Access this article online
Website:
www.internationalshoulderjournal.org
DOI:
10.4103/0973-6042.94307
Quick Response Code:

# Validation of the Stanmore percentage of normal shoulder assessment

Ali M. Noorani, David J. S. Roberts, Alex A. Malone[1], Tim S. Waters[2], Anju Jaggi,
Simon M. Lambert, Ian Bayley

Department of Shoulder and Elbow, Royal National Orthopaedic Hospital, Brockely Hill, Stanmore, Middlesex, HA7 4LP, [1]Department of Orthopaedic, Christchurch Hospital, Riccarton Avenue, Christchurch, New Zealand, [2]Department of Orthopaedic, Hemel Hempstead Hospital, Hillfield Road, Hemel Hempstead, Hertfordshire, HP2 4AD, UK

**Address for correspondence:**
Mr. Ali M. Noorani,
119 Clarence Gate Gardens,
Glentworth St, London, NW1 6AL.
E-mail: alinoorani@yahoo.com

## ABSTRACT

**Background and Purpose:** The Stanmore Percentage of Normal Shoulder Assessment (SPONSA) is a patient-reported outcome measure (PROM). The score assesses pain, range of movement, strength, stability and function of the shoulder. The aim of this work was to formally validate the SPONSA.
**Materials and Methods:** Validation of this score was carried out by measuring reproducibility, construct validity and sensitivity to change. Time to completion was also recorded. The Oxford Shoulder Score (OSS) and Constant Score (CS) were used for comparison. These assessments were performed with 61 individuals undergoing shoulder interventions.
**Results:** There was excellent preoperative reproducibility in both intra- and inter-observer groups. The SPONSA had a 0.79 correlation with the OSS and 0.78 with the CS. The overall effect size of the SPONSA was 0.72, which was comparable to OSS (0.65) and greater than CS (0.34), implying equal or better sensitivity to change.
**Conclusions:** The SPONSA is practical and quick to perform and also a reproducible and a sensitive instrument. This simple PROM is a commendable addition to the existing validated scoring methods for the shoulder.
**Level of Evidence:** I; testing of previously developed diagnostic criteria on consecutive patients (with universally applied reference "gold" standard).

**Key words:** Assessment, outcome, score, shoulder, validation

## INTRODUCTION

Numerous shoulder assessment systems are used as tools for clinical practice and research.[1-7] A percentage of normal assessment tool is used as a simple patient-reported outcome measure (PROM) for the shoulder at our institution and the use of similar tools has been described by other authors.[8-12] Such a tool is simple and rapid to perform and provides a useful adjunct to the existing shoulder assessments. This study aims to formally validate a percentage subjective shoulder assessment tool by assessing reproducibility, construct validity and sensitivity to change.

## MATERIALS AND METHODS

A single-question, the Stanmore Percentage of Normal Shoulder Assessment (SPONSA), for obtaining a percentage outcome score for the shoulder has been developed by the senior author through discussion with patients, surgeons and shoulder physiotherapy colleagues. It asks about symptoms most noted to affect the quality of life of individuals with shoulder complaints. It, therefore, includes statements about pain, range of movement, strength, stability and function. The following script is read to the patient and a verbal response obtained:

"A normal shoulder is one which is pain-free, with a full range of movement, normal strength and stability, and allows you to do what you feel your shoulder, if normal, should allow you to do. A normal shoulder is scored as 100 percent, while a completely useless shoulder is scored as 0 percent. Overall where would you rate your shoulder between 0 and 100 percent, at this present time?"

Please cite this article as: Noorani AM, Roberts DJ, Malone AA, Waters TS, Jaggi A, Lambert SM, *et al*. Validation of the Stanmore percentage of normal shoulder assessment. Int J Shoulder Surg 2012;6:9-14.

Pre-testing was undertaken to ensure the question was understood. Ten patients with shoulder complaints were asked the question and they reported no difficulty in comprehension. In order to assess the validity of our tool, we analyzed the scores of patients undergoing treatment.

