REVIEW

Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature

S D M Bot, C B Terwee, D A W M van der Windt, L M Bouter, J Dekker, H C W de Vet

Ann Rheum Dis 2004;**63**:335–341. doi: 10.1136/ard.2003.007724

Appendix W1 and tables W1–W3 are available at http://www.annrheumdis.com/supplemental

See end of article for authors' affiliations

Correspondence to: Ms S D M Bot, van der Boechorststraat 7, Amsterdam 1081 BT, The Netherlands; s.bot@vumc.nl

Accepted 9 September 2003

Objective: To identify all available shoulder disability questionnaires designed to measure physical functioning and to evaluate evidence for the clinimetric quality of these instruments.

Methods: Systematic literature searches were performed to identify self administered shoulder disability questionnaires. A checklist was developed to evaluate and compare the clinimetric quality of the instruments.

Results: Two reviewers identified and evaluated 16 questionnaires by our checklist. Most studies were found for the Disability of the Arm, Shoulder, and Hand scale (DASH), the Shoulder Pain and Disability Index (SPADI), and the American Shoulder and Elbow Surgeons Standardised Shoulder Assessment Form (ASES). None of the questionnaires demonstrated satisfactory results for all properties. Most questionnaires claim to measure several domains (for example, pain, physical, emotional, and social functioning), yet dimensionality was studied in only three instruments. The internal consistency was calculated for seven questionnaires and only one received an adequate rating. Twelve questionnaires received positive ratings for construct validity, although depending on the population studied, four of these questionnaires received poor ratings too. Seven questionnaires were shown to have adequate test-retest reliability (ICC >0.70), but five questionnaires were tested inadequately. In most clinimetric studies only small sample sizes (n<43) were used. Nearly all publications lacked information on the interpretation of scores.

Conclusion: The DASH, SPADI, and ASES have been studied most extensively, and yet even published validation studies of these instruments have limitations in study design, sample sizes, or evidence for dimensionality. Overall, the DASH received the best ratings for its clinimetric properties.

Function of the shoulder has conventionally been assessed with objective measures such as range of motion and strength. However, objective measures can be impractical in some settings, because they are time consuming and require face to face contact. Besides, although shoulder disorders are often associated with restricted range of motion and muscle weakness, these measures have no direct clinical meaning to patients, who just want to be free of pain and perform their daily activities. Nowadays, the efficacy of treatment is more often evaluated using outcomes that are directly relevant to patients. Both in clinical practice and research, using subjective measures that assess the ability to function in daily life ensures that the treatment and evaluations focus on the patient rather than on the disease.[1]

In the past decade a large number of shoulder disability questionnaires have been developed, which are designed to assess physical functioning (that is, the performance of daily activities).[2–17] The choice of which questionnaire to use may be based on the study group, the purpose of the questionnaire, its clinimetric quality as shown by validity, reproducibility, responsiveness, and on practical considerations (for example, ease of scoring, and how long it takes to complete). Different questionnaires may be required for different patient groups, but this should be balanced against the need to standardise results from different studies by the use of a single instrument.[18] Studies comparing the content and clinimetric quality of these shoulder disability questionnaires are lacking. Consequently, little evidence is available to guide the clinician and researcher during questionnaire selection.

Garratt *et al* stated that structured reviews are prerequisites for standardisation.[19] We developed a checklist to evaluate the clinimetric quality of the instruments as shown by their validity, reproducibility, responsiveness, interpretability, and practical burden. The purpose of this paper is to systematically review the content and clinimetric quality of all published shoulder disability questionnaires in order to provide guidelines for clinicians and researchers enabling them to choose the most appropriate measure for different purposes.

METHODS

Study selection

Initially, studies were identified by searches of the computerised bibliographic database Medline (1966–July 2002). Subsequently, other databases—that is, CINAHL (1988–July 2002), SPORTDiscus (1949–July 2002), and PsychINFO were searched for additional studies. The following keywords were used to identify eligible studies: shoulder, upper-extremity, disability, functional status, questionnaire, self-report, self-assessment, outcome measure, outcome assessment (MESH term or text word). The names of identified instruments were used as terms for a further search of the electronic databases. References of retrieved articles were screened for additional relevant studies.

Inclusion criteria

Instruments were included in the review if they were self assessed, condition-specific (shoulder or combined shoulder-upper limb problems), and included items on disability or

Abbreviations: CTT, classical test theory; ICC, intraclass correlation coefficient; IRT, item response theory; MCID, minimal clinically important difference

physical functioning (that is, the performance of activities of daily living). Studies were eligible when the main focus of the study was the development and/or the clinimetric evaluation of a shoulder disability questionnaire. Furthermore, only studies that were written as full report (that is, no abstract or letter to the editor) and had been published in English were included. No restrictions were put on the year of publication. Instruments that were developed for groups whose primary complaint did not concern shoulder disorders (for example, wheelchair users, patients with cancer) were excluded.

Data abstraction and quality assessment

A checklist was composed to evaluate and compare the clinimetric properties of the questionnaire. The checklist was partly based on the review criteria developed by the Scientific Advisory Committee of the Medical Outcome Trust[20] and the checklist developed by Bombardier and Tugwell.[21] After testing the checklist on papers about other condition-specific questionnaires the final version of the checklist was completed. (This checklist can be found in appendix W1, available at http://www.annrheumdis.com/supplemental) Two reviewers independently scored the clinimetric quality of each study, according to the checklist. If an instrument had more than one scale, only the subscales containing items on physical functioning were reviewed. Disagreements between the reviewers were discussed and resolved during a consensus meeting.

Description of the questionnaires

Descriptive data extracted from the publications included the target population, domains to which the items could be classified (pain, symptoms, physical functioning, emotional functioning, and social functioning), number of scales, number of items, response options, range of minimal and maximal score, time needed to complete the questionnaire, and study groups used in the clinimetric studies about the questionnaires.

Practical burden

For administrative burden the scoring method was rated: easy, when the items were simply summed; moderate, when a visual analogue scale or simple formula was used; and difficult, when either a visual analogue scale in combination with a formula or a complex formula was used. For respondent burden a positive rating was given when the questionnaires could be completed within 10 minutes.

Validity

Validity is the degree to which an instrument measures what it is supposed to measure.[20] The instruments were evaluated for content and construct validity. Content validity examines the extent to which the domain of interest is comprehensively sampled by the items in the questionnaire.[22] Items on the questionnaire must reflect areas that are important to patients with shoulder disorders. Therefore, studies achieved a positive rating for content validity when patients were involved during item selection. A positive rating for readability or comprehension was given when patients tested the questionnaire in a pilot study.

Internal consistency is a measure of the homogeneity of a (sub)scale. It indicates the extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct. Factor analysis should be applied to determine the dimensionality of the items—that is, to determine whether or not they formed only one overall dimension or more than one. A positive rating for internal consistency was achieved when the dimensional structure of the questionnaire was explored by factor analysis and Cronbach's α for each dimension separately was between 0.70 and 0.90.[23]

Construct validity refers to the extent to which scores on a particular instrument relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the constructs that are measured.[24] The associations of the questionnaire with other variables that measured disability or physical functioning were abstracted from the studies. Construct validity was considered to be adequately tested if hypotheses were specified and the results corresponded with these hypotheses.

Floor and ceiling effects were considered present if more than 15% of respondents achieved the highest or lowest possible score, respectively.[25] Therefore, authors had to provide descriptive statistics of the distribution of scores, which included information on the presence of floor or ceiling effects.

