

Reliability by Surgical Status of Self-Reported Outcomes in Patients Who Have Shoulder Pathologies

Karon F. Cook, PhD¹

Toni S. Roddey, PT, PhD, OCS, FAAOMPT²

Sharon L. Olson, PT, PhD³

Gary M. Gartsman, MD⁴

Franz Felix T. Valenzuela, PT, MS⁵

William P. Hanten, PT, EdD⁶

Study Design: A test-retest design was used to evaluate the reliability of the self-report sections of 4 shoulder pain and disability scales.

Objective: The objective of the study was to compare interitem consistency and test-retest reliability by surgical status (postoperative versus nonoperative) and to evaluate the effect of surgical status in the prediction of retest scores.

Background: Patients and healthcare providers evaluate shoulder status based on self-evaluations of pain and disability. Shoulder outcome measures have been developed that include self-reports, but the properties of these measures have not been assessed by surgical status.

Methods and Measures: A questionnaire containing self-report sections of 4 shoulder scales was administered to study participants twice with 1 week between administrations. The outcome measures examined were the: (1) University of California at Los Angeles (UCLA) Shoulder Score; (2) Constant-Murley Scale (CMS); (3) American Shoulder and Elbow Society (ASES) Shoulder Index; and (4) Shoulder Pain and Disability Index (SPADI).

Intraclass correlation coefficients (ICC) were calculated to estimate the test-retest reliability of each of the scales and subscales. The interitem consistencies of the multi-item subscales were assessed using Cronbach's alpha. The effect of surgical status on shoulder outcome scale reliability was evaluated using a general linear models approach.

Results: The interitem consistency estimates for the multi-item scales were high with both operative and nonoperative participants (0.88 to 0.96). With the exception of the satisfaction subscale of the UCLA Shoulder Score for the nonsurgical group, the estimated intraclass coefficients ranged from 0.51 to 0.91. The prediction of UCLA-satisfaction and ASES-disability, pain, and total retest scores was improved with the addition of surgical status into a regression model.

Conclusions: The examined scales exhibited good internal consistency across surgical status. The postsurgical sample's reproducibility estimates tended to be higher than those of the nonsurgical sample. Reliability of shoulder outcome scales can be affected by patient surgical status. *J Orthop Sports Phys Ther* 2002;32:336-346.

Key Words: *outcome assessment (healthcare), psychometrics, reliability, shoulder, validity*

In many patient populations, shoulder-related dysfunction is a common and debilitating health problem. For example, shoulder problems have been found to affect as many as 34% of persons 65 and older,⁴ 39% of persons with paraplegia,¹³ 78% of those with quadriplegia,²⁰ and 64% of those with a stroke.²² Shoulder dysfunction can result in substantial disabilities affecting a person's activities of daily living (eg, bathing, dressing, and toileting) and the ability to function independently (eg, ability to board public transportation).^{2-4,22} Though there are no published studies that directly estimate the yearly financial expenditures for

¹ Associate director for research, Houston VA Parkinson's Disease Research, Education, and Clinical Center, Houston, TX.

² Assistant professor, Texas Woman's University School of Physical Therapy, Houston, TX.

³ Associate professor, Texas Woman's University School of Physical Therapy, Houston, TX.

⁴ Orthopaedic surgeon, Fondren Orthopedic Group, and clinical associate professor, University of Texas Health Science Center, Houston, TX.

⁵ Advanced master's physical therapist, Houston, TX.

⁶ Professor, Texas Woman's University School of Physical Therapy, Houston, TX.

This study was approved by the Institutional Review Board, Texas Woman's University, Houston, TX. Funding was provided by Veterans Affairs Health Services Research and Development IIR 98-077-1 and HCA/Columbia and Texas Orthopedic Hospital Research Fellowship.

Send correspondence to Karon F. Cook, VAMC (153)/2002 Holcombe Boulevard, Houston, TX. E-mail: karonc@bcm.tmc.edu

shoulder-related dysfunction, the cost information that exists, coupled with prevalence data, suggests that the amount is quite large. For example, the costs associated with rotator cuff injury averaged more than \$50,000 per person in 1 study of persons who sustained shoulder injuries at work.¹⁹

Measures such as joint range of motion and manual muscle testing frequently are used to assess shoulder impairments. Shoulder status may also be assessed based on self-evaluations,¹⁷ and a number of shoulder outcome measures include, or are comprised entirely of, patient self-reports. There is wide range in the specificity of the outcomes measured by these scales. For example, the Rotator Cuff Quality of Life (RC-QOL) scale¹⁰ was designed to evaluate outcomes specifically related to rotator cuff pathologies. In contrast, the Disabilities of the Arm, Shoulder, and Hand (DASH)¹¹ scale and the Upper Extremity Function scale¹⁵ were developed to measure a wider range of upper extremity musculoskeletal conditions.

