The Royal College of Surgeons of England

REVIEW

Ann R Coll Surg Engl 2010; **92**: 9–13
doi 10.1308/003588410X12518836440522

Critical appraisal of subjective outcome measures used in the assessment of shoulder disability

ARAVIND S DESAI, ASTERIOS DRAMIS, ANTHONY J HEARNDEN

Department of Upper Limb Surgery, Wrightington Hospital, Wigan, UK

ABSTRACT

INTRODUCTION: Objective measures can be impractical in some settings, because they are time consuming and require face-to-face contact. More recently, there is an increasing trend towards the use of subjective outcome measures. Hence, in this article, five common subjective shoulder outcome measures are critically appraised in terms of their development, validity, reliability, responsiveness and clinical application.

MATERIALS AND METHODS: Following an extensive literature search, five common shoulder patient-based scores were identified: Disability of Arm, Shoulder and Hand (DASH), Oxford Shoulder Score (OSS), Shoulder Disability Questionnaire (SDQ-UK), Shoulder Pain and Disability Index (SPADI), and the Shoulder Rating Questionnaire (SRQ). These questionnaires were then critically appraised in terms of their development process, validity, reliability, responsiveness, and clinical application.

RESULTS: The SDQ-UK has shown good construct validity but there is no data available regarding internal consistency, reliability and responsiveness. The SPADI has good internal consistency, fair reliability with adequate criterion and construct validity. The DASH has shown to have good construct validity, excellent test–re-test reliability and responsiveness to change. The OSS has good sensitivity, validity and responsiveness. Though SRQ has good internal consistency, its reproducibility and responsiveness are poor.

CONCLUSIONS: Based on this critical appraisal, the DASH received the best ratings for its clinimetric properties followed by the OSS.

KEYWORDS

Shoulder – Disability – Subjective – Outcome measure

CORRESPONDENCE TO

Asterios Dramis, 38 Pakenham Road, Edgbaston, Birmingham B15 2NE, UK
E-mail: ad199@doctors.org.uk

Function of the shoulder has traditionally been assessed with measures such as range of motion, strength and pain. However, these objective measures can be impractical in some settings, because they are time consuming and require face-to-face contact. More recently, there is an increasing trend towards the measures of health-related quality of life both generic and disease-specific, often with use of questionnaires completed by the patients (subjective outcome measures).[1] Patient-based outcome measures provide a feasible and appropriate method for addressing the concerns of patients in the context of controlled clinical trials.[2] In general terms, patient-perceived health status and health-related quality of life are now generally accepted as the most important outcomes – barring mortality – from surgical intervention.[3] Furthermore, patients are expecting to be involved in decisions regarding their healthcare as they become increasingly interested in how their symptoms and their treatments are likely to affect them. Various subjective scoring systems are being developed and used to evaluate the results of treatment of shoulder joint disability. Some of the commonly used subjective outcomes measures, identified by our literature search, are Disability of Arm, Shoulder and Hand (DASH),[4] Oxford Shoulder Score (OSS),[5] Shoulder Disability Questionnaire (SDQ-UK),[6] Shoulder Pain and Disability Index (SPADI),[7] and the Shoulder Rating Questionnaire (SRQ)[8] (Table 1). In this article, we critically appraised these subjective outcome measures in terms of their development process, validity, reliability, responsiveness, and clinical application.

Materials and Methods

We conducted an extensive literature search to identify the entire subjective outcome measures used in the assessment of patients with shoulder disability symptoms.

We searched Medline (1950 to present), EMBASE (1980 to present) and CINAHL (1981 to present) using the OVID interface. We also searched other interfaces like Proquest, EBSCO and Academic Search Premier. We used search terms such as ‘shoulder’, ‘disability’, ‘upper extremity’, 'outcome measures', 'patient-generated questionnaire', 'self-assessment' either alone or in different combinations. References cited in retrieved articles were screened for additional relevant studies.