Reproducibility

Reproducibility is the extent to which an instrument is free of measurement error. It was assessed by rating test-retest reliability and agreement.[26] We considered calculation of the intraclass correlation coefficient (ICC) for each domain as an adequate method for test-retest reliability.[26] An ICC >0.70 for group comparisons was rated as positive.[20 23 27] Confidence intervals should be presented as an index of the expected random variation. Application of Pearson correlation coefficients to estimate test-retest reliability was rated as doubtful, as it neglects systematic errors if present.[28]

A measure of agreement is important to quantify measurement error and detect systematic differences between two measurements. Calculation of the 95% limits of agreement,[29] the κ coefficient,[30] or the standard error of measurement (SEM) was regarded as an adequate measure of agreement. It was not possible to define adequate cut off points for the result of an agreement study. Therefore, a positive rating was received when an adequate method for agreement was used.

Responsiveness

Responsiveness refers to an instrument's ability to detect important change over time in the concept being measured.[31–33] There is no single agreed method to assess responsiveness. Calculating change scores for a group of patients whose health is expected to have changed and to examine the correlation with corresponding changes in a reference measure or transition was considered to be a suitable method to assess responsiveness.[32] This method requires predictions about how the results of the questionnaire should correlate with other related measures. Responsiveness was considered adequately tested if hypotheses were specified and when the results corresponded with these hypotheses.

Validity, reproducibility, and responsiveness depend on the setting and the population in which they are assessed. Therefore, the description of the design of each individual clinimetric study was rated. A clear description included characteristics of the study group (including diagnosis and clinical features), measurements, testing conditions, and data analysis. Furthermore, methodological weaknesses in the design or execution of a study were recorded. When an adequate description was lacking or methodological weaknesses were found, validity was rated as doubtful.

Interpretability

Interpretability was defined as the degree to which one can assign qualitative meaning to quantitative scores.[20] The investigators should provide information about what (difference in) score would be clinically meaningful. We rated if the authors had presented a minimal clinically important difference (MCID) or if other information was present that could aid in interpreting the questionnaires' scores—for instance, (*a*) presentation of means and standard deviations

(SD) of patients scores before and after treatment; (*b*) comparative data on the distribution of scores in relevant subgroups; (*c*) information on the relationship of scores to well known functional measures or to clinical diagnosis; and (*d*) relating changes in disability score to patients' global ratings of the magnitude of change they have experienced. Investigators had to provide at least two of the previously described types of information for a positive rating of interpretability.

RESULTS

The Medline search identified 553 publications, in which 22 self administered shoulder disability questionnaires were reported. The additional searches in CINAHL, SPORTDiscus, and PsychINFO identified one additional questionnaire (SPADI) and two additional studies about the clinimetric characteristics of included questionnaires.[16 34] Reference tracking resulted in four additional clinimetric studies.[11 15 35 36] Of the 23 identified questionnaires, 17 met the inclusion criteria. One questionnaire, the University of Pennsylvania Shoulder Scale (Upenn),[37] was not evaluated because the clinimetric properties were obtained by Rasch analysis. We based our checklist on classical test theory (CTT) and did not accommodate an evaluation of rate item response theory (IRT) methods. Therefore, we were not able to evaluate this questionnaire. Four questionnaires were excluded because they were developed for special groups (that is, wheelchair users, patients with bone and soft tissue sarcoma, and athletes).[38–41] Two questionnaires were excluded as they had no items on physical functioning.[42 43] Finally, a total of 28 studies referring to 16 shoulder disability questionnaires were included in the review.

Description of questionnaires

Table 1 presents a description of the 16 included questionnaires (full names are given in appendix 1). The DASH, UEFS, and UEFL were designed as upper extremity questionnaires, but can be used for the evaluation of any joint or condition of the upper extremity, and all have been applied in patients with shoulder problems. Two questionnaires were developed for shoulder instability (SIQ, WOSI), one for rotator cuff tears (RC-QOL), and one for osteoarthritis (WOOS). The other questionnaires were developed for shoulder disorders in general. The ASES consist of a self administered and a performance based part. Only the self administered part was reviewed. The RC-QOL had the most items (n = 34), followed by the SSI (n = 31), and the DASH (n = 30), while the SSRS had the smallest number of items (n = 5).

Validity

Most questionnaires were developed out of a pool of items generated by patients, experts, and/or the investigator(s). The dimensional structure of only three questionnaires (SST, UEFS, and SPADI) was studied by factor analysis. In two studies the factor analysis of the SPADI showed loading on one factor only, although the questionnaire claims to measure two constructs: pain and disability.[16 44] In contrast, factor analysis supported a two factor solution for the SST, while the SST claims to measure a single construct.[44] Information on internal consistency was found for seven questionnaires. Cronbach's α ranged from 0.71 to 0.96. The SIQ and the disability subscale of the SPADI had a Cronbach's α above 0.90.

Construct validity was studied by correlating the score of the questionnaire with other disability questionnaires, with the physical function dimension of general health instruments, or with a global rating system for shoulder disorders. Six of 19 studies that investigated construct validity did not present hypotheses relating to the magnitude and direction of expected relationships with other instruments. The SSRS had moderate correlations with other shoulder disability questionnaires (0.47−0.50). The correlations between the SST, SSI, ASES, SPADI, and DASH were high (>0.74).

Three questionnaires showed a floor or ceiling effect: The SDQ-UK showed a ceiling effect in a community sample of people with shoulder pain, the UEFL a floor effect for older women in the community, and the SDQ-NL showed a ceiling effect for primary care patients.

Reproducibility

Information on the test-retest reliability was found for 10 questionnaires. A Pearson correlation coefficient was used to calculate test-retest reliability of the SIQ and SRQ, while an ICC was reported for the other questionnaires. Except for the SPADI all coefficients were >0.70. Test-retest reliability of the SPADI was investigated in four studies and the ICC for the disability subscale ranged from 0.57 to 0.84.

Six studies presented information on agreement of, in total, 10 questionnaires. Methods used were the coefficient of reliability,[29] the SEM, and the percentage of agreement on repeated measures.

Responsiveness

The responsiveness of 13 questionnaires was evaluated in 14 studies. Four responsiveness studies presented hypotheses. Most studies compared scale scores before and after the treatment and presented mean change scores only. Furthermore the standardised response mean was used frequently. No data on responsiveness were found for the SDQ-UK, RC-QOL, and UEFL. The number of patients used to measure responsiveness was small (n<43) in eight of 14 responsiveness studies.

Interpretability

Five studies paid attention to interpretability of scores, and for three questionnaires (SRQ, SPADI, and SDQ-NL) an MCID was presented. Information on scores of different shoulder disability groups was available for the SST.[35] Means and SD (or equivalent) of baseline and follow up scores or scores of relevant subgroups were available for nine questionnaires. No data on the distribution of scores from the SDQ-UK, WOSI, WOOS, SSI, and UEFL were found.

Detailed information on the clinimetric properties of the questionnaires (that is, validity, reproducibility, responsiveness, and interpretability) can be found in tables W1–W3 available at http://www.annrheumdis.com/supplemental.

Overall quality

Only a few studies gave an adequate description of the study design and population characteristics. Eight studies did not adequately describe its study group and in five studies information about data analyses was missing. Nine publications provided insufficient information on the methodological aspects to enable a good evaluation of the study design. Furthermore, information about non-response, subjects lost to follow up, and missing data were often lacking.

Table 2 shows the quality assessment of the 16 shoulder disability questionnaires, summarising each item as good, doubtful, or poor quality. A question mark indicates insufficient information about an aspect of quality. As results are dependent on the population studied, the kind of population is presented (that is, community, primary care, outpatient, or hospital patients). Overall, the DASH received the best ratings for its clinimetric properties (that is, 9 positive scores out of 12).