The current study compares the patient self-report portions of several scales that were developed to measure shoulder outcomes specifically, without reference to particular shoulder conditions: (1) University of California at Los Angeles (UCLA) Shoulder Score,⁷ (2) Constant-Murley Scale (CMS),⁶ (3) American Shoulder and Elbow Society (ASES) Shoulder Index,¹⁶ and (4) Shoulder Pain and Disability Index (SPADI).¹⁷ These scales were chosen because, at the time of data collection, they were among the more frequently employed in studies examining outcomes in patients with shoulder pathologies. The purposes of the study were to evaluate the reliability of the patient-report sections of the 4 scales and to compare the scale reliability estimates obtained for the subsamples of patients who had and had not had an operation on their affected shoulder.

METHODS

Subjects

Patients were recruited from the office of a private-practice orthopaedic surgeon who treats shoulder dysfunction. Criteria for participation in the study included being at least 18 years of age and being able to read and understand English. Patients who had undergone shoulder surgery were excluded if they were less than 8 weeks postsurgery. Patients were excluded if they had received a cortisone injection in their affected shoulder the day of their appointment with their physician or if changes or progression in their rehabilitation were scheduled to commence during the study period. The Institutional Review Board at Texas Woman's University approved the study proposal. Informed written consent was obtained from all participants.

Study participants completed a questionnaire that contained the self-report sections of the UCLA, CMS, ASES, and SPADI scales, as well as questions regarding surgical status and demographics. After completing the questionnaire, participants were given a blank second copy. They were asked to return this copy 1 week later in the provided preaddressed, stamped envelope. The time interval of 1 week between administrations was chosen because it was thought to be long enough that participants would be unlikely to recall their answers to questionnaire items and brief enough that substantial changes in status would be unexpected. To control for changes in patient status, the second copy of the questionnaire differed from the first by the inclusion of an item that asked patients whether their shoulder was "the same," "better," or "worse" compared to the previous week. Only the scores of participants who reported being "the same" were included in the test-retest analyses.

Measures

UCLA Shoulder Score In addition to subscales for strength and active range of motion, the UCLA Shoulder Score includes 3 single-item subscales designed to evaluate pain, function, and satisfaction with the shoulder (Appendix). In response to the pain and function items, patients circle a number from 1 to 10 to indicate level of shoulder pain or function.⁷ Higher scores indicate less pain and greater function. In response to the satisfaction item, patients indicate they are "satisfied and better" (5 points) or "not satisfied and worse" (0 points). Summed scores across the 3 self-report sections of the UCLA Shoulder Score range from 0 to 25. The instructions for completing the UCLA Shoulder Score do not specify for the patient a particular time frame to reference (eg, in the past week).

Constant-Murley Scale The CMS consists of 3 subscales. A clinician evaluates the patient's range of motion and strength, and patients report their pain levels (Appendix). The pain subscale was evaluated in this study. On this subscale, patients indicate whether their levels of pain are "none" (15 points), "mild" (10 points), "moderate" (5 points), or "severe" (0 points).⁶ As was the case for the UCLA, the CMS does not give a particular time frame for respondents to reference in completing the self-report.

ASES Shoulder Index The ASES Shoulder Index measures pain and function by self-report (Appendix). The pain subscale is a 1-item, visual analogue scale (VAS) approximately 10-cm long anchored by the descriptors, "no pain at all" (left anchor) and "pain as bad as it can be" (right anchor). The pain score is calculated by measuring the distance from the left anchor to the mark placed on the line by the subject, subtracting this distance from the length of the visual analogue, and multiplying by 5. The

pain subscale score can range from 0 to 50 points. Lower scores indicate higher pain.

The function subscale consists of 10 items, each describing a functional activity. Respondents report their levels of difficulty in completing each activity by circling 0 (unable to do), 1 (very difficult to do), 2 (somewhat difficult to do), or 3 (not difficult to do). Function subscale scores are obtained by multiplying the sum of the item scores by 5/3. As with the pain subscale, scores can range from 0 to 50, and higher scores indicate better function. The total ASES Shoulder Index ranges from 0 to 100 and is obtained by adding the scores of the pain and function subscales.¹⁶ For the pain scale, respondents are asked to mark their answers based on their "pain today." For the function scale, no specific time frame is given.