All studies that contained information on self-reporting and performance-based measures were included in the review. They were further screened in order to identify those which included items on physical functioning or disability and the main focus was either the development or clinimetric assessment of a shoulder disability outcome measure. We restricted the search to studies written as a full report and published in English. Questionnaires that were developed for patients whose main complaint did not involve a shoulder condition (*e.g.* wheelchair users) were excluded. Finally, objective outcome measures (*e.g.* solely performance-based measures with or without self-rated measures) were excluded as they did not form part of the study. We aimed to evaluate five common subjective shoulder outcome measures as we did not intend to perform a systematic review of all the self-rated shoulder outcome measures available in the medical literature.

The literature search identified 241 publications in which 22 self-rated shoulder disability questionnaires were reported. Six questionnaires did not fulfil our inclusion criteria and they were excluded. The studies which contained information on the remaining questionnaires were screened in order to identify five common subjective outcome measures used currently in clinical practice. Therefore, we selected 11 published articles which contained information on the following commonly used instruments – the DASH questionnaire, Oxford Shoulder Score, Shoulder Disability Questionnaire, Shoulder Pain and Disability Index and the Shoulder Rating Questionnaire. These questionnaires were evaluated in terms of the development process, validity, reliability, responsiveness and clinical application of each outcome measure.

Critical appraisal

Development

Each questionnaire has been developed to measure slightly different aspects of pathology, intervention or population, *i.e.* a measurement tool which has been developed to identify changes in one pathology or population could miss any changes in another (Table 1).

In the SDQ-UK, the authors generated the items after consultation with both field experts (epidemiologists, physiotherapists, and occupational therapists) and patients. The authors tested the questionnaire on two community population groups. They failed to mention the types of interviews used to generate the questions, the data analysis and the demographics. It is then likely, that there were demographic variations in consultation behaviour or differences in the perception of severity of shoulder pain.

In SPADI, the authors used a similar approach where they selected the items after consultations with a panel that included three rheumatologists and a physical therapist. There was no patient involvement. During the development of this questionnaire, the population selected were all males and primarily of old age. Because of this, skewed demographic items of this scale relate mostly to self-care and dressing (therefore, do not address occupational and recreational disability), meaning the score will not change when applied to other population groups, *i.e.* young patients with instability.

In DASH, the generation of the items was comprehensive. It was a collaborative project where members of the Upper Extremity Collaborative Group reviewed all possible items on symptoms and functional status of upper limb pathology. They pre-tested the questionnaire on 20 patients but failed to mention the baseline characteristics and pathologies of those patients. The authors administered the questionnaire to a large number of patients in different

Table 1 Description of the subjective shoulder outcome measures

Outcome measures	Target population[a]	Domains[b]	Number of items	Number of scales[c]	Range of scores
DASH	Upper extremity	Pain symptoms, physical, emotional, social	30	1	0–100
OSS	Shoulder operation	Pain, physical	12	1	12–60
SPADI	Shoulder pain	Pain, physical	13	2	0–100
SRQ	Shoulder disorders	Pain, function, social	21	6	17–100
SDQ	Shoulder symptoms	Physical, emotional, social	22	1	0–22

[a] Population for which the questionnaire has been developed.

[b] Domains: pain, other symptoms, physical functioning, emotional functioning, and social functioning.

[c] Scales: a sub score within a questionnaire.

parts of the world involving a sample of patients with various demographic factors and upper limp pathology.

In OSS, the study design used for the development of questionnaire was satisfactory. The development of this questionnaire was methodical and the authors took care with patient involvement in the drafting this outcome measure. There was equal distribution of men and women in the study. The main drawback was that two-thirds of the cohort suffered from impingement-type pathologies, while only one-third had arthritis or frozen shoulder. Hence, this outcome measure is more applicable and useful in evaluating outcome of people suffering from impingement pathology. This could have been avoided by recruiting and studying all groups of patients equally for its general applicability.

In the SRQ, it is not known how the items on the instrument were selected or established. Regarding the design of the study, there are many pitfalls. Of the study population, 73% were men and 64% had dominant-side pathology. This is an unequal distribution of the study. In terms of occupation, 75% of patients involved in the study were paid workers leading to unequal recruitment of the patients. The majority of patients were suffering from impingement syndrome and instability. This again makes the SRQ more applicable to this pathological conditions rather than general shoulder disability.

