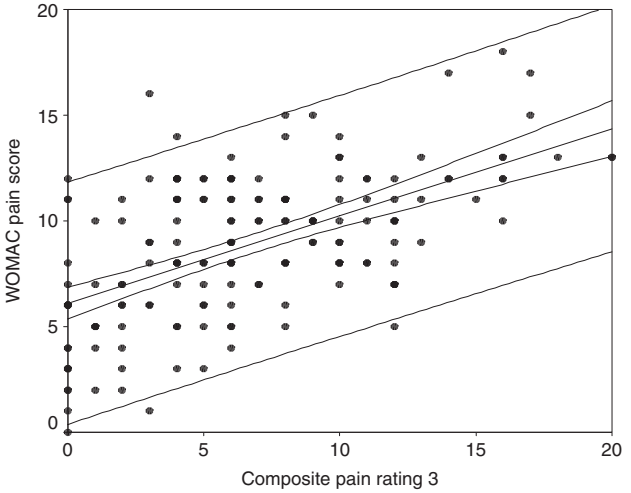

Figure 3 Scatter plot with regression line and 95% (narrow curved lines) and prediction bands (wide curved lines) for WOMAC pain sub-scale and composite of two functional pain measures (self-paced walk, TUG)

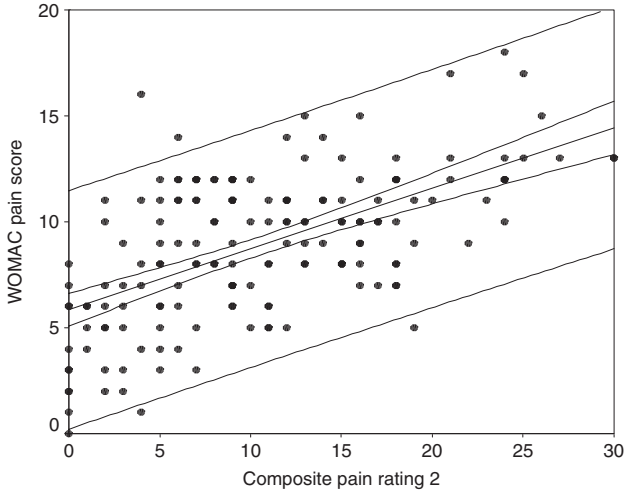

Figure 2 Scatter plot with regression line and 95% confidence (narrow curved lines) and prediction bands (wide curved lines) for WOMAC pain sub-scale and composite of three functional pain measures (self-paced walk, stair test, TUG)

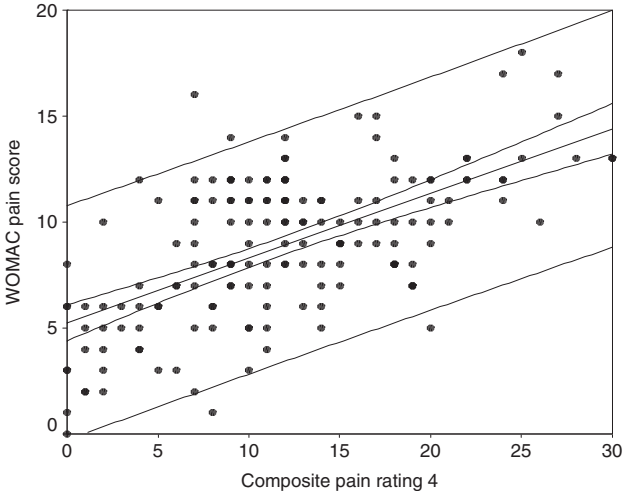

Figure 4 Scatter plot with regression line and 95% confidence (narrow curved lines) and prediction bands (wide curved lines) for WOMAC pain sub-scale and composite of three functional pain measures (self-paced walk, TUG, 6-minute walk)

hypothesis-testing studies. Parameter-estimation studies help determine whether the measurement properties of the instrument are ''in the ballpark'' of competing measures and provide investigators with data useful in calculating sample-size estimates for subsequent hypothesis-testing studies. The purposes of our study were (1) to estimate the test–retest reliability of performance-specific assessments of pain, (2) to estimate the test–retest reliability of the WOMAC pain sub-scale, (3) to estimate the difference in reliability between the composite performance-specific pain ratings, and (4) to examine the association between the composite performance-specific pain scores and the WOMAC pain scores in patients awaiting THA or TKA as a result of OA.