SPADI The SPADI is comprised of self-report items that assess shoulder pain and disability (Appendix). For these subscales, respondents are asked to answer with regard to their pain and disability "in the past week." The 5 questions of the pain subscale ask patients to report their level of pain performing 5 activities of daily living (ADL) by placing marks on 10-cm visual analogues. The analogues are anchored by the descriptors "no pain" (left anchor) and "worst pain imaginable" (right anchor). The 8 items of the disability subscale evaluate shoulder function during ADL. Respondents indicate their level of difficulty with each activity by placing a mark on 10-cm visual analogues that are anchored with the descriptors "no difficulty" (left anchor) and "so difficult required help" (right anchor). Item scores are calculated by measuring the distance in centimeters from the left anchor to the respondent's mark. To obtain the subscale scores, item scores are summed and then divided by the summed lengths of all the visual analogues to which the participant responded. This mean item score is then multiplied by 10. Scores on each of the 2 subscales range from 0 to 100. The SPADI total score also ranges from 0 to 100 and is obtained by averaging the pain and disability subscale scores. Higher scores indicate greater pain and greater disability.¹⁷

Data Analysis

Reproducibility We evaluated the reproducibility of the scales using only the scores from participants who completed both the initial and follow-up questionnaires and reported their shoulder status was "the same as last week." For descriptive purposes, we calculated, by surgical status, the average scores and standard deviations of scores for both the test and retest administrations of the questionnaire. In addition, we compared first administration subscale scores by surgical status. An independent samples *t*-test was computed for the ASES pain and for the SPADI subscale scores. For all other subscales, we

calculated the Mann-Whitney *U*. The Mann-Whitney *U* statistic is a nonparametric equivalent of the *t*-test that compares ranks in 2 samples to estimate the probability that the samples come from the same population.

Intraclass correlation coefficients (ICC) were calculated to estimate the test-retest reliability of each of the scales and subscales. The ICC (3,1)¹⁴ and their respective confidence intervals were obtained using SPSS for Windows, version 10.0.5 (SPSS Inc, Chicago, IL).

Interitem Consistency The interitem consistencies of the multi-item subscales were assessed using Cronbach's alpha. These calculations were made for the entire study population as well as separately for the surgical and nonsurgical groups.

Effect of Surgical Status on Scale Reliability The central question for this study was, "Do shoulder outcome scores differ in reliability between surgical and nonsurgical groups?" We evaluated this question using a general linear models approach. For each of the scales we compared the amount of variance in retest scores explained by 2 different regression models: (1) a restricted model in which first administration scale scores were the only predictor variable, and (2) a full model that included first administration scores and a dummy variable for surgical status. A restricted model will account for less variance than its corresponding full model, but the pertinent question is whether the decrease is statistically significant. We evaluated this question using an *F*-test of variance (R^2) that tests the statistical significance of change in R^2 value between full and restricted models. In the current study, the probability value obtained is an estimate of the likelihood that the magnitude of increase in R^2 obtained by adding surgical status to the regression model would have occurred by chance.

RESULTS

Demographics

Of 110 participants, 56 reported being about the same as the previous week; 11 reported being worse; and 21 reported being better. Of the remaining 22 participants, 2 did not respond to this question; 1 reported being the same, but did not complete the remainder of the second form; and 19 failed to return a second form. Table 1 reports demographic data and surgical history for the 110 participants who enrolled in the study and for the subgroup of 56 participants who returned a second questionnaire by mail and reported their shoulder to be "about the same as last week." Table 2 categorizes participants by diagnoses and surgical status.

Descriptive Data

For descriptive purposes, average scores and standard deviations of subscale scores by surgical status were calculated and are reported in Table 3 for patients who completed the questionnaire twice and reported that their shoulder was “the same” compared to enrollment. In addition, we compared first administration subscale scores by surgical status.

An independent samples *t*-test was computed for the ASES pain and for the SPADI subscale scores. For all other subscales, we calculated the Mann-Whitney *U*. As the results reported in Table 4 indicate, scores of the surgical and nonsurgical groups on the initial administration of the questionnaires were significantly different (at $\alpha < 0.05$) on UCLA satisfaction, CMS pain, ASES pain, and SPADI pain, disability, and total score.

TABLE 1. Number of subjects included in the study.

Subjects	Age (y) Mean \pm SD (range)	Sex		Surgical Status	
		Men	Women	Postsurgical	Nonsurgical
Initially Recruited (n = 110)	49.2 \pm 15.7 (18–78)	71 (65%)	39 (35%)	64 (58%)	46 (42%)
Test-Retest Analysis (n = 56)	49.2 \pm 15.7 (18–76)	34 (61%)	22 (39%)	31 (55%)	25 (45%)

TABLE 2. Number of subjects in each diagnostic group.

Groups	All Participants (n = 110)		Test-Retest (n = 56)	
	Postsurgical (n = 64)	Nonsurgical (n = 46)	Postsurgical (n = 31)	Nonsurgical (n = 25)
RC* tear	34	17	20	7
AC† arthritis	0	2	0	1
Instability	18	4	9	2
RC* strain	0	11	0	6
GH‡ arthritis	4	5	0	5
Impingement	4	5	0	3
MS§ strain	1	0	0	0
Adhesive cap	2	2	1	1
Fracture	1	0	1	0

* RC = rotator cuff

† AC = acromioclavicular joint

‡ GH = glenohumeral joint

§ MS = musculoskeletal

|| Adhesive cap = adhesive capsulitis

TABLE 3. Mean \pm standard deviation for each self-report scale used in this study. Data presented for initial test and retest by surgical status (n = 56).