Table 1 description of the shoulder disability questionnaires

Questionnaire	Target population*	Domains†	Number of scales‡	Number of items	Number of response options	Range of scores	Time to administer (min)	Ease of scoring	Study population(s)§
SDQ-UK	Shoulder symptoms	Physical, emotional, social	1	22	2	0–22	?	Easy	Community/general practice[2]
SIQ	Shoulder instability	Pain symptoms, physical, emotional	1	12	5	12–60	?	Easy	Outpatient clinic[3]
OSQ	Shoulder operation	Pain, physical	1	12	5	12–60	?	Easy	Outpatient clinic[4]
SDQ-NL	Soft tissue, shoulder disorders	Pain, physical, emotional	1	16	3	0–100	5–10	Easy	Rehabilitation centre,[5] general practice[62]
RC-QOL	Rotator cuff disease	Pain symptoms, physical, emotional, social	1	34	VAS	0–100	?	Moderate	Sports medicine centre[6]
DASH	Upper extremity	Pain symptoms, physical, emotional, social	1	30	5	0–100	<5	Moderate	Medical centre,[63 64] sports medicine clinic,[8] upper extremity clinic[65]
WOSI	Shoulder instability	Pain symptoms, physical, emotional, social	5	21	VAS	0–2100	?	Moderate	Sports medicine clinic[8]
SSRS	Shoulder problems	Pain symptoms, physical	1	5	3; 4; 5	0–100	<5	Easy	Orthopaedic clinic,[9] upper extremity clinic[66 67]
SRQ	Shoulder disorders	Pain, function, social	6	21	5; VAS	17–100	5–10	Difficult	hospital for special surgery[10]
SST	Shoulder	Physical	1	12	2	?	<3	?	Shoulder and elbow clinic,[11] medical centre,[64] private practice surgeon,[49] upper extremity clinic[66 67]
WOOS	Osteoarthritis of the shoulder	Pain symptoms, physical, emotional, social	4	19	VAS	0–1900	10	Moderate	Sports medicine clinic[12]
SSI	Shoulder pain	Pain, physical	?	31	3; 5; VAS	1–100	7	Difficult	Upper extremity clinic[66 67]
UEFS	Upper extremity (occupational)	Physical	1	8	11	8–80	<5	Easy	Upper extremity clinic[14]
ASES	Shoulder problems	Pain, physical, symptoms	2	16, m-ASES: 15	2; 4; VAS	0–100	3–5	Difficult	Medical centre[41 64]
SPADI	Shoulder pain	Pain, physical	2	13	VAS; 11	0–100	3–10	Difficult	Ambulatory care clinic,[16] medical centre,[63 68] upper extremity clinic,[66 67] physical therapy clinic,[39] private practice surgeon [49 59]
UEFL	Upper extremity	Function	1	3	5	?	?	Easy	Older community women[17]

*Population for which the questionnaire has been developed; †domains: pain, other symptoms, physical functioning, emotional functioning, and social functioning; ‡scales: a subscore within a questionnaire; §population(s) used in the clinimetric studies.
VAS, visual analogue scale; m-ASES, modified ASES[67]; ?, no data published.

Table 2 Summary of the quality assessment of the shoulder disability questionnaires

Questionnaire	Time to administer	Ease of scoring	Readability and comprehension	Content validity	Internal consistency	Construct validity	Floor/ceiling effect	Reliability	Agreement	Responsiveness	Interpretability	MCID
SDQ-UK	?	+	?	+	?	+	+ (b); − (a)	?	?	?	?	?
SIQ	?	+	+	+	o	+	+	o	+	+	+	?
OSQ	?	+	+	+	o	+	+	?	+	+	+	?
SDQ-NL	+	+	+	o	?	+	− (b)	?	?	+ (b); o (b)	+	+
RC-QOL	?	o	+	+	?	+	+	?	o	?	o	?
DASH	+	o	+	+	?	+ (c, d); o (c)	+ (c)	+ (c)	+ (c)	+ (c); o (d)	+	?
WOSI	?	o	+	+	?	+	?	+	?	o	?	?
SSRS	+	+	?	−	?	o (c)	+ (d)	o (d)	o (d)	o (d)	+	?
SRQ	+	−	+	+	o	o	?	o	o	o	o	o
SST	+	?	?	+	o	+ (c); o (d)	+ (c)	o (d)	o (c, d)	o (c, d)	+	?
WOOS	+	o	+	+	?	o	?	o	?	o	?	?
SSI	+	−	?	?	?	+ (c)	+ (c)	o (d)	o (d)	o (d)	?	?
UEFS	+	+	?	−	+	o	+	?	?	o	+	?
ASES	+	−	?	−	o	+ (c); o (d)	+ (d); o (c)	o (c, d)	o (d)	o (c, d)	+	?
SPADI	+	−	?	−	o	+ (c); o (b, c)	+ (b, c)	o (c, d)	+ (c); o (d)	+ (c); o (b, d)	+	+
UEFL	?	+	?	−	?	+	− (a)	?	?	?	?	?

MCID, minimal clinically important difference; method or result was rated as: + good; o doubtful; − poor; ? no data available.
Kinds of study population(s) used in the clinimetric studies: (a) community; (b) primary care; (c) outpatient clinic; and (d) hospital patient.

DISCUSSION

We identified 16 condition-specific questionnaires for the evaluation of physical functioning in patients with shoulder disorders for which the clinimetric characteristics had been evaluated. None of the questionnaires demonstrated satisfactory results for all categories. Overall, the DASH received the best ratings for its clinimetric properties.

When constructing a questionnaire, one should specify beforehand which constructs it is supposed to measure (that is, if the questionnaire will be a unidimensional or multidimensional instrument). Subsequently, the theoretically dimensional structure should be tested using factor analysis. We found that this is not properly done, or not done at all. One may assume that the number of scales corresponds with the number of dimensions, but only five questionnaires had an equal number of scales and dimensions. Seven questionnaires claimed to cover more than one dimension, but had one scale only, and the SST and SPADI appeared to have different structures than stated. Seven questionnaires claimed to cover more than one dimension, but had one scale only. When the dimensionality of a questionnaire is not analysed, Cronbach's α may not be interpretable.

The DASH and WOSI were the only questionnaires with a positive rating for test-retest reliability. Test-retest reliability of the SRQ, SSRS, SST, WOOS, and SSI was done using small sample sizes (n = 22–41). When statistical estimates are derived from very small populations, confidence intervals will be wide. This indicates the high degree of uncertainty in the precision of the reliability coefficient.

Our checklist was developed to evaluate the measurement properties of questionnaires based on CTT. A relatively new method to develop and evaluate health status questionnaires is IRT.[45] IRT has a number of potential advantages over CTT[46] and can be helpful in developing health outcome measures with better clinimetric properties. IRT makes it possible to calibrate a large number of physical functioning items on the same scale, which allows different tests to be meaningfully compared with one another, even if they are administered to completely different groups.[47] Cook *et al* used an IRT model to investigate the trait-specific reliability of the DASH, ASES, SST, and Upenn.[37] They showed that the questionnaires did not measure all levels of shoulder functioning with equal precision (that is, the questionnaires were unable to measure accurately patients with very low or very high levels of shoulder functioning). The evaluation of shoulder disability questionnaires may be improved by using IRT.[37]

Validity studies were available for all questionnaires. It is important to formulate hypotheses before validity testing. These hypotheses should specify both magnitude and direction of the expected correlation. The same accounts for studies on responsiveness. Most authors looked at the treatment effect, but the magnitude of the treatment effect tells us little about the ability of an instrument to detect clinically relevant change.[32]

The presence of floor and ceiling effects may influence the responsiveness of an instrument. An intervention effect will be missed for people who occupy the lower levels of the scale before the intervention. Floor and ceiling effects are dependent upon the population being studied. The SDQ-UK had a ceiling effect for community people with shoulder pain, but not for primary care patients; the SDQ-NL showed a ceiling effect for patients with shoulder pain receiving physiotherapy treatment, but not for patients with shoulder disorders visiting their general practitioner.