Established outcome tools were used for comparison to determine validity. The assessments used for comparison had to be validated shoulder scores that were in common use across Europe. The Constant Score (CS) is partly subjective (patient-based), with questions about pain and activities of daily living, and partly objective (clinician-based) involving examination of range of movement and power measured using a spring balance or dynamometer. The CS[2] has been validated through repeated use and is the only outcome tool recommended by the European Society for Surgery of the Shoulder and the Elbow (ESSSE). The Oxford Shoulder Score (OSS)[1] is a validated, widely used PROM (Patient-Reported Outcome Score) with subjective questions about pain and activities of daily living. All three scores were administered in English. Although all patients had good command of the English language, we had provisionally planned to exclude anyone who required translation.

Sixty-one consecutive patients were recruited at the preoperative assessment clinic at our institution. All patients had previously undergone clinical assessment and were to undergo shoulder intervention under the care of the senior authors. All patients were included irrespective of age, gender, pathology and type of intervention. Patients were assessed at three separate intervals. Interview A was conducted in the preoperative assessment clinic (2–3 weeks pre-operatively), interview B immediately preoperatively and interview C at 3–6 months postoperative outpatient review. Interview B was conducted in two ways: for group B1, assessment was performed by the same observer to assess (test–retest) intra-observer reliability, while group B2 was assessed by different observers at the two interviews to assess inter-observer reliability.

At interviews A and C, the patients completed an OSS, answered the Stanmore (SPONSA) question and had the CS performed by one of three independent assessors (DR, AN, TW). The time taken to perform each test was recorded. Before interview B (immediately preoperatively), the patients were randomized into two groups. Randomization was by a computer-generated list indicating if the patient was to be assessed by the same observer as in interview A or by a different observer. The randomization process was carried out by a research physiotherapist who ensured allocation concealment and informed the assessors of their allocated interview B on the morning of the task. At interview B, all patients were asked if there had been any change in their symptoms since interview A at the preoperative assessment clinic. The purpose of interview B was to collect data to test reproducibility of the SPONSA score. It was not expected that the subject's symptoms would change significantly in the weeks between interview A and B, but for the purposes of testing reproducibility, if a change in symptoms was reported (e.g. "my shoulder is more/less painful"), then that subject's data were excluded from calculations to assess the reproducibility of the score. Data collected from that subject and interviews A and C were still used for the purposes of other statistical analyses.

At interview C, the SPONSA, OSS and CS were administered and performed by the same assessor as in interview A. Patients were also asked if the operation or intervention was successful or not (yes or no response).

## Statistical methods

A combination of SPSS™ (IBM Corp., Armonk, NY, USA) and Microsoft Excel™ (Microsoft, Redmont, WA, USA) was used for statistical analysis.

### *Reproducibility*

We assessed test–retest reliability of the SPONSA by measuring and comparing the SPONSA scores at points A and B. Before SPONSA B, all patients were asked if there was a change in their shoulder symptoms. Only those who reported no change were included so that we could measure the true reproducibility of the score. Reproducibility was assessed using the method described by Bland and Altman.[13] Those patients with scores collected by the same observer at interviews A and B (group B1) were analyzed to provide information about the patient's consistency of rating or the intra-observer reproducibility. The influence of the assessor or the inter-observer reproducibility was determined by comparing scores for patients assessed by a different observer (group B2). Differences in scores between points A and B of less than 2 standard deviations from the mean difference represent good agreement.

### *Construct validity*

The Pearson Correlation Coefficient was calculated to determine the correlation between the SPONSA and the OSS and CS. The correlation of the SPONSA score with the OSS and CS was calculated for data collected preoperatively (point A) and postoperatively (point C). When calculating the correlation between the SPONSA and OSS, the 10 subjects with shoulder instability were excluded from calculations as the OSS is not validated for assessing instability. Significance level was set at $P<0.05$.