Scale/Subscale	Postsurgical (n = 31)		Nonsurgical (n = 25)	
	Test	Retest	Test	Retest
UCLA*				
Pain	5.4 \pm 2.8	5.4 \pm 2.8	4.0 \pm 1.9	4.1 \pm 1.9
Function	6.2 \pm 2.3	6.0 \pm 2.3	5.1 \pm 2.3	4.8 \pm 2.1
Satisfaction	4.1 \pm 1.9	4.5 \pm 1.4	1.0 \pm 2.0	1.0 \pm 2.1
CMS†				
Pain	7.7 \pm 3.6	8.1 \pm 3.1	4.6 \pm 3.2	5.2 \pm 3.4
ASES‡				
Pain	13.5 \pm 13.0	14.0 \pm 12.5	21.5 \pm 13.5	25.0 \pm 12.0
Function	17.3 \pm 7.3	19.8 \pm 6.8	16.0 \pm 7.2	15.9 \pm 7.4
Total	65.7 \pm 22.7	66.4 \pm 20.7	55.1 \pm 22.9	50.9 \pm 21.0
SPADI§				
Pain	37.4 \pm 23.5	37.5 \pm 30.2	61.7 \pm 23.6	56.7 \pm 24.0
Disability	28.5 \pm 25.6	26.5 \pm 23.1	47.9 \pm 24.6	47.3 \pm 24.2
Total	32.7 \pm 23.5	31.7 \pm 22.8	54.8 \pm 22.9	52.0 \pm 22.9

* UCLA = University of California Los Angeles Shoulder Score

† CMS = Constant-Murley Scale

‡ ASES = American Shoulder and Elbow Society Shoulder Index

§ SPADI = Shoulder Pain and Disability Index

TABLE 4. Statistical comparison of first administration scale and subscale scores between the postsurgical and nonsurgical groups.

Scale/Subscale	Statistical Test	P Value
UCLA*		
Pain	Mann-Whitney <i>U</i>	0.06
Function	Mann-Whitney <i>U</i>	0.10
Satisfaction	Mann-Whitney <i>U</i>	< 0.001
CMS†		
Pain	Mann-Whitney <i>U</i>	< 0.001
ASES‡		
Pain	Independent Samples <i>t</i> -Test	0.03
Function	Mann-Whitney <i>U</i>	0.53
Total	Mann-Whitney <i>U</i>	0.09
SPADI§		
Pain	Independent Samples <i>t</i> -Test	< 0.001
Disability	Independent Samples <i>t</i> -Test	0.01
Total	Independent Samples <i>t</i> -Test	0.01

* UCLA = University of California Los Angeles Shoulder Score
 † CMS = Constant-Murley Scale
 ‡ ASES = American Shoulder and Elbow Society Shoulder Index
 § SPADI = Shoulder Pain and Disability Index

Reproducibility

Scale and subscale ICC are reported in Table 5. With the exception of the CMS pain subscale, the ASES function subscale, and the SPADI disability subscale, the calculated ICC values were higher for the postsurgical study population than for the nonsurgical participants. Many of the disparities between the postsurgical and nonsurgical samples were quite large. This was the case for the UCLA subscales, in particular. The values for the CMS pain subscale were similar in the 2 groups: 0.80 for postsurgical and 0.87 for nonsurgical. A valid ICC could not be calculated for the UCLA satisfaction subscale in the nonsurgical group secondary to the restriction in range of the data and the number of individuals who reversed their satisfaction rating between the 2 test administrations.

The 95% confidence intervals for the ICC were quite large. These results are displayed graphically (Figure). The confidence intervals overlap in all but 1 of the post- and nonsurgical comparisons. For example, even though there was substantial disparity between the ICC values obtained for the SPADI disability subscale by surgery status (0.57 for the postsurgical group and 0.84 for the nonsurgical group), the confidence intervals for these 2 ICC overlap indicating that the difference between them was not statistically significant.²¹ Despite the large confidence intervals, however, there were statistically significant differences between the postsurgical and nonsurgical ICC of the UCLA function subscale ($P < .05$).