More information is needed on the interpretation of scores. Only five studies paid attention to interpretability of the outcome scores and an MCID was stated for only three questionnaires (SRQ, SPADI, and SDQ-NL). When investigators do not provide an indication of how to interpret

Table 3 Full names of the questionnaires included

SDQ-UK	Shoulder Disability Questionnaire	Croft P *et al* (1994)[2]
SIQ	Shoulder Instability Questionnaire	Dawson J *et al* (1999)[3]
OSQ	(Oxford) Shoulder Questionnaire	Dawson J *et al* (1996)[4]
SDQ-NL	Shoulder Disability Questionnaire	van der Heijden GJ *et al* (1996)[5]
RC-QOL	Rotator Cuff Quality of Life Measure	Hollinshead RM *et al* (2000)[6]
DASH	Disabilities of the Arm, Shoulder, and Hand Scale	Hudak PL *et al* (1996)[7]
WOSI	Western Ontario Shoulder Instability Index	Kirkley A *et al* (1998)[8]
SSRS	Subjective Shoulder Rating Scale	Kohn D *et al* (1997)[9]
SRQ	Shoulder Rating Questionnaire	L'Insalata JC *et al* (1997)[10]
SST	Simply Shoulder Test	Lippitt SB *et al* (1993)[11]
WOOS	Western Ontario Osteoarthritis of the Shoulder index	Lo IKY *et al* (2001)[12]
SSI	Shoulder Severity Index	Patte D (1987)[13]
UEFS	Upper Extremity Function Scale	Pransky G *et al* (1997)[14]
ASES	American Shoulder and Elbow Surgeons Standardised Shoulder Assessment Form	Richards RR *et al* (1994)[15]
SPADI	Shoulder Pain and Disability Index	Roach KE *et al* (1991)[16]
UEFL	Upper Extremity Functional Limitation Scale	Simonsick EM *et al* (2001)[17]

changes in health related quality of life score, the findings are of limited use to clinicians.[48] Among others, Lydick and Epstein have described different approaches for interpretation of health related quality of life changes.[49] It should be recognised that interpretation of the results is questionable when the clinimetric quality of an instrument is unknown or has not been adequately tested.

It is important to realise that the clinimetric properties of a questionnaire are not fixed, and may vary among different settings and populations.[50] The use of various methods and various populations helps in ''building'' these properties. The DASH, SST, ASES, and SPADI have been studied most often. Besides rating the clinimetric properties of a questionnaire, the choice of a questionnaire depends on its purpose and applicability. An easy scoring method and information about acceptable levels of missing data enhances applicability.

Interest in using patient based instruments in clinical practice for assessment and treatment monitoring of individual patients is growing. These instruments enable clinicians to detect and treat functional and psychological problems that previously may have been missed. Furthermore, they promote shared decision making and facilitate doctor-patient communication.[51] Questionnaires with fewer items and shorter administration may be more practical for routine use in clinical practice.[25] Clearly, questionnaires used for clinical assessment of individual patients demand higher measurement standards than those used in groups. For test-retest reliability, an ICC >0.70 was regarded as adequate for group comparisons, yet for individual comparisons an ICC of $\geqslant 0.90$ should be required.[20 23 25 27 52] This means that the SSRS and SPADI may not be applicable for individual patients. In addition, small confidence intervals around an individual patient score are needed to make the questionnaire useful for evaluating treatment results in individual patients. McHorney *et al* suggested that score confidence intervals must be fully documented before standardised health measures are routinely incorporated into clinical practice for assessment of individual patients.[53] Cook *et al* were the only authors who presented confidence intervals around the reliability coefficients. Their result showed wide confidence intervals for the reliability coefficients of both the SPADI and ASES.[54]

This review provides information for researchers and clinicians to facilitate the choice among the existing questionnaires for shoulder disability. The ''best'' scale is always best for a particular purpose, where purpose is defined by the disease, the population, and the treatment.[55] The DASH, SPADI, and ASES have been evaluated most often and, overall, the DASH received the best ratings for its clinimetric properties. The DASH and SPADI are recommended for evaluative purposes in outpatient clinics. These questionnaires received positive ratings for responsiveness and have no floor or ceiling effects. The SIQ is recommended for evaluation of patients with shoulder instability and the OSQ for evaluation of patients having a shoulder operation other than stabilisation. The SST is a short, unidimensional questionnaire that had an ICC of 0.99 for test-retest reliability in one study.[60] Hence, for discriminative purposes the SST is suggested for patients with shoulder complaints in general. The SSRS and SPADI should not be used for assessment of individual patients.

There are no standardised criteria to evaluate the quality of subjective health measurement questionnaires. The criteria we used to evaluate the quality of the questionnaires may be disputed. However, it was not our intention to create a standardised evaluation checklist, but to provide information about the questionnaires' clinimetric properties in order to facilitate the choice between questionnaires. Guidelines are needed to set standards and define the criteria by which these instruments should be assessed Continuing accumulation of research evidence for the clinimetric properties of a scale is important for demonstrating the scale's usefulness in both clinical practice and research applications.

Authors' affiliations

S D M Bot, C B Terwee, D A W M van der Windt, L M Bouter, J Dekker, H C W de Vet, VU University Medical Centre, Amsterdam, The Netherlands

APPENDIX 1

Table 3 shows the full names of the questionnaires included.

REFERENCES

1 **Higginson IJ**, Carr AJ. Measuring quality of life: using quality of life measures in the clinical setting. *BMJ* 2001;**322**:1297–300.
2 **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.
3 **Dawson J**, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br* 1999;**81**:420–6.
4 **Dawson J**, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. *J Bone Joint Surg Br* 1996;**78**:593–600.
5 **van der Heijden GJ**, Leffers P, Bouter LM. Shoulder disability questionnaire design and responsiveness of a functional status measure. *J Clin Epidemiol* 2000;**53**:29–38.
6 **Hollinshead RM**, Mohtadi NG, Van de 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**:373–81.
7 **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.