### *Sensitivity to change*

Sensitivity to change was assessed by calculating the effect size (z-score) for the SPONSA, OSS and CS. Effect size is a means of measuring the extent of change detected by outcome tools.[14] It is calculated by dividing the difference between the mean preoperative and postoperative scores by the standard deviation of the preoperative scores. An effect size of 1.0 is equivalent to a change of 1 standard deviation in the sample. When calculating the effect size for the OSS, the 10 subjects with shoulder instability were excluded. When calculating the effect size, we firstly included all preoperative and postoperative scores for the SPONSA, OSS and CS. As a second analysis for the effect size, we split the data into two groups according to the answer to the question, "Was your operation or intervention successful or not" (yes or no response).

*Binary logistic regression analysis*

It was used to determine the odds ratio of true success for a change in 5, 7, 15, 20, 25 and 30 points (%) in the SPONSA between preoperative and postoperative analysis.

Time taken to perform the SPONSA was compared to the OSS and CS for each subject using a paired *t*-test.

We aimed to recruit 60 patients as this number was used to validate the OSS and our methods for validation were to be similar. It was calculated that this sample size gave us 80% power to detect an effect size of 0.8 or greater at the 5% level.[15]

## RESULTS

Sixty-one consecutive subjects were recruited. Fifty-five underwent the planned treatment. Four procedures were postponed as planned preoperative medical optimization was not fully completed and two procedures were cancelled due to symptomatic improvement. These patients did undergo assessment at point B, so they could be included in determining reproducibility of the SPONSA. Data collected at point A from the six patients who did not undergo treatment could not be used for calculation of construct validity or effect size. The diagnosis and operative procedures of the patients are summarized in Table 1.

Overall, 11 patients were not assessed at point C. This was due to unavailability of patient to attend the final postoperative follow-up assessment in the research clinic. The overall follow-up rate was 82%.

The differences in the scores between points A and B were plotted against the mean difference and 2 standard deviations in Figures 1 and 2. The mean score difference for those patients assessed by the same observer was 1.42 and for those assessed by different observers was −1.1. These estimated means of score difference were not significantly different from 0. Bland and Altman have described that good reproducibility (test–retest reliability) is represented if 95% of the individual differences

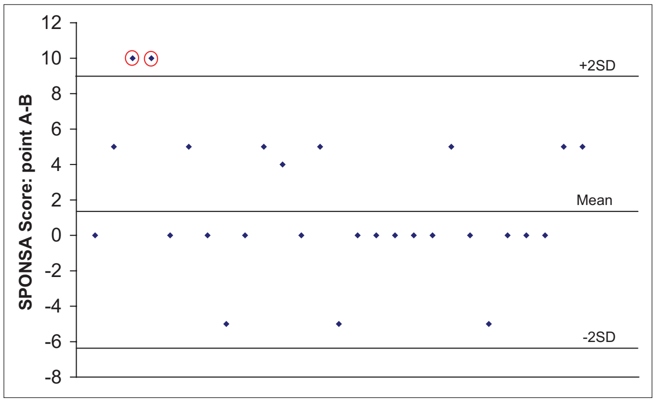

**Figure 1:** Intra-observer reliability of the SPONSA. SPONSA performed by observer 1 at points A and B (n=31). Bland and Altman plot of scores for intra-observer reliability showing that only two subjects in each group were outside 2 standard deviations of the mean difference and 93% of the subjects were within 2 standard deviations

**Table 1: Diagnoses and procedures**

| Diagnosis | No. | Procedure |
|---|---|---|
| Osteoarthritis | 13 | 11 Primary TSA |
| | | 2 Postponed |
| Rheumatoid arthritis | 4 | 3 CADCAM TSA |
| | | 1 Postponed |
| Revision arthroplasty | 11 | 2 First stage revision TSA |
| | | 3 Second stage revision CADCAM TSA |
| | | 1 Single stage revision CADCAM TSA |
| | | 1 Reduction of displaced prosthetic humeral head |
| | | 2 Single stage revision to Bayley-Walker TSA |
| | | 1 Single stage revision to unconstrained TSA |
| | | 1 Postponed |
| Subacromial impingement | 11 | Arthroscopic subacromial decompression |
| Instability | 10 | 5 Bankhart repair |
| | | 1 Broca/Bankhart repair |
| | | 3 Diagnostic arthroscopy and specialist physiotherapy |
| | | 1 Rotator cuff reconstruction |
| Rotator cuff tear | 7 | 5 Open rotator cuff reconstruction |
| | | 2 Symptoms improved, non-operative management continued |
| Non-union of fracture | 3 | 1 CADCAM distal humerus replacement |
| | | 1 Plate and bone graft to clavicle |
| | | 1 Removal of Philos plate and insertion of unconstrained TSA |
| Adhesive capsulitis | 2 | MUA and arthroscopy |