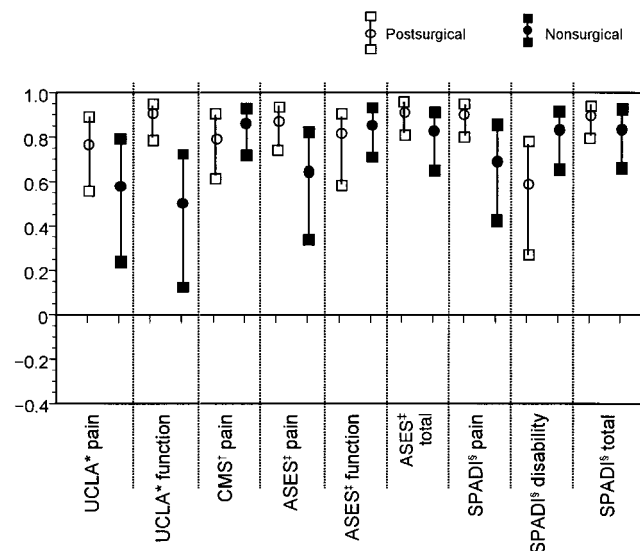
Interitem Consistency

Cronbach's alpha values for the SPADI pain subscale, SPADI disability subscale, and ASES function subscale were calculated for the first and the

TABLE 5. Test-retest reliability intraclass correlation coefficients (ICC) for subscales and scales.

Scale/Subscale	Surgical Status	
	Postsurgical (n = 31) ICC (95% CI*)	Nonsurgical (n = 25) ICC (95% CI*)
UCLA†		
Pain	0.78 (0.58–0.89)	0.59 (0.25–0.80)
Function	0.89 (0.78–0.94)	0.51 (0.14–0.75) ‡
Satisfaction	0.79 (0.59–0.89)	
CMS§		
Pain	0.80 (0.63–0.90)	0.87 (0.73–0.94)
ASES		
Pain	0.88 (0.76–0.94)	0.65 (0.35–0.83)
Function	0.78 (0.59–0.89)	0.86 (0.72–0.94)
Total	0.91 (0.82–0.96)	0.84 (0.66–0.92)
SPADI¶		
Pain	0.91 (0.81–0.95)	0.70 (0.43–0.86)
Disability	0.57 (0.27–0.77)	0.84 (0.66–0.92)
Total	0.91 (0.81–0.95)	0.84 (0.67–0.93)

* CI = confidence interval (lower bound-upper bound)
 † UCLA = University of California Los Angeles Shoulder Score
 ‡ An ICC value could not be calculated. See text for details.
 § CMS = Constant-Murley Scale
 || ASES = American Shoulder and Elbow Society Shoulder Index
 ¶ SPADI = Shoulder Pain and Disability Index



* UCLA = University of California Los Angeles Shoulder Score
 † CMS = Constant-Murley Scale
 ‡ ASES = American Shoulder and Elbow Society Shoulder Index
 § SPADI = Shoulder Pain and Disability Index

FIGURE. 95% confidence intervals of scale and subscale intraclass correlation coefficients (ICC) by surgical status.

second administrations. These are reported in Table 6 by surgical groups as well as for the combined sample. As indicated in the table, all alpha values were quite high; none was below 0.88.

Effect of Surgical Status on Scale Reliability

We evaluated the effect of surgical status on shoulder outcome scale reliability using a general linear

TABLE 6. Cronbach's alpha values for multi-item subscales.

Subjects	ASES* Function		SPADI† Pain		SPADI† Disability	
	Test	Retest	Test	Retest	Test	Retest
Postsurgical	0.91 (n = 23)	0.91 (n = 23)	0.90 (n = 24)	0.96 (n = 25)	0.94 (n = 24)	0.95 (n = 22)
Nonsurgical	0.89 (n = 23)	0.92 (n = 23)	0.88 (n = 25)	0.93 (n = 23)	0.92 (n = 25)	0.93 (n = 23)
Total	0.90 (n = 46)	0.91 (n = 46)	0.90 (n = 49)	0.95 (n = 48)	0.94 (n = 49)	0.95 (n = 45)

* ASES = American Shoulder and Elbow Society Shoulder Index

† SPADI = Shoulder Pain and Disability Index

models approach. Based on an *F*-test of variance, we compared the amount of variance in retest scores explained by a full model that included surgical status to that of a restricted model that did not include surgical status. The probability values obtained indicate whether there is a statistically significant increase in variance accounted for when surgical status is included in the model. Table 7 reports these results. Because R^2 values tend to overestimate model fit, we report adjusted R^2 values, which are corrected to more closely estimate the actual fit of the regression model. As indicated in Table 7, inclusion of surgical status in the regression model improved significantly the variance accounted for in UCLA satisfaction scores ($P = 0.05$) and in the ASES total and its 2 subscales ($P = 0.01$).

DISCUSSION

The examined scales exhibited good internal consistency in both the postsurgical and nonsurgical groups. Test-retest reliability estimates tended to be higher for the postsurgical group than for the nonsurgical group. There were exceptions to this trend (eg, SPADI disability subscale). Most of the calculated test-retest values were moderately high in both groups. Only the postsurgical SPADI disability ICC and the 3 nonsurgical UCLA subscale ICC were below 0.65. All others were in the range of 0.65 and 0.91. Very low values (0.51 to 0.59) were obtained for the UCLA scales in the nonsurgical group. This could be due in part to the fact that 1-item scales tend to be less reliable than multi-item scales;¹ though, in the postsurgical group, the UCLA pain, function, and satisfaction scales, which also are 1-item scales, had reasonably high test-retest values (0.78, 0.89, and 0.79, respectively).