8 **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.
9 **Kohn D**, Geyer M. The subjective shoulder rating system. *Arch Orthop Trauma Surg* 1997;**116**:324–8.
10 **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.
11 **Lippitt SB**, Harryman DTI, Matsen FAI. A practical tool for evaluation of function: the simple shoulder test. In: Matsen FA III, Fu FH, Hawkins RJ, eds. *The shoulder: a balance of mobility and stability*. Rosemont, Illinios: The American Academy of Orthopaedic Surgeons, 1993:501–18.
12 **Lo IK**, Griffin S, Kirkley A. The development of a disease-specific quality of life measurement tool for osteoarthritis of the shoulder: the Western Ontario Osteoarthritis of the Shoulder (WOOS) index. *Osteoarthritis Cartilage* 2001;**9**:771–8.
13 **Patte D**. Directions for the use of the index severity for painful and/or chronically disabled shoulders. Abstracts of the First Open Congress of the European Society of Surgery of the Shoulder and the Elbow. 1987:36–41.
14 **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**:1195–202.
15 **Richards RR**, An K-N, 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.
16 **Roach KE**, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. *Arthritis Care Res* 1991;**4**:143–9.
17 **Simonsick EM**, Kasper JD, Guralnik JM, Bandeen-Roche K, Ferrucci L, Hirsch R, *et al*. Severity of upper and lower extremity functional limitation: scale development and validation with self-report and performance-based measures of physical function. *J Gerontol B Psychol Sci Soc Sci* 2001;**56**:S10–19.
18 **Croft P**. Measuring up to shoulder pain. *Ann Rheum Dis* 1998;**57**:65–6.
19 **Garratt A**, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. *BMJ* 2002;**324**:1417.
20 **Lohr KN**, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, *et al*. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther* 1996;**18**:979–92.
21 **Bombardier C**, Tugwell P. Methodological considerations in functional assessment. *J Rheumatol* 1987;**14**(suppl 15):6–10.
22 **Guyatt GH**, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med* 1993;**118**:622–9.
23 **Nunnally JC**. *Psychometric theory*, 2nd ed. New York: McGraw-Hill, 1978.
24 **Kirshner B**, Guyatt G. A methodological framework for assessing health indices. *J Chronic Dis* 1985;**38**:27–36.
25 **McHorney CA**, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res, 1995;**4**:293–307.
26 **Bland JM**, Altman DG. Measurement error and correlation coefficients. *BMJ* 1996;**313**:41–2.
27 **Fitzpatrick R**, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. *Health Technol Assess* 1998;**2**:i–iv, 1–74.
28 **Deyo RA**, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials* 1991;**12**(suppl):142–58S.
29 **Bland JM**, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**i**:307–10.
30 **Cohen J**. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960;**20**:37–46.
31 **de Bruin AF**, Diederiks JP, de Witte LP, Stevens FC, Philipsen H. Assessing the responsiveness of a functional status measure: the Sickness Impact Profile versus the SIP68. *J Clin Epidemiol* 1997;**50**:529–40.
32 **Terwee CB**, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res* 2003;**12**:349–62.
33 **Testa MA**, Simonson DC. Assesment of quality-of-life outcomes. *N Engl J Med* 1996;**334**:835–40.
34 **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.
35 **Gerber C**. Integrated scoring systems for the functional assessment of the shoulder. In: Matsen FA III, Fu FH, Hawkins RJ, eds. *The shoulder: a balance of mobility and stability*. Rosemont, Illinois: The American Academy of Orthopedic Surgeons, 1993:531–50.
36 **Romeo AA**, Bach BR Jr, O'Halloran KL. Scoring systems for shoulder conditions. *Am J Sports Med* 1996;**24**:472–6.
37 **Cook KF**, Gartsman GM, Roddey TS, Olson SL. The measurement level and trait-specific reliability of 4 scales of shoulder functioning: an empiric investigation. *Arch Phys Med Rehabil* 2001;**82**:1558–65.
38 **Curtis KA**, Roach KE, Applegate EB, Amar T, Benbow CS, Genecco TD, *et al*. Development of the Wheelchair User's Shoulder Pain Index (WUSPI). *Paraplegia* 1995;**33**:290–3.
39 **Davis AM**, Wright JG, Williams JI, Bombardier C, Griffin A, Bell RS. Development of a measure of physical function for patients with bone and soft tissue sarcoma. *Qual Life Res* 1996;**5**:508–16.
40 **Marino RJ**, Shea JA, Stineman MG. The capabilities of upper extremity instrument: reliability and validity of a measure of functional limitation in tetraplegia. *Arch Phys Med Rehabil* 1998;**79**:1512–21.
41 **Soldatis JJ**, Moseley JB, Etminan M. Shoulder symptoms in healthy athletes: a comparison of outcome scoring systems. *J Shoulder Elbow Surg* 1997;**6**:265–71.
42 **Salerno DF**, Franzblau A, Armstrong TJ, Werner RA, Becker MP. Test-retest reliability of the upper extremity questionnaire among keyboard operators. *Am J Ind Med* 2001;**40**:655–66.
43 **Winters JC**, Sobel JS, Groenier KH, Arendzen JH, Meyboom-De Jong B. A shoulder pain score: a comprehensive questionnaire for assessing pain in patients with shoulder complaints. *Scand J Rehabil Med* 1996;**28**:163–7.
44 **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.
45 **Reise SP**, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol Bull* 1993;**114**:552–66.
46 **Hays RD**, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. *Med Care* 2000;**38**(suppl):II28–42.
47 **McHorney CA**, Cohen AS. Equating health status measures with item response theory: illustrations with functional status items. *Med Care* 2000;**38**(suppl):II43–59.
48 **Guyatt GH**. Making sense of quality-of-life data. *Med Care* 2000;**38**(suppl):II175–9.
49 **Lydick E**, Epstein RS. Interpretation of quality of life changes. *Qual Life Res* 1993;**2**:221–6.
50 **Streiner DL**, Norman GR. *Health measurement scales: a practical guide to their development and use*, 2nd ed. Oxford: Oxford University Press, 1995.
51 **Greenhalgh J**, Meadows K. The effectiveness of the use of patient-based measures of health in routine practice in improving the process and outcomes of patient care: a literature review. *J Eval Clin Pract* 1999;**5**:401–16.
52 **Hays RD**, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* 1993;**2**:441–9.
53 **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.
54 **Cook KF**, Roddey TS, Olson SL, Gartsman GM, Valenzuela FF, Hanten WP. Reliability by surgical status of self-reported outcomes in patients who have shoulder pathologies. *J Orthop Sports Phys Ther* 2002;**32**:336–46.
55 **Hyland ME**. Recommendations from quality of life scales are not simple. *BMJ* 2002;**325**:599.
56 **van der Windt DA**, van der Heijden GJ, de Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. *Ann Rheum Dis* 1998;**57**:82–7.
57 **Beaton DE**, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the disabilities of the arm, shoulder and hand outcome measure in different regions of the upper extremity. *J Hand Ther* 2001;**14**:128–46.
58 **Skutek M**, Fremerey RW, Zeichen J, Bosch U. Outcome analysis following open rotator cuff repair. Early effectiveness validated using four different shoulder assessment scales. *Arch Orthop Trauma Surg* 2000;**120**:432–6.
59 **SooHoo NF**, McDonald AP, Seiler JG 3rd, McGillivary GR. valuation of the construct validity of the DASH questionnaire by correlation to the SF-36. *J Hand Surg Am* 2002;**27**:537–41.
60 **Beaton D**, Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. *J Shoulder Elbow Surg* 1998;**7**:565–72.
61 **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.
62 **Williams JW Jr**, Holleman DR Jr, Simel DL. easuring shoulder function with the Shoulder Pain and Disability Index. *J Rheumatol* 1995;**22**:727–32.
63 **Meenan RF**, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. *Arthritis Rheum* 1992;**35**:1–10.
64 **Constant CR**, Murley AH. A clinical method of functional assessment of the shoulder. *Clin Orthop* 1987:160–4.
65 **Rowe CR**, Patel D, Southmayd WW. The Bankart procedure: a long-term end-result study. *J Bone Joint Surg Am* 1978;**60**:1–16.
66 **Ware JE Jr**, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;**30**:473–83.
67 **Fries JF**, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;**23**:137–45.
68 **Bergner M**, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981;**19**:787–805.
69 **Amstutz HC**, Sew Hoy AL, Clarke IC. UCLA anatomic total shoulder arthroplasty. *Clin Orthop* 1981:7–20.
70 **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**:1136–44.
71 **Guyatt G**, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987;**40**:171–8.