In order to be clinically useful, an instrument must be capable of discriminating among patients and display small errors of measurement. The ICC provides a representation of the extent to which an instrument is capable of discriminating among patients, while the SEM reports the measurement error in the same units as the original measurement. In our study ($n = 22$), the ICCs for individual performance-specific pain ratings varied from 0.70 to 0.86 and the SEMs varied from 0.97 to 1.33 pain points. These ICC and SEM estimates are consistent with values previously reported for patients with a spectrum of musculoskeletal conditions.[20]

Our estimate of the test–retest reliability for the WOMAC pain sub-scale, 0.57, is lower than that

reported previously. Typically, literature-based estimates of test–retest reliability for the WOMAC pain sub-scale vary between 0.77 and 0.86.[16,17] A possible explanation for this difference is that literature-based estimates applied a retest interval of less than 22 days, whereas our retest interval was less than 90 days.

Our study also estimated the difference in reliability between the WOMAC pain sub-scale and three performance-specific composite pain ratings. The study sample and the interval between retest assessments for these comparisons were identical for all measures. Although the point estimates of reliability were substantially greater for the performance ratings of pain, only two CIs on the difference between instruments excluded the value of zero. A conservative interpretation of this finding is that one cannot be confident that the test–retest reliability of these instruments differs. However, the CIs on the observed differences are large, and this will be addressed below in the discussion of limitations and directions for future investigation.

In addition to estimating ICCs for performance-specific pain measures, we calculated the SEM as an estimate of measurement error in the same units as the original measurements. In clinical practice, the SEM is extremely useful, as it provides information concerning the extent to which a clinician can be confident in a measured value. Moreover, the SEM can also be applied to help a clinician determine whether it is likely that a patient has truly changed over time. To illustrate these two applications, we will consider a clinical vignette in which a patient reports a SPW pain score of 6 at the time of an initial assessment and a pain score of 3 at a follow-up assessment. The confidence in a measured value at a given point can be obtained by applying a multiple of the SEM. For example, in our study the SEM for the SPW pain score was 1.14, and a 90% CI on our patient's pain score of 6 can be obtained by multiplying the SEM by 1.65, the z-value associated with a 90% confidence value. Accordingly, rather than viewing the patient's initial SPW pain score as 6, one would conceptualize the patient's pain score to be $6 \pm (1.14 \times 1.65)$, or somewhere between 4 and 8. To answer the question, ''Is it likely that this patient has truly changed between assessments?'' one could apply information obtained from MDC_{90}. MDC_{90} represents the variability between measurements (e.g., between initial assessment and follow-up assessment) in patients whose true pain has not changed. As defined previously, it is obtained by multiplying SEM $\times$ z-value for 90% CI $\times \sqrt{2}$. In this example, MDC_{90} is calculated to be 2.66 pain points $(1.14 \times 1.65 \times \sqrt{2})$. The interpretation of MDC_{90} is that 90% of truly stable patients will display random fluctuations within 2.66 points in reporting their pain. Accordingly, a change of 3 points, as observed in the vignette patient, would be interpreted as a true reduction in pain.

The relationship between the WOMAC pain sub-scale scores and the composite pain scores appears to be linear (see Figures 1–4). The confidence and prediction bands provide additional information. As previously noted, the confidence bands represent where the mean WOMAC pain score is likely to lie for a given composite pain rating value. This is useful to researchers, who are examining a larger sample of patients, but not of particular use to clinicians, who deal with one patient at a time. The prediction bands, on the other hand, indicate where an individual's WOMAC pain score is likely to lie for a reported composite pain score. This is far more important for clinicians, but, as observed in this study (see Figures 1–4), it is far less accurate. The wide prediction bands indicate that the composite pain scores cannot accurately predict a WOMAC pain score for an individual patient and thus cannot be used interchangeably with the WOMAC pain sub-scale in evaluating pain outcomes.