Gupta and colleagues⁹ have critiqued the wording and format of the single-item UCLA satisfaction subscale. The item dichotomizes patient responses into the categories "satisfied and better" and "not satisfied and worse." Limiting patient responses to these 2 options results in only gross estimates of patients' levels of satisfaction. In addition, the multidimensionality of patient satisfaction is not well repre-

TABLE 7. Scale and subscale adjusted R^2 and *P*-value for *F*-test comparing differences in R^2 values between full and restricted models.

Scale/Subscale	Adjusted R^2		<i>P</i> Value
	Restricted	Full	
UCLA*			
Pain	0.56	0.56	0.77
Function	0.55	0.56	0.29
Satisfaction	0.53	0.57	0.05
CMS†			
Pain	0.75	0.75	0.49
ASES‡			
Pain	0.64	0.68	0.01
Function	0.66	0.70	0.01
Total	0.77	0.80	0.01
SPADI§			
Pain	0.73	0.74	0.35
Disability	0.55	0.58	0.07
Total	0.81	0.82	0.31

* UCLA = University of California Los Angeles Shoulder Score

† CMS = Constant-Murley Scale

‡ ASES = American Shoulder and Elbow Society Shoulder Index

§ SPADI = Shoulder Pain and Disability Index

sented with a single-item scale.¹⁸ These explanations, plausible with our inability to calculate a valid ICC value for the nonsurgical group, do not explain the higher ICC value (0.79) for the postsurgical group. A case-by-case comparison of test and retest satisfaction scores provides some insight. Of the persons who took both administrations of the satisfaction item, the majority of the postsurgical group responded that they were "satisfied and better" (24 of 27; 89%). Only 3 (11%) reported being "not satisfied and worse." Among these respondents, only 1 patient changed satisfaction response between administrations. In contrast, 5 of the 24 nonsurgical cases (21%) reported being "satisfied and better," and 19 reported being "not satisfied and worse." Eight patients changed their satisfaction responses between administrations. It may be that postsurgical patients on the whole were very satisfied and that nonsurgical respondents were closer to the threshold between being satisfied and not satisfied. If this were the case,

a small decrement in satisfaction among postsurgical patients would not cause them to endorse the alternative, "not satisfied and worse." Among those who were "borderline," however, reversals of responses between administrations would be more likely.

The satisfaction test-retest results obtained in the current study may be considered also in the context of the nature of self-reported satisfaction. We limited our analyses to persons who reported that their shoulders were "the same" in comparison to the previous week. However, constancy in shoulder status does not guarantee constancy in satisfaction with shoulder status. Research in cognitive psychology indicates that persons' expressed happiness with an outcome are related not only to how positive the outcome is, but also to how the outcome compares to a "counter-factual," the expected or alternative outcome a person imagines.⁸ It may be that patients judge their satisfaction with their shoulder by comparing their status to some anticipated outcome. For example, persons who expect their shoulder to improve may become dissatisfied if their shoulder condition remains the same.

Though values for the shoulder subscale ICC differed substantially, the confidence intervals for the ICC estimations were wide, and only 1 of the differences we observed were statistically significant at $\alpha < 0.05$. Estimation error and wide confidence intervals (CI) are not unique to the present study. For example, Roach and colleagues¹⁷ reported CI for the SPADI subscales. They calculated the 95% CI for the SPADI pain subscale to range from 0.40 to 0.80 and for the SPADI disability subscale to range from 0.41 to 0.80.

The strongest evidence for group differences in scale reliability was obtained by comparing a full regression model in which both first-administration scores and surgical status are used to predict retest scores to a restricted model in which surgical status was not taken into account. We found that, in the 3 ASES subscales and in the UCLA satisfaction subscale, prediction was improved by including surgical status. Explanations for group differences in the UCLA satisfaction scale have been offered above. We could not explain plausibly, however, the group differences in the 3 ASES scales. Interpreting this effect is particularly difficult since including surgical status in the regression model did not improve prediction with the other pain or function/disability scales.