Appendix W1. Checklist for rating the clinimetric quality of self-assessment questionnaires

Clinimetric property	Definition	Criteria used to rate the clinimetric quality
Content validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire.	1) patients were involved during item selection and/or item reduction 2) patients were consulted for reading and comprehension. Rating: + patients and (investigator or expert) involved ± patients only - no patient involvement ? no information found on content validity
Readability & comprehension	The questionnaire is understandable for all patients	Rating: + readability tested; result was good - inadequate readability ? no information found on readability and comprehension
Internal consistency	The extent to which items in a (sub)scale are intercorrelated; a measure of the homogeneity of a (sub)scale	1) Factor analysis was applied in order to provide empirical support for the dimensionality of the questionnaire. 2) Cronbach's alpha between 0.70 and 0.90 for every dimension/subscale Rating: + adequate design & method; factor analysis; alpha 0.70-0.90 ± doubtful method used - inadequate internal consistency ? no information found on internal consistency
Construct validity	The extent to which scores on the questionnaire relate to other measures in a manner that is consistent with theoretically derived hypothesis concerning the domains that are measured.	1) hypotheses were formulated 2) results were acceptable in accordance with the hypotheses 3) an adequate measure was used Rating: + adequate design, method, and result ± doubtful method used - inadequate construct validity ? no information found on construct validity
Floor & ceiling effects	The questionnaire fails to demonstrate a worse score in patients clinically deteriorated and an improved score in patients who clinically improved	1) descriptive statistics of the distribution of scores were presented 2) 15% of respondents achieved the highest or lowest possible score Rating: + no floor / ceiling effects - more than 15% in extremities ? no information found on floor and ceiling effects
Test-retest reliability	The extent to which the same results are obtained on repeated administrations of the same questionnaire when no change in physical functioning has occurred	1) calculation of an intraclass correlation coefficient (ICC); ICC > 0.70 2) time interval and confidence intervals were presented Rating: + adequate design, method, and ICC > 0.70 ± doubtful method was used - inadequate reliability ? no information found on test-retest reliability
Agreement	the ability to produce exactly the same scores with repeated measurements	1) for evaluative questionnaires reliability agreement should be assessed 2) limits of agreement, Kappa, or standard error of measurement (SEM) was presented Rating: + adequate design, method and result ± doubtful method used - inadequate agreement ? no information found on agreement
Responsiveness	The ability to detect important change over time in the concept being measured	1) for evaluative questionnaires responsiveness should be assessed 2) hypotheses were formulated and results were in agreement 3) an adequate measure was used (ES, SRM, comparison with external standard) Rating: + adequate design, method and result ± doubtful method used - inadequate responsiveness ? no information found on responsiveness
Interpretability	The degree to which one can assign qualitative meaning to quantitative scores	Authors provided information on the interpretation of scores: 1. presentation of means and SD of scores before and after treatment 2. comparative data on the distribution of scores in relevant subgroups 3. information on the relationship of scores to well-known functional measures or clinical diagnosis 4. information on the association between changes in score and patients' global ratings of the magnitude of change they have experienced Rating: + 2 or more of the above types of information was presented ± doubtful method used or doubtful description ? no information found on interpretation
Minimal clinically important difference (MCID)	The smallest difference in score in the domain of interest which patients perceive as beneficial and would mandate a change in patient's management	Information is provided about what (difference in) score would be clinically meaningful. Rating: + MCID presented - no MCID presented
Time to administer	Time needed to complete the questionnaire	Rating: + less than 10 minutes - more than 10 minutes ? no information found on time to complete the questionnaire
Administration burden	Ease of the method used to calculate the questionnaire's score	Rating: + easy: summing up of the items ± moderate: visual analogue scale (VAS) or simple formula - difficult: VAS in combination with formula, or complex formula ? no information found on rating method

Table W1: content and construct validity of the shoulder disability questionnaires

questionnaire		*content validity* item selection*	item reduction*	level of reading examined*	dimensionality studied?	internal consistency	*construct validity* hypothesis	(main) results	floor / ceiling effect	study size
SDQ-UK	w1	patients experts investigator	no	no	?	?	yes	score GP-patients > score community ; restricted ROM -> higher disability	ceiling†	54; 67
SIQ	w2	patients investigator	yes	yes	?	α = 0.91	yes	Constant: r = -0.56 Rowe: r = -0.51 SF36 physical: r = -0.71	no	92
OSQ	w3	patients investigator	yes	yes	?	α = 0.89	yes	Constant: r = -0.74 SF36 physical: r = -0.61 HAQ disability: r = 0.86	no	111
SDQ-NL	w4	experts investigator	yes	yes	?	?	?	?	ceiling‡	180
RC-QOL	w5	patients experts investigator	yes	yes	?	?	yes	SF36: r = 0.78; ASES: r = 0.84	no	70
DASH	w6	patients	yes	yes	?	?				
	w7	experts					yes	SPADI function: r = 0.85	no	138
	w8	investigator					yes	SF36 physical: r=-0.73	no	23
	w9						no	Constant: r = -0.76	?	23
WOSI	w10	patients experts investigator	yes	yes	?	?	yes	DASH: r = 0.77; Constant: r = 0.59; Rowe: r = 0.61; ASES: r = 0.55; SF12 physical: r = 0.66	?	47
SSRS	w11	investigator	no	no	?	?	no	Constant: r = 0.83	no	200
	w12						yes	SF36 physical: r = 0.12; SST: r = 0.47; SPADI: r = 0.50; m-ASES: r = 0.50; SSI: r = 0.48	no	90
SRQ	w13	patients investigator	yes	yes	?	α = 0.77-0.90¶	yes	AIMS: r = -0.84	?	97
SST	w14	patients	no	no						
	w12	investigator					yes	SF36 physical: r = 0.58; SSRS: r = 0.47; SPADI: r = 0.74; m-ASES: r = 0.73; SSI: r = 0.80	no	90
	w9						no	Constant: r = 0.49	?	23
	w15				yes	α = 0.85	no	SPADI: r = -0.80	?	192

questionnaire		*content validity* item selection*	item reduction*	level of reading examined*	dimensionality studied?	internal consistency	*construct validity* hypothesis	(main) results	floor / ceiling effect	study size
WOOS	w16	patients experts investigator	yes	yes	?	?	yes	Constant: r = 0.73; ASES: r = 0.59; SF12 physical: r = 0.65	?	41
SSI	w12	?	?	?	?	?	yes	SF36 physical: r = 0.59; SSRS: r = 0.48; SST: r = 0.80; SPADI: r = 0.79; m-ASES: r = 0.79	no	90
UEFS	w17	experts	yes	no	yes	α = 0.83-0.93**	no	significant difference between levels of severity	no	
ASES	w18	experts	yes	no	?					
	w9	investigator					no	Constant: r = 0.87	?	23
	w12						yes	SF36 physical: r = 0.60; SST: r = 0.73; SSRS: r = 0.50; SPADI: r =0.77; SSI: r = 0.79	no	90
	w10						yes	WOSI: r = 0.55	?	
	w16						yes	WOOS: r = 0.59	?	
	w19						no	Rowe: r = 0.82; UCLA: r = 0.50	?	
	w20					α = 0.90				
SPADI	w21	experts	yes	no	yes	α = 0.93††	yes	ROM: r = -0.54 - -0.80	?	37
	w22						no	HAQ: r = 0.61; SF-20 physical: r = -0.50	no	102
	w23						yes	SIP: r = 0.21 - 0.57	no	94
	w15				yes	α = 0.95††	no	UCLA function: r = -0.64	?	192
	w12						yes	SF36 physical: r = 0.58; SST: r = 0.74; SSRS: r = 0.50; m-ASES: r = 0.77; SSI : r = 0.79	no	90
	w7						yes	DASH: r = 0.85	?	138
	w20					α = 0.94††				
UEFL	w24	investigator	no	no	yes	?	yes	prevalence of self-reported difficulty at each level of functional limitation	floor	1002