Validity

It is important to formulate hypotheses before validity testing and the authors in SDQ-UK have done this well. Their first hypothesis was that subjects attending general practitioners with shoulder pain would have higher levels of disability than those identified by population screening. Their second hypothesis was that disability scores would reflect more objective measures of shoulder restriction. Both these hypotheses were confirmed and the authors specify clearly the magnitude and the direction of the expected correlation. Even though there are no agreed standards for how high correlations should be between a new questionnaire and other variables in order to establish construct validity, a value of 0.6 may be strong evidence in support of construct validity.[5] In this questionnaire the correlation coefficient was 0.84 giving strong evidence of this questionnaire measure.

In SPADI, the results showed that the division between the two dimensions might not be indicated. Though there is some evidence to support the construct validity, the factor analysis suggested that the scale might not reflect two separate dimensions.[9] Roach *et al.*[7] originally reported a two-factor solution to factor analysis when varimax rotation was used. However, that study sample was much lower than typically required to conduct factor analyses, which would suggest that there might be some instability in their results.[10] The authors did not compare their questionnaire with a health-related quality of life measure with an established validity and reliability. This could have provided stronger evidence for the validity of their study. It is interesting to note that the validity is strong for the out-patient setting but doubtful for the primary care and hospital patients. Furthermore, 10% of the patients had shoulder pain of radicular origin that may be not associated with decreased shoulder range of motion (ROM). Therefore, the inclusion of these patients in the study may have influenced the association between active shoulder ROM and SPADI scores. On the contrary, there is not enough research in the medical literature to suggest a positive correlation between shoulder active ROM and function. Beaton and Richards[1] found that active shoulder ROM measurements correlated poorly with scores on the SPADI questionnaire, *i.e.* patients with better movement had less pain and disability.

It is important to formulate hypotheses before validity testing. These hypotheses should specify both magnitude and direction of the expected correlation. In DASH, this is demonstrated very well by showing moderate-to-high correlations of scores with other outcome measures of shoulder disability. In both DASH and SPADI, there is no record regarding internal consistency, so the question on the extent to which items in a scale are intercorrelated is doubtful.

In OSS the authors ensured content validity by deriving the items in the questionnaire by exploratory interviews with patients. The draft versions were tested and re-tested; the final version was established only when all participants agreed after complete understanding. Construct validity was also adequately tested by examining the level of agreement of the questionnaire with the clinical data and with the scales from existing health status questionnaires like SF-36 and HAQ (Health Assessment Questionnaire) with a good correlation noted amongst these scores. Correlation was highest in the assessment of pain and physical function. Thus, the validity of the OSS was well tested and had good correlation with other well-established questionnaires. Although predictions as to how the instruments should correlate were not made, the results, which show modest correlation with the other shoulder instruments and the appropriate domains of the global tools, seem appropriate.[11]

In SRQ, the authors describe validation process by correlating the scores with domains of arthritis impact measurement scales. The authors fail to describe any predictions or observed correlations mentioned in the study. Though a construct tested showed significant difference in all the four domains, construct validation between this measure and the other outcome measures have not been explained.

Reliability

The reliability of an outcome measure is its ability to generate the same score on the same group of patients at a later date. The test–re-test coefficients need to exceed 0.90 or 0.95 before an interpretation on an individual level can be considered. The intraclass correlation coefficient (ICC) is the test of choice, as the Pearson correlation coefficient can neglect systematic errors if present.[12]

In the SDQ-UK questionnaire, there is no data available regarding reliability. It is essential to establish that any changes observed in a trial are due to the intervention and not to problems in the measuring instrument. SDQ-UK simply fails to prove this. In SPADI, the results indicate both that the items of these scales are measuring consistently with each other and each of the constructs being measured (pain and function or disability) are relatively homogeneous constructs.[13] Bot *et al.*[12] have argued that, because ICC is low, the SPADI may not be applicable for individual patients.