Our results suggest that, with some shoulder outcome scales, test-retest reliability can differ by surgical status. This finding highlights an aspect of outcomes measurement not commonly appreciated; namely that the psychometric properties of a measure are sample- and purpose-dependent.¹⁹ Often, the use of a scale is justified by stating that the measure has been "found to be reliable and valid" in a previous study. Our data suggest that it is important to

determine whether the sample or purposes of a study are similar to those of the referenced study. The psychometric properties of a measure, however, can vary markedly depending upon the sample in which the measure is used. McHorney and colleagues¹² found the reliability of the scales of the MOS 36-item Short-Form Health Survey (SF-36) were quite different in different patient subgroups disaggregated by diagnosis and severity. For example, coefficient alpha values for the Social Function scale ranged from 0.72 to 0.90; for the General Health Perceptions scale, values ranged from 0.65 to 0.80. Clayton and Chubon⁵, summarizing studies of the reliability of the Life Situation Survey (LSS), reported Cronbach's alpha values as ranging "from the low 0.70s to the mid 0.90s" for most population samples.

There are a number of limitations to the current study. Chief among these is the lack of a gold standard for comparing estimates of shoulder outcome variables. Self-reported function, disability, and satisfaction may be compared across measure and across time, but there is no external referent by which the scales' external validity can be established. The small sample size also limited the current study. A larger pool of participants would have narrowed the confidence intervals around the calculated ICC values and may have broadened the conclusions that could be drawn from the results. Another weakness in the study is that we did not have control over when patients actually completed the second questionnaire. After completing the first questionnaire, participants were given a blank second copy and asked to return this copy 1 week later. It is possible that some patients completed the second questionnaire after less than the requested amount of time had elapsed.

We limited our test-retest analysis to persons who reported that they had experienced no change in the previous week. This was a necessary inclusion criterion since test-retest statistics are intended to evaluate the stability of scores in the absence of change in the trait being measured. Asking patients to report how well they are doing in comparison to the previous week, however, requires them not only to evaluate their current shoulder status, but also to remember their shoulder status of the previous week. We cannot be certain of the accuracy with which respondents were able to accomplish these tasks.

CONCLUSIONS

Our results have both clinical and methodological implications. The large ICC confidence intervals we obtained highlight both the impact of error on reliability estimation and the importance of reporting these intervals. Future studies should be conducted with larger sample sizes to improve the precision of test-retest reliability estimates.

The disparities we obtained in test-retest reliability estimates suggest that some of the self-report sections on the shoulder scales may be differentially reliable by surgery status. The extremely low test-retest coefficients obtained in the nonsurgical group for the UCLA pain, function, and satisfaction scores suggest that this scale may be inappropriate for use in that patient population. Future research should evaluate the effect of surgical status and other clinical variables on the psychometric properties of shoulder outcome measures.

ACKNOWLEDGMENTS

The authors of this study would like to thank the Rehabilitation Department of Texas Orthopedic Hospital for their support and to acknowledge the unwavering assistance of the employees of the Fondren Orthopedic Group with this project.

REFERENCES

- Anastasi A. *Psychological Testing*. New York, NY: Macmillan; 1988.
- Anderson HI. The epidemiology of chronic pain in a Swedish rural area. *Qual Life Res*. 1994;3:19-26.
- Bergstrom G, Bjelle A, Sorensen LB, Sundh V, Svanborg A. Prevalence of symptoms and signs of joint impairment at age 79. *Scand J Rehabil Med*. 1985;17:173-182.
- Chakravarty KK, Webley M. Disorders of the shoulder: an often unrecognised cause of disability in elderly people. *BMJ*. 1990;300:848-849.
- Clayton KS, Chubon RA. Factors associated with the quality of life of long-term spinal cord injured persons. *Arch Phys Med Rehabil*. 1994;75:633-638.
- Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. *Clin Orthop*. 1987;214:160-164.
- Ellman H, Hanker G, Bayer M. Repair of the rotator cuff. End-result study of factors influencing reconstruction. *J Bone Joint Surg Am*. 1986;68:1136-1144.
- Gulliksen H. *Theory of Mental Tests*. New York, NY: Wiley; 1950.
- Gupta R, Leggin BG, Iannotti JP. Results of surgical repair of full-thickness tears of the rotator cuff. *Orthop Clin North Am*. 1997;28:241-248.
- Hollinshead RM, Mohtadi NG, Vande Guchte RA, Wadey VM. Two 6-year follow-up studies of large and massive rotator cuff tears: comparison of outcome measures. *J Shoulder Elbow Surg*. 2000;9:373-381.
- Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med*. 1996;29:602-608.
- McHorney CA, Ware JE, Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Med Care*. 1994;32:40-66.
- Pentland WE, Twomey LT. Upper limb function in persons with long term paraplegia and implications for independence: Part I. *Paraplegia*. 1994;32:211-218.
- Portney LG, Watkins MP. Statistical measures of reliability. In: *Foundations of Clinical Research: Applications and Practice*. Norwalk, CT: Appleton & Lange; 1993.
- Pransky G, Feuerstein M, Himmelstein J, Katz JN, Vickers-Lahti M. Measuring functional outcomes in work-related upper extremity disorders. Development and validation of the Upper Extremity Function Scale. *J Occup Environ Med*. 1997;39:1195-1202.
- Richards RR, An KN, Bigliani LU, et al. A standardized method for the assessment of shoulder function. *J Shoulder Elbow Surg*. 1994;3:347-352.
- Roach K, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. *Arthritis Care Res*. 1991;4:143-149.
- Roush SE, Sonstroem RJ. Development of the physical therapy outpatient satisfaction survey (PTOPS). *Phys Ther*. 1999;79:159-170.
- Savoie FH, 3rd, Field LD, Jenkins RN. Costs analysis of successful rotator cuff repair surgery: an outcome study. Comparison of gatekeeper system in surgical patients. *Arthroscopy*. 1995;11:672-676.
- Silfverskiold J, Waters RL. Shoulder pain and functional disability in spinal cord injury patients. *Clin Orthop*. 1991;272:141-45.
- Sim J, Reid N. Statistical inference by confidence intervals: issues of interpretation and utilization. *Phys Ther*. 1999;79:186-195.
- Wanklyn P, Forster A, Young J. Hemiplegic shoulder pain (HSP): natural history and investigation of associated features. *Disabil Rehabil*. 1996;18:497-501.