* results based on first reference; † ceiling effect in community sample with shoulder disorders; ‡ ceiling effect in people who got physiotherapy treatment for soft tissue disorders; floor effect in healthy community dwelling woman and moderately to sever disabled woman (age > 65 years); ¶ subscales daily activities, recreational and athletic activities, and work; ** range across study groups; †† value of subscale disability; α = chronbach's alpha; ? = no data published; r = correlation coefficient; AIMS = Arthritis Impact Measurement Scales Health Status Questionnaire [w25]; Constant = Constant Score [w26]; Rowe Rating Scale [w27]; SF-36 = Medical Outcome Study Short-Form 36 [w28]; HAQ = Health Assessment Questionnaire [w29]; SIP = Sickness Impact Profile [w30]; UCLA = University California - Los Angeles Shoulder Scale [w31;w32]

Table W2: reproducibility of the shoulder disability questionnaires

questionnaire		reliability (I)	time interval	agreement (II)	study size*
SDQ-UK	w1	?	?	?	?
SIQ	w2	r = 0.97	24 hours	CoR = 5.7	34 (I, II)
OSQ	w3	?	24 hours	CoR = 6.8; MD = -0.12 (out of score 1-5)	60 (II)
SDQ-NL	w4	?	?	?	?
RC-QOL	w5	?	2 weeks	MD = 5.05 (out of score 0-100)	30 (II)
DASH	w6	ICC = 0.96	3-5 days	SEM = 4.6 (score 0-100)	73 (I); 56 (II)
WOSI	w10	ICC = 0.91†	2 weeks	?	51 (I)
SSRS	w33	ICC = 0.71	1 week	% = 63 [71]	41 (I, II)
SRQ	w13	r = 0.89-0.96‡	± 3 days	Kappa = 0.73 - 0.97	40 (I)
SST	w33	ICC = 0.99	1 week	% = 80 [95]	41 (I, II)
	w15			SEM = 11.65 (score 0-100)	192 (II)
WOOS	w16	ICC = 0.94†	3 months	?	22 (I)
SSI	w33	ICC = 0.97	1 week	% = 24 [NA]	41 (I, II)
UEFS	w17	?	?	?	
ASES	w33	ICC = 0.96	1 week	% = 31 [51]	41 (I, II)
	w20	ICC = 0.78 (0.59-0.89) post-surgical ; ICC = 0.86 (0.72-0.94) non-surgical	1 week		31(I) 25 (I)
SPADI	w21	ICC = 0.64¶	24 hours		23 (I)
	w15			SEM = 5.78¶	192 (II)
	w33	ICC = 0.91	1 week	% = 5 [23]	41 (I, II)
	w20	ICC = 0.57 (0.27-0.77) post-surgical ; ICC = 0.84 (0.66-0.92) non-surgical	1 week		31 (I) 25 (I)
UEFL	w24	?	?	?	?

* study size for study of study of reliability (I) and study of agreement (II); † subscale "sport, recreation and work"; ‡ subscales "daily activities", "recreational and athletic activities", and "work"; value of subscale function/disability with confidence intervals (in brackets) for post- and non-surgical patients"; ¶ value of subscale disability; ? = no data published; r = correlation coefficient; CoR = coefficient of reliability [w34]; MD = mean difference; SEM = standard error of measurement; Kappa = the proportion of the observed agreement that exceeds the agreement that is expected by chance alone; % perfect agreement: percent of subjects having identical scores; % perfect agreement within 1 response category (in brackets); NA = not applicable

Table W3: responsiveness and interpretability of the shoulder disability questionnaires

questionnaire		*responsiveness* treatment	time to follow-up	Hypothesis	(main) results	study size	*interpretability* attention for interpretability	baseline and follow up scores	scores of relevant subgroups	MCID
SDQ-UK	w1	?		?	?	?	no	?	?	no
SIQ	w2	physiotherapy / surgery	6 months	no	ES = 0.8 sign. difference between improved - not-improved	64	no	36.6; 95%CI: 34.4-38.8 (baseline); 28.3; 95%CI: 35.6-31.1 (follow-up)*	comparison of change scores with regard to the patients assessment of change	no
OSQ	w3	surgery	6 months	no	ES = 1.2 sign. difference between improved - not-improved	56	no	36.3; 95%CI: 34.6-37.9 (baseline); 26.0; 95%CI: 23.0-28.9 (follow-up)*	comparison of change scores with regard to the patients assessment of change	no
SDQ-NL	w4	physiotherapy	6 weeks	no	CRR = 1.14 ROC curve; AUC = 0.72	180	no	74 (63, 85) (stable); 70 (58, 78) (improved)†	median + percentiles of stable / improved patients for shoulder pain, chief complaint, symptoms and mobility	no
	w35	general practice	1 and 6 months	no	MCS: 20; 35 CRR = 2.22; 1.89 ROC curve; AUC = 0.84; 0.88	308	yes	67 (±23) (baseline); 47 (±31) (1 month); 32 (±31) (6 months)‡	mean change scores for clinical stable, improved and deteriorated patients	3 items
RC-QOL	w5	surgery	42 months (range 25-71)	?	?	?	no	69.9 (4.4-100) (follow up)	scores of large and massive rotator cuff tears	no
DASH	w7	surgery	3 months	yes	MCS: -13.4 (SD 16.6) SRM = 0.81 ES = 0.64 functional status: r =0.69 ROC curve	138	no	48.8 (±21.0) (baseline); 35.3 (± 21.3) (follow-up)‡	mean + SD wrist/hand patients; transition scale; comparison of change scores with regard to the patients assessment of change	no
	w6	?	3 months	yes	SRM =0.71 WOSI: r = 0.76	47	no	?	?	NA
	w9	surgery	57.8 weeks (±15.7)‡	no		23	no	49.6 (±8.5) (pre-operative); 21.6 (±13.0) (post-operatiave)	?	no
WOSI	w10	non specified treatment OA	3 months	yes	SRM = 0.93 DASH: r = 0.76 Constant: r =0.69 ASES: r = 0.50	47	no	?	?	no
SSRS	w11	surgery	3 and 12 months	no		?	yes	47(pre), 83 (post)¶; SA** 72 (pre),95 (post)¶; Bankart** 42 (pre), 52 (post)¶; MUA**	median+range diagnostic groups; comparison of change scores with regard to the patients assessment of treatment results	no
	w33	surgery	6 months	yes	MCS: 16.4 SRM = 0.65	33	no	52.2 (baseline); 69.1 (follow up)	?	no
SRQ	w13	surgery	12 months	no	SRM = 1.9 (1.1 - 1.8)* MCS: 26.7 (1.7 - 4.9)* IoR = 1.6 (1.1 - 2.0)*	30	yes	61.6 (±13.4) (pre-operative); 88.3 (±10.0) (post-operative)‡	overall score, scale-scores; initial score; score at one year follow up	2 points / domain