There are several limitations associated with the present study. The first is that the sample size was one of convenience and not based on formal sample-size calculations. Consequently, the CIs on the reliability coefficients are somewhat larger than one would desire. However, the study's results can be used to provide sample-size estimates for subsequent research questions. One such question would address whether the reliability of a pain measure exceeds a particular value. For example, if one were interested in testing whether the reliability exceeds 0.65, the following approach could be taken: null hypothesis: $R \leq 0.65$; alternate hypothesis: $R > 0.65$; expected R from study $= 0.75$; Type I error probability $= 0.05$, 1-tailed; Type II error probability $= 0.20$; and two pain measures (test and retest). Given these assumptions, a sample size of 160 patients would be required.[23] A second research question could address whether the test–retest reliability of the composite performance pain assessment exceeds that of the WOMAC pain sub-scale. Applying the estimates from the current study, one might form the following assumptions: null hypothesis: the difference in reliability coefficients $d_r \leq 0.10$; alternate hypothesis: the $d_r > 0.10$ in favour of the composite pain measure; expected WOMAC and composite R-values of 0.65 and 0.80 respectively; Type I error probability $= 0.05$, 1-tailed; Type II error probability $= 0.20$; two pain measures (test and retest); the correlation between performance and WOMAC reliability coefficients is 0.26. Given these assumptions, a sample size of 90 patients would be required.[24]

A second potential limitation of this study is the interval between assessments. Conceivably, the random error for this interval may be greater than that for shorter periods. It is also possible that several patients in our study sample may have truly changed over the test–retest assessment interval, and this would have reduced the reliability estimates. We attempted to explore the

extent to which this may have influenced our results by removing two patients who, in our opinion, had truly changed clinically. As one would expect, ICCs increased and SEMs decreased for all pain measures.

CONCLUSION

The results reported here suggest that the test–retest reliability of performance-specific pain ratings applying an 11-point NPRS are comparable with published values for patients with orthopaedic conditions of the neck, low back, and extremities. Moreover, our findings demonstrate that the test–retest reliability of the performance-specific pain ratings is equal to or better than that of the WOMAC pain sub-scale. Finally, the level of association between the WOMAC pain sub-scale and the various performance-specific pain ratings suggests that the scores can be used interchangeably when applied to groups, but not when applied to individual patients.

KEY MESSAGES

What Is Already Known on This Subject

Historically, self-report measures of pain and physical function have not been successful in distinguishing between these important health concepts. Conversely, specific measures of actual physical performance have been more adept at making this distinction. Of the measures available to assess pain intensity, the NPRS is often preferred.

What This Study Adds

We examined the test–retest reliability of using NPRS to measure reported pain associated with specific physical performance tasks that require joint loading in individuals with osteoarthritis awaiting total joint arthroplasty. We examined also the extent to which these pain ratings related to the WOMAC pain sub-scale. Results demonstrated reliability for all measures consistent with previous estimates for a variety of musculoskeletal conditions. Point estimates of the difference in reliability between the WOMAC pain sub-scale and composites of the performance-specific pain ratings favoured the performance-specific tests, though confidence intervals indicated that they may not, in fact, differ. We also provide SEM and MDC_{90} values that can be used to determine the degree of confidence the clinician should have in a given score or to estimate the likelihood that a patient has truly changed over time. The wide confidence bands shown in Figures 1–4 indicate that a composite pain measure cannot be used interchangeably with the WOMAC pain sub-scale in an individual patient. Therefore, both sets of measures must be used in evaluating patient outcomes.