TSA – Total shoulder arthroplasty; CADCAM – Computer aided design computer aided manufacture; MUA – Manipulation under anaesthetic between the two measurements lie within 2 standard deviations of the mean difference in scores. Scores of only two subjects in each group were outside 2 standard deviations of the mean, which is equivalent to 93% of subjects (intra-observer group 2/30, inter-observer group 2/31) within 2 standard deviations. In these two cases, the answers varied by 10 points only. In 93% of the cases, the score differences were 0±5 points. Our findings represent good reproducibility.

The Pearson Correlation Coefficient (r) was calculated to assess correlation of the SPONSA with OSS and CS with both preoperative and postoperative scores. Perfect correlation is indicated if r=1 or −1, and in the case of no correlation, r=0. As the OSS is a negative score, increasing in value with worsening shoulder function, a negative coefficient implies good correlation. The overall correlation of SPONSA with OSS was r=−0.79 ($P$<0.001) and the overall correlation of SPONSA with CS calculated was r=0.78 ($P$<0.001). This trend is summarized in Figures 3 and 4.

The effect size (representing sensitivity to change between preoperative and postoperative scores) of the SPONSA was 0.72 ($n$=44), OSS was −0.65 ($n$=44) and CS was 0.34 ($n$=34). These results demonstrate comparable sensitivity to change to the OSS and greater sensitivity to change than the CS.

We examined the relationship between responses when individuals were asked whether surgery was a success or not and the effect size. For those who reported the operation was a success, the mean increase in SPONSA was 29 points (SD 24 points) and the effect size for change in SPONSA score was 1.2 and for OSS was −1.1. For those reporting the operation to be unsuccessful, the mean SPONSA score dropped as expected by 7 points (SD 24 points) and the effect sizes for SPONSA and OSS were −0.24 and 0.33. In this particular aspect, the SPONSA was also comparable to OSS. It also showed that the SPONSA is very sensitive to change after successful treatment and not very sensitive to change if the treatment was unsuccessful.

Binary logistic regression analysis was used to determine the odds ratio of true success. These values are summarized in Table 2. So, when a SPONSA score increases by 10 points, the chances of that change representing a successful treatment as opposed to unsuccessful treatment are 2.7:1 (95% CI 1.4–5.1).

The mean time to complete the SPONSA was 31.9 s (SD 7.2 s), OSS 129.5 s (SD 40.6 s) and CS 163.3 s (SD 41.5 s) [Figure 5]. Time to perform the SPONSA was significantly less compared to the OSS and the CS ($P$<0.0001 for both).

## DISCUSSION

A recent review identified 44 outcome scores for the shoulder.[16] Many are used inappropriately or modified and not tested for validity, reproducibility or sensitivity to change.

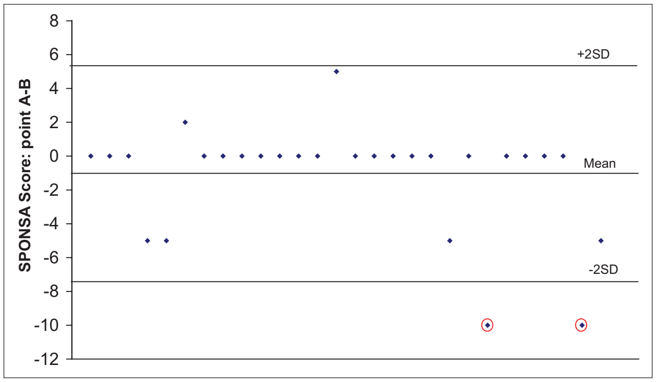

**Figure 2:** Inter-observer reliability of the SPONSA. SPONSA performed by observer 1 at point A and observer 2 at point B (n=30). Bland and Altman plot of scores for inter-observer reliability showing that only two subjects in each group were outside 2 standard deviations of the mean difference and 93% of the subjects were within 2 standard deviations