questionnaire		*responsiveness* treatment	time to follow-up	hypothesis	(main) results	study size	*Interpretability* attention for interpretability	baseline and follow up scores	scores of relevant subgroups	MCID
SST	w14			no	% progress per item	9 - 29	yes	% item score	% score diagnostic groups	no
	w33	surgery	6 months	yes	MCS: 17.2 SRM = 0.87	33	no	36.0 (baseline); 53.8 (follow up)	?	no
	w15	surgery	57.8 weeks (±15.7)‡	no		23	no	3.30 (±1.82) (pre-operative), 6.97 (±1.80) (post-operative)	?	no
WOOS	w16	surgery	3 months	yes	SRM = 1.91 Constant: r = 0.69 ASES: r = 0.43	41	no	?	?	no
SSI	w33	surgery	6 months	yes	MCS: 20.1 SRM = 1.05	33	no	47.0 (baseline); 67.3 (follow up)	?	no
UEFS	w17	?	19 months (12-24)	no	SRM = -1.33 average pain: r = 0.58	16	no	43.3 (3.3-75.9) (baseline); 31.5 (0.0-62.0) (follow up)	working status; duration symptoms	no
ASES	w9	surgery	57.8 weeks (±15.7)‡	no		23	no	33.9 (± 15.9) (pre-operative); 71.9 (± 16.8) (post-operative)	?	no
	w33	surgery	6 months	yes	MCS: 17.6 SRM = 0.93	33	no	49.4 (baseline); 68.0 (follow up)	?	no
	w10	non-defined treatment OA	3 months	yes	SRM = 0.54 WOSI: r = 0.50	47	no	?	?	no
	w16	surgery	6 months	yes	SRM = 1.29 WOOS: r = 0.43	41	no	?	?	no
	w20	?	?	?	?	31; 25	no	65.7 (± 22.7) (post-surgical); 66.4 (± 22.9) (non-surgical)‡	?	
SPADI	w21	medication or injection	30 days	no	ROM: r = -0.52 - -0.70 MCS: -25.3††	30	no	?	?	no
	w22	?	2, 4 and 12 weeks	no	overall status: r = 0.73; r = 0.76; r = 0.79 ROC curve; AUC = 0.91	75	yes	57.6 (22.5) (baseline)‡; -21.9-6.5 (change score)	change score 2-4-12 weeks / overall status (improved, same, worse).	>10 points
	w23	physiotherapy	±10 weeks	no	MCS: -28.4†† SRM = 1.04)††	34	no	33.9 (±28.1) (baseline); -28.4 (±27.2) (change score)‡	consensus between therapist and patient judgement on meaningful improvement in shoulder function	no
	w33	surgery	6 months	yes	MCS: 25.6 SRM = 1.23	33	no	39.9 (baseline); 66.4 (follow up)	?	no
	w7	surgery	3 months	yes	SRM = 0.71††	138	no	?	?	no
	w20	?	?	?	?	31; 25	no	28.5 (±25.6) (post-surgical); 47.9 (±24.6) ((non-surgical)‡	?	
UEFL	w24	?		?	?	?	no	?	?	no

* mean and 95% confidence interval; † mean and 25th and 75th percentiles;‡ mean and SD; mean and range; ¶ median score pre-operatiave (pre) and post-operative (post); ** SA = subacromial decompression; Bankart = Bankart repair of anterior shoulder reconstruction; MUA = manipulation under anesthesia; †† disability scale; MCID = minimal clinically important difference; IoR = Index of Responsiveness [w36]; CRR = calibrated responsiveness ratio; MCS = mean change score; SRM = standardized response mean; ES = effect size

Reference List

w1. 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.

w2. Dawson J, Fitzpatrick R, Carr A. The assessment of shoulder instability. The development and validation of a questionnaire. *J Bone Joint Surg Br* 1999;**81**:420-6.

w3. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. *J Bone Joint Surg Br* 1996;**78**:593-600.

w4. Heijden GJvd, Leffers P, Bouter LM. Shoulder disability questionnaire design and responsiveness of a functional status measure. *J Clin Epidemiol* 2000;**53**:29-38.

w5. 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**:373-81.

w6. 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.

w7. Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. *J Hand Ther* 2001 ;**14**:128-46.

w8. SooHoo NF, McDonald AP, Seiler JG, McGillivary GR. Evaluation of the construct validity of the DASH questionnaire by correlation to the SF-36. *J Hand Surg [Am]* 2002;**27**:537-41.

w9. Skutek M, Fremerey RW, Zeichen J, Bosch U. Outcome analysis following open rotator cuff repair. Early effectiveness validated using four different shoulder assessment scales. *Arch Orthop Trauma Surg* 2000;**120**:432-6.

w10. 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.

w11. Kohn D,.Geyer M. The subjective shoulder rating system. *Arch Orthop Trauma Surg* 1997;**116**:324-8.

w12. 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.

w13. 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.

w14. Lippit SB, Harryman DTI, Matsen FAI. A practical tool for evaluation of function: the simple shoulder test. *The Shoulder: a Balance of Mobility and Stability.Matsen FA III, FU FH, Hawkins RJ (ed).Rosemont, Illinios, The American Academy of Orthopaedic Surgeons* 1993;501-18.

w15. 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.

w16. Lo IK, Griffin S, Kirkley A. The development of a disease-specific quality of life measurement tool for osteoarthritis of the shoulder: The Western Ontario Osteoarthritis of the Shoulder (WOOS) index. *Osteoarthritis Cartilage* 2001;**9**:771-8.

w17. 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**:1195-202.

w18. Richards RR, An K-N, Bigliani LU, Friedman RJ, Gartsman GM, Gristina AG *et al.* A standardized method for the assessment of shoulder function. *Journal of Shoulder and Elbow Surgery* 1994;**3**:347-52.

w19. Romeo AA, Bach BR, O'Halloran KL. Scoring systems for shoulder conditions. *Am J Sports Med* 1996;**24**:472-6.

w20. Cook KF, Roddey TS, Olson SL, Gartsman GM, Valenzuela FF, Hanten WP. Reliability by surgical status of self-reported outcomes in patients who have shoulder pathologies. *J Orthop Sports Phys Ther* 2002;**32**:336-46.

w21. Roach KE, Budiman-Mak E, Songsiridej N, Lertratanakul Y. Development of a shoulder pain and disability index. *Arthritis Care Res* 1991;**4**:143-9.

w22. Williams JW, Holleman DR, Simel DL. Measuring shoulder function with the Shoulder Pain and Disability Index. *J Rheumatol* 1995;**22**:727-32.

w23. 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.

w24. Simonsick EM, Kasper JD, Guralnik JM, Bandeen-Roche K, Ferrucci L, Hirsch R *et al*. Severity of upper and lower extremity functional limitation: scale development and validation with self-report and performance-based measures wof physical function. *J Gerontol B Psychol Sci Soc Sci* 2001;**56**:10-9.

w25. Meenan RF, Mason JH, Anderson JJ, Guccione AA, Kazis LE. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. *Arthritis Rheum* 1992;**35**:1-10.

w26. Constant CR,.Murley AH. A clinical method of functional assessment of the shoulder. *Clin Orthop* 1987;160-4.

w27. Rowe CR, Patel D, Southmayd WW. The Bankart procedure: a long-term end-result study. *J Bone Joint Surg Am* 1978;**60**:1-16.

w28. Ware JE,.Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;**30**:473-83.

w29. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;**23**:137-45.

w30. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. *Med Care* 1981;**19**:787-805.

w31. Amstutz HC, Sew Hoy AL, Clarke IC. UCLA anatomic total shoulder arthroplasty. *Clin Orthop* 1981;7-20.

w32. Ellman H, Hanker G, Bayer M. Repair of the rotator cuff. End-result study of factors influencing reconstruction. *J wBone Joint Surg Am* 1986;**68**:1136-44.

w33. Beaton D,.Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. *J Shoulder Elbow Surg* 1998;**7**:565-72.

w34. Bland JM,.Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1**:307-10.

w35. Windt DAvd, Heijden GJvd, Winter AFd, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. *Ann Rheum Dis* 1998;**57**:82-7.

w36. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987;**40**:171-8.