Appendix

Self-Report Measures

*University of California at Los Angeles (UCLA) Shoulder Scale**

Please circle the one number that best fits for each category:

I. Pain

Present all of the time and unbearable; strong medication frequently	1
Present all of the time but bearable; strong medication occasionally	2
	3
None or little at rest, present during light activities; medication frequently	4
	5
Present during heavy or particular activities only; medication occasionally	6
	7
Occasional and slight	8
	9
None	10

II. Function

Unable to use limb	1
Only light activities possible	2
	3
Able to do light housework or most activities of daily living	4
	5
Most housework, shopping, and driving possible; able to do hair and dress and undress, including fastening brassiere	6
	7
Slight restriction only; able to work above shoulder level	8
	9
Normal activities	10

III. Satisfaction

Satisfied and better	5
Not satisfied and worse	0

Constant-Murley Scale (CMS)[†]

Please circle the best word that describes your pain:

None	15
Mild	10
Moderate	5
Severe	0

American Shoulder and Elbow Society (ASES) Shoulder Index[‡]

I. Pain

How bad is your pain today? (mark on line)^{||}

0	10
No pain at all	Pain as bad as it can be

II. Function

Circle the number that indicates your ability to do the following activities with your painful shoulder.

- 0 = unable to do
- 1 = very difficult to do
- 2 = somewhat difficult to do
- 3 = not difficult to do

1. Put on a coat	0	1	2	3	
2. Sleep on your painful side	0	1	2	3	
3. Reach up behind back	0	1	2	3	
4. Manage toileting	0	1	2	3	
5. Comb hair	0	1	2	3	
6. Reach a high shelf	0	1	2	3	
7. Lift 10 lbs. above shoulder	0	1	2	3	
8. Throw ball overhead	0	1	2	3	Have not tried
9. Do usual work	0	1	2	3	Have not tried
10. Do usual sport	0	1	2	3	Have not tried

Shoulder Pain and Disability Index (SPADI)^S

Mark on the line to show how much pain you have had in the past week for each question.^{||}

Example: No Pain _____ / _____ Worst Pain Imaginable

Pain Scale

A. How severe is your pain

1. At its worst?
No Pain _____ Worst Pain Imaginable
2. When lying on the involved side?
No Pain _____ Worst Pain Imaginable
3. When reaching for something on a high shelf?
No Pain _____ Worst Pain Imaginable
4. When touching the back of your neck?
No Pain _____ Worst Pain Imaginable
5. When pushing with the involved arm?
No Pain _____ Worst Pain Imaginable

Place a mark to show how much difficulty you have had in the past week to do each activity.^{||}

Example: No Difficulty _____ / _____ So Difficult Required Help

Disability Scale

B. How much difficulty did you have

1. Washing your hair?
No Difficulty _____ So Difficult Required Help
2. Washing your back?
No Difficulty _____ So Difficult Required Help
3. Putting on an undershirt or pullover shirt?
No Difficulty _____ So Difficult Required Help
4. Putting on a shirt that buttons down the front?
No Difficulty _____ So Difficult Required Help

5. Putting on your pants?

No Difficulty _____ So Difficult Required Help

6. Placing an object on a high shelf?

No Difficulty _____ So Difficult Required Help

7. Carrying a heavy object of 10 pounds or more?

No Difficulty _____ So Difficult Required Help

8. Removing something from your back pocket?

No Difficulty _____ So Difficult Required Help

* Reproduced with permission of The Journal of Bone and Joint Surgery.⁷

† Reproduced with permission of Clinical Orthopaedics and Related Research.⁶

‡ Reproduced with permission of The Journal of Shoulder and Elbow Surgery.¹⁶

§ Reproduced with permission of Arthritis Care Research.¹⁷

|| Lines should be 10 cm in length.