REFERENCES

1. Bellamy N, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, et al. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. J Rheumatol. 1997;24:799–802.
2. Terwee CB, van der Slikke RM, van Lummel RC, Benink RJ, Meijers WG, de Vet HC. Self-reported physical functioning was more influenced by pain than performance-based physical functioning in knee-osteoarthritis patients. J Clin Epidemiol. 2006;59:724–31.
3. Maly MR, Costigan PA, Olney SJ. Determinants of self-report outcome measures in people with knee osteoarthritis. Arch Phys Med Rehabil. 2006;87:96–104.
4. Stratford PW, Kennedy DM, Woodhouse LJ. Performance measures provide assessments of pain and function in people with advanced osteoarthritis of the hip or knee. Phys Ther. 2006;86:1489–96; discussion 1496–500.
5. Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord. 2005;6:3.
6. Bellamy N. Osteoarthritis clinical trials: candidate variables and clinimetric properties. J Rheumatol. 1997;24:768–78.
7. Guermazi M, Poiraudeau S, Yahia M, Mezganni M, Fermanian J, Habib Elleuch M, et al. Translation, adaptation and validation of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) for an Arab population: the Sfax modified WOMAC. Osteoarthr Cartilage. 2004;12:459–68.
8. Thumboo J, Chew LH, Soh CH. Validation of the Western Ontario and McMaster University Osteoarthritis Index in Asians with osteoarthritis in Singapore. Osteoarthr Cartilage. 2001;9:440–6.
9. Faucher M, Poiraudeau S, Lefevre-Colau MM, Rannou F, Fermanian J, Revel M. Assessment of the test–retest reliability and construct validity of a modified WOMAC index in knee osteoarthritis. Joint Bone Spine. 2004;71:121–7.
10. Kennedy D, Stratford PW, Pagura SMC, Wessel J, Gollish JD, Woodhouse LJ. Exploring the factorial validity and clinical interpretability of the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC). Physiother Can. 2003;55:160–8.
11. Parent E, Moffet H. Comparative responsiveness of locomotor tests and questionnaires used to follow early recovery after total knee arthroplasty. Arch Phys Med Rehabil. 2002; 83:70–80.
12. Stratford PW, Kennedy DM, Hanna SE. Condition-specific Western Ontario McMaster Osteoarthritis Index was not superior to region-specific Lower Extremity Functional Scale at detecting change. J Clin Epidemiol. 2004;57: 1025–32.
13. Messick S. Validity. In: Linn RL, editor. Educational measurement. 3rd ed. Phoenix, Arizona: ORYZ Press; 1993. p. 14.
14. Kennedy DM, Stratford PW, Hanna SE, Wessel J, Gollish JD. Modeling early recovery of physical function following hip and knee arthroplasty. BMC Musculoskelet Disord. 2006;7:100.

15. Bellamy N. Pain assessment in osteoarthritis: experience with the WOMAC Osteoarthritis Index. Semin Arthritis Rheu. 1989;18:14–7.
16. Roorda LD, Jones CA, Waltz M, Lankhorst GJ, Bouter LM, van der Eijken JW, et al. Satisfactory cross cultural equivalence of the Dutch WOMAC in patients with hip osteoarthritis waiting for arthroplasty. Ann Rheum Dis. 2004;63:36–42.
17. Salaffi F, Leardini G, Canesi B, Mannoni A, Fioravanti A, Caporali R, et al. Reliability and validity of the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index in Italian patients with osteoarthritis of the knee. Osteoarthr Cartilage. 2003;11:551–60.
18. Stratford PW, Kennedy DM, Woodhouse LJ, Spadoni GF. Measurement properties of the WOMAC LK 3.1 pain scale. Osteoarthr Cartilage. 2007;15:266–72.
19. Spadoni GF, Stratford PW, Solomon PE, Wishart LR. The evaluation of change in pain intensity: a comparison of the P4 and single-item numeric pain rating scales. J Orthop Sport Phys. 2004;34:187–93.
20. Stratford PW, Spadoni G. The reliability, consistency, and clinical application of a numeric pain rating scale. Physiother Can. 2001;53:88–91, 114.
21. Guyatt GH, Pugsley SO, Sullivan MJ, Thompson PJ, Berman L, Jones NL, et al. Effect of encouragement on walking test performance. Thorax. 1984;39:818–22.
22. Nunnally JC. Psychometric theory. Toronto: McGraw-Hill; 1978.
23. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17:101–10.
24. Stratford PW, Spadoni GF. Sample size estimation for the comparison of competing measures' reliability coefficients. Physiother Can. 2003;55:225–9.