The DASH scoring system exceeded the recommended standards for test-retest reliability.

In OSS, reproducibility (test–re-test reliability) was assessed in only 60 out of 111 patients. The main criticism of this is that the authors did the reliability test within 24 h. Due to the recall period being very short, there is always a possibility that the patients were able to remember their earlier scores, which, in turn, would have increased the *r*-value. Though the ICC was not calculated, the Pearson correlation coefficient closely approximates it.

In the SRQ, the reliability was tested in only 40 patients who repeated the questionnaire after a mean of 3 days. The reproducibility of the total scale and its subscales was assessed using the Spearman–Brown test–re-test reliability test. Spearman rank correlation coefficient was used as an additional measure of reproducibility. According to Kirkley *et al.*,[11] a strong criticism of this approach is that values may have been falsely elevated for two reasons. First, 3 days is unlikely to be long enough for patients to forget their original score. Second, testing reliability in such diverse population increases the numerator, giving higher reliability.

Responsiveness

Responsiveness refers to an instrument's ability to detect important changes over time within individuals that might reflect therapeutic effects.

In SDQ-UK, the authors do not tell us about the ability of this questionnaire to detect clinically relevant changes. There are also floor and ceiling effects. This questionnaire reduces the likelihood of further improvement or deterioration being recorded beyond a certain point and, hence, making it impossible to report most favourable or worst health states. It has a ceiling effect for community people with shoulder pain (subjects with shoulder pain in past month lasting for > 24 h), but not for primary care patients (subjects who attended primary care practice with a new episode of shoulder pain).

In SPADI, the responsiveness was very good for the outpatient clinic but doubtful for primary care and hospital patients.

The DASH showed responsiveness to change observed before and after treatment and change in those patients who said they were better. The authors also showed that the DASH has potential in the role of monitoring physical function and symptoms in shoulder, wrist and hand disorders. Neither the DASH nor SPADI have shown any floor and ceiling effects.

In OSS, the sensitivity to change was examined by comparing scores before and 6 months after operation by one patient satisfaction item and three transition items. The authors demonstrated the sensitivity to change of this outcome measure both at the end of 6 months and at medium term (4 years) when compared with SF-36 and HAQ. They also suggest that reliability and sensitivity of the OSS was significantly reduced over the long term.

In SRQ, the responsiveness, as measured by the standardised response mean, compares favourably with the values reported for several health-status questionnaires.[8]

Responsiveness of this outcome measure has not been tested or compared with any other established shoulder outcome measures. It has no ceiling or floor effects. The sensitivity of this outcome measure is well documented.

Clinical application

Clinical application of an outcome measure depends on the MCID (minimal clinical important difference); for example, a difference of five points may be relevant in some scores but not in others. When authors do not provide an indication of how to interpret changes in their outcome score, the findings are of limited use to clinicians.[14] Only in SPADI and SRQ were outcome measures interpretability of the outcome scores and MCID reported.

The DASH and SPADI are recommended for use in the out-patient clinic. These questionnaires received positive ratings for responsiveness and have no floor or ceiling effects. For test–re-test reliability, an ICC 0.70 is regarded as adequate for group comparisons, yet for individual comparisons an ICC of > 0.90 should be required. This means that the SPADI (ICC 0.57–0.84) may not be applicable for individual patients. Finally, OSS should not be used for evaluation of patients having shoulder stabilisation as it has not been evaluated in this group.

Discussion

Table 2 presents a summary of the five outcome instruments discussed in detail in this article and shows that the DASH and OSS questionnaires have received the best ratings for their clinimetric properties.