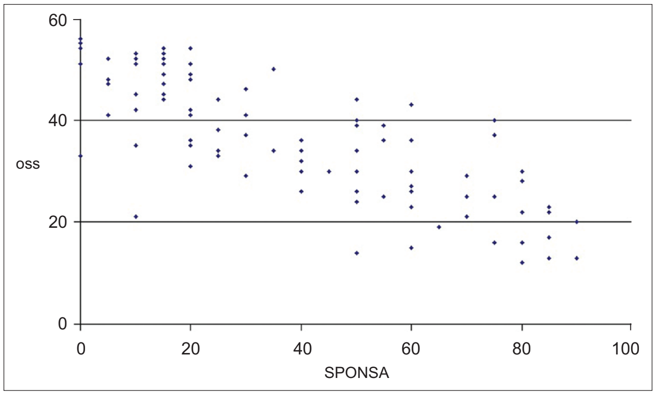

**Figure 3:** Correlation of SPONSA with Oxford Shoulder Score. Combined pre- and post-op treatment. Pearson correlation coefficient=–0.79 (n=105). A graphical representation of the correlation between the SPONSA and the Oxford Shoulder Score

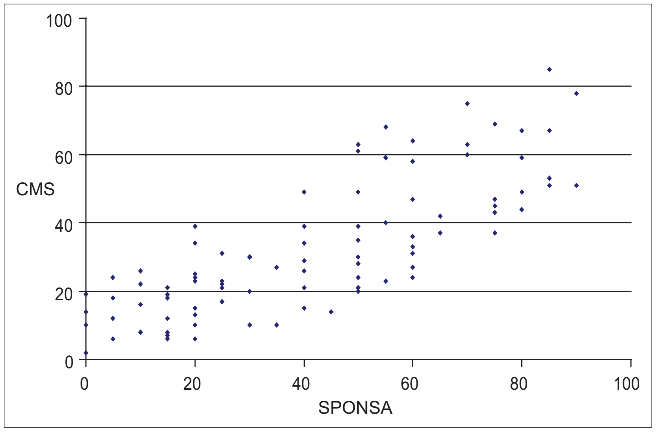

**Figure 4:** Correlation of SPONSA with Constant Score. Combined pre- and post-treatment. Pearson correlation coefficient=0.78 (n=95). A graphical representation of the correlation between the SPONSA and the Constant Score

Full validation of a scoring system is essential before it can be recommended for clinical or research use. There remain methodological inconsistencies and difficulties with some widely used scores.[6]

**Table 2: Percentage increase in SPONSA score after treatment and the odds ratio of successful treatment**

| Percentage increase in SPONSA score | Odds ratio for successful treatment (95% confidence intervals) |
|---|---|
| 5 | 1.6 (1.2–2.3) |
| 7 | 2.0 (1.3–3.1) |
| 10 | 2.7 (1.4–5.1) |
| 15 | 4.4 (1.7–11.5) |
| 20 | 7.2 (2.0–26.0) |
| 25 | 11.9 (2.4–58.8) |
| 30 | 19.5 (2.9–133.8) |