The OSS is a simple, reliable scoring system. It has good sensitivity, validity and responsiveness. It is specific to shoulder pathologies except for instability. The main weakness is its

Table 2 **Summary of the critical appraisal of the subjective shoulder outcome measures**

Outcome measures	Development	Validity	Reliability	Responsiveness
DASH	Good	Good	Good	Good
OSS	Good	Good	Fair	Good
SPADI	Fair	Poor	Poor	Good[a]
SRQ	Poor	Fair	Poor	Poor
SDQ	Poor	Good	Poor	Poor

[a]Out-patient clinic only.

many versions and confusing score interpretation. Though SRQ has good internal consistency, its reproducibility and responsiveness are poor. It has many limitations as discussed earlier making it not generalisable. The SDQ-UK is a simple and easy-to-score questionnaire. It has shown good construct validity but there is no data available regarding internal consistency, reliability and responsiveness. Its clinimetric properties are not good enough for evaluation of patients with shoulder disability. The SPADI appears to have functioned well on community patients, which were primarily of older men. The degree to which these results can be generalised to women and younger patients with shoulder symptoms is not clear. Patients must be instructed in the proper use of the SPADI. Though it demonstrates good internal consistency, it has poor reliability and validity. It also appears to be able to detect change in patient status over time. The DASH has shown to have good construct validity, test–re-test reliability and responsiveness to change and, along with OSS, would constitute our recommended subjective outcome measures for use in the clinical setting.

Selection of a subjective shoulder questionnaire from the numerous scales can be overwhelming. We did not intend to perform a full systematic review of all the subjective shoulder questionnaires available but rather focus on few common ones used currently in clinical practice. This is one of the limitations of our study where the number of questionnaires had to be limited due to the busy clinical setting and some other frequently used measures were not included in our critical review. Furthermore, there are no standardised criteria to appraise healthcare outcome measures critically and our evaluation may be disputed. However, our aim was to provide information on clinimetric properties of some common shoulder subjective questionnaires in order to facilitate the choice between the questionnaires.

Acknowledgement

The authors wish to thank Dr L Goodacre for her valuable contribution in this study.

References

1. Beaton DE, Richards RR. Measuring function of the shoulder: a cross-sectional comparison of five questionnaires. *J Bone Joint Surg Am* 1996; **78**: 882–90.
2. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Tech Ass* 1998; **2**: 1–86.
3. McDowell I, Newell C. *Measuring health: a guide to rating scales and questionnaires*, 2nd edn. New York: Oxford University Press, 1996.
4. 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 Ind Med* 1996; **29**: 602–8.
5. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. *J Bone Joint Surg Br* 1996; **78**: 593–600.
6. Croft P, Pope D, Zonca M, O'Neill T, Silman A. Measurement of shoulder related disability: results of a validation study. *Ann Rheum Dis* 1994; **53**: 525–8.
7. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanaku L. Development of a shoulder pain and disability index. *Arthritis Care Res* 1991; **4**: 143–9.
8. L'Insalata JC, Warren RF, Cohen SB, Altchek DW, Peterson MG. A self–administered questionnaire for assessment of symptoms and function of the shoulder. *J Bone Joint Surg Am* 1997; **79**: 738–48.
9. Heald SL, Riddle DL, Lamb RL. The shoulder pain and disability index: the construct validity and responsiveness of a region specific disability measure. *Phys Ther* 1997; **77**: 1079–89.
10. MacDermid JC, Solomon P, Prkachin K. The Shoulder Pain and Disability Index demonstrates factor, construct and longitudinal validity. *BMC Musculoskel Disord* 2006; **7**: 12.
11. Kirkley A, Griffin S, Dainty K. Scoring system for the functional assessment of the shoulder. *Arthroscopy* 2003; **19**: 1109–20.
12. Bot SDM, Terwee CB, van der Windt DAWM, Bouter LM, Dekker J, de Vet HCW. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. *Ann Rheum Dis* 2004; **63**: 335–41.
13. Roddey TS, Olson SL, Cook KF, Gartsman GM, Hanten W. Comparison of the University of California–Los Angeles Shoulder Scale and the Simple Shoulder Test With the Shoulder Pain and Disability Index: single-administration reliability and validity. *Phys Ther* 2000; **80**: 759–68.
14. Guyatt GH. Making sense of quality–of-life data. *Med Care* 2000; **38 (Suppl)**: 175–79.