SPONSA – Stanmore percentage of normal shoulder assessment

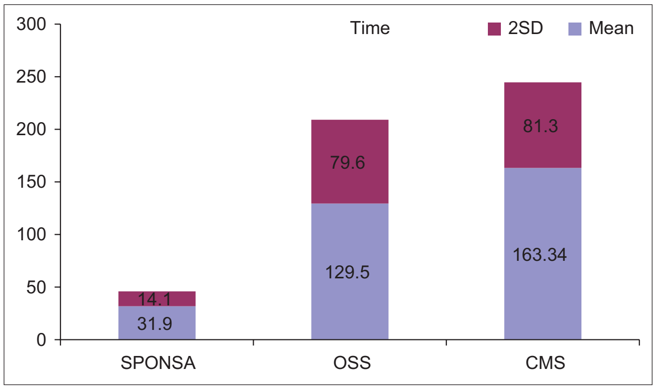

**Figure 5:** Time to complete scores: The time to complete the SPONSA, OSS and CM scores with 2 SD

The Clinical Effectiveness Unit of The Royal College of Surgeons of England recommends subjective, patient-based outcome measures (PROMs) in the assessment of surgical outcomes. Such measures minimise the confusion of which physiological components are to be assessed and it is considered that the patient's perceived health status and quality of life are the most important outcomes in surgery besides mortality.[17]

We have demonstrated a verbally administered percentage of normal outcome assessment to be quick, reliable, valid construct which is sensitive to change. It is reproducible if performed by either the same or different observers. We have also demonstrated good correlation of the SPONSA with the existing scoring systems.

Good correlation of subjective percentage shoulder scoring with the existing shoulder assessment tools has also been demonstrated by other authors.[9,10] Williams *et al.* have demonstrated good correlation with the Rowe[18] and Society of American Shoulder and Elbow Surgeons[5] outcome scores in young patients following surgery for shoulder instability.[9]

Other widely used shoulder outcome measures are well-established tools in research and clinical practice. The more detailed information that is obtained with multi-point scores may be useful to the clinician in guiding treatment. However, a significant number of these scores are time consuming and require special equipment. The existing subjective, patient-based assessment tools vary in their complexity. The Disabilities of the Arm Shoulder and Hand outcome measure is a 30-point questionnaire on common daily activities and pain, with answers on a 1–5 scale.[19] The Shoulder Pain and Disability Index (SPADI) is a 13-point questionnaire with sections on pain (5) and the performance of everyday tasks (8).[20] The OSS is a validated subjective tool.[3] Four of 12 questions are specifically about pain. Tools for assessment of shoulder instability such as the Western Ontario Shoulder Instability Index (WOSI), a 21-point questionnaire with visual analogue scale response, are also fully validated. As no clinical assessment is required, these subjective assessments may be completed by the patient alone as a paper questionnaire.

The question posed to the patient when performing the SPONSA asks about subjects, similar to other outcome scores, i.e. pain, range of movement, strength, stability and function. We speculate the SPONSA is very sensitive to change because unlike other scores, we do not artificially allocate a specific weight to each of these symptoms. Most outcome tools score components of shoulder morbidity (e.g. pain, range of movement, ability to perform daily tasks) individually, summing them to reach a final score. Careful item identification, reduction, weighting of items and testing in the development of a questionnaire as a shoulder outcome tool, demonstrated by Kirley *et al.*,[7] must be undertaken in the development of a reliable, responsive shoulder questionnaire. In such a system with multiple questions, each carries a fixed weight in the overall score. However, one particular shoulder symptom may have a larger effect on the quality of life as perceived by the individual. There will, therefore, be a "ceiling" to the overall score for the most severe symptom regardless of the overall real morbidity. This is one of the reasons why the CS and the OSS perform poorly for shoulder instability. The SPONSA is a percentage of normal score that allows the patient to weight the score to reflect their most problematic symptoms, therefore allowing it to be more sensitive to change after successful treatment. The odds ratio in Table 2 is a useful indicator of the chance of a successful treatment for any given percentage increase in SPONSA score.

A potential weakness of this study is that our sample contained no subjects with normal shoulders. If an individual experiences no shoulder symptoms, perhaps it is reasonable to assume that when our assessment is performed, the individual would report a score of 100% or near 100%. In such individuals, it is possible that an outcome tool with an objective, clinician-based component may not give maximal scores as correction for age and sex might not account for functionally normal variations in power and range of movement.

## CONCLUSIONS

This study adds to the weight of evidence for the value and utility of a subjective percentage shoulder assessment tool by the process of formal validation. It is relevant to note that the SPONSA as assessment tool is not diagnosis specific, and can therefore be used to evaluate comparative values of interventions for shoulder conditions. With an increasing emphasis on patient-reported outcomes in everyday clinical practice, a percentage assessment tool will be advantageous as it is simple and fast to administer. It may act as a useful adjunct to the existing outcome tools.

## REFERENCES

1. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. J Bone Joint Surg Br 1996;78:593-600.
2. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. Clin Orthop 1987;214:160-4.
3. Richards RR, An KN, Bigliani LU, Friedman RJ, Gartsman GM, Gristina AG, *et al*. A standardized method for the assessment of shoulder function. J Shoulder Elbow Surg 1994;3:347-52.
4. Lippitt SB, Harryman DT, Matsen FA. A practical tool for evaluation function: The simple shoulder test. In: Matsen FA, Fu FH, Hawkins RJ, Editors. The shoulder: A balance of mobility and stability. Rosemont: American Academy of Orthopaedic Surgeons; 1993. p.501-18.
5. Barrett WP, Franklin JL, Jackins SE, Wyss CR, Matsen FA. Total shoulder arthroplasty. J Bone Joint Surg Am 1987;69:865-72.
6. Conboy VB, Morris RW, Kiss J, Carr AJ. An evaluation of the Constant Murley Shoulder Assessment. J Bone Joint Surg Br 1996;78:229-32.
7. Kirkley A, Griffin S, McLintock H, Ng L. The Development and evaluation of a disease-specific quality of Life Measurement Tool for Shoulder Instability. The Western Ontario Shoulder Instability Index (WOSI). Am J Sports Med 1998;26:764-72.
8. Gerber C, Maquieira G, Espinosa N. Latissimus dorsi transfer for the treatment of irreparable rotator cuff tears. J Bone Joint Surg Am 2006;88:113-20.
9. Williams GN, Gangel TJ, Arciero RA, Uhorchak JM, Taylor DC. Comparison of the Single Assessment Numeric Evaluation method and two shoulder rating scales. Outcomes measures after shoulder surgery. Am J Sports Med 1999;27:214-21.
10. Gilbart MK, Gerber C. Comparison of the subjective shoulder value and the Constant score. J Shoulder Elbow Surg 2007;16:717-21.
11. Williams GN, Taylor DC, Gangel TJ, Uhorchak JM, Arciero RA. Comparison of the single assessment numeric evaluation method and the Lysholm score. Clin Orthop Relat Res 2000;373:184-92.
12. Fuchs B, Jost B, Gerber C. Posterior-inferior capsular shift for the treatment of recurrent, voluntary posterior subluxation of the shoulder. J Bone Joint Surg Am 2000;82:16-25.
13. Bland JM, Altman DG. Statistical method for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
14. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care 1989;27:178-89.
15. Machin D, Campbell M, Fayers M, Pinol A. Sample size for clinical studies 2nd ed). Hoboken, New Jersey: Blackwell science; 1997.
16. Harvie P, Pollard TC, Chennagiri RJ, Carr AJ. The use of outcome scores in surgery of the shoulder. J Bone Joint Surg Br 2005;87:151-4.
17. The Royal College of Surgeons of England, Clinical Effectiveness Unit. Available from: http://www.rcseng.ac.uk/media/media-background-briefings-and-statistics/measuring-surgical-outcomes [Last cited on 23 October 2011].
18. Rowe CR, Patel D, Southmayd WW. The Bankart procedure: A long term end-results study. J Bone Joint Surg Am 1978;60:1-16.
19. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: The DASH (disabilities of the arm, shoulder and hand). The Upper Extremity Collaborative Group (UECG). Am J Indian Med 1996;30:602-8.
20. Pranksy G, Feuerstein M, Himmelstein J, Katz JN, Vickers-Lahti M. Measuring functional outcomes in work-related upper extremity disorders. J Occup Environ Med 1997;39:1195-202.

**Source of Support:** Nil, **Conflict of Interest:** None declared.