Evaluation of Shoulder-Specific Patient-Reported Outcome Measures: A Systematic, Standardized Comparison of Available Evidence

Stefanie Schmidt[1,2,4], Kalliopi Vrotsou[5], Antonio Escobar[5], Marta González, Esther Villalonga[1], Silvia López, Nerea González, Amado Ribero, Miren Orive Caldaza, Carlota Lashayas, José Maria Valderas[6], Jordi Alonso[1,2,4], Montse Ferrer[1,3,4] in the name of the Scientific Committee of BIBLIOPRO.

Affiliation:

1 IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain

2 Universitat Pompeu Fabra (UPF), Barcelona, Spain

3 Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

4 CIBER Epidemiología y Salud Pública (CIBERESP), Spain

5 REDISSEC, Bilbao, Spain

6 Health Services and Policy Research Group, Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

FUNDING: This study was funded by the Department of Health, Government of the Basque Country (Project Nº 2010111156).

Corresponding author:

Montse Ferrer

Health Service Research Unit, IMIM (Hospital del Mar Medical Research Institute)

Doctor Aiguader, 88, 08003 Barcelona, Spain

Telephone +34 933 160 763, Fax +34 933 160 797, Email: mferrer@imim.es

Background:

The shoulder is one of the most complex joints of the human body. Shoulder related disorders account for substantial medical, economic and social costs[19,42,40] and comprise a wide spectrum of problems. Shoulder problems are mostly accompanied by pain and a restricted movement of the hand, arm or shoulder that leads to difficulties in performing certain activities.[19,32,1] A recent research suggests that shoulder pain not only affects function in work and leisure time activities, but may also interfere with psychological and social wellbeing.[28] A systematic review showed that the estimated prevalence of shoulder pain in the general population varies greatly among studies, with a lifetime prevalence from 7 to 67%.[22] In fact, shoulder or neck pain is one of the most frequent work-related complaints and a frequent reason for work absent.[24] Data from a study conducted in the Netherlands showed that 30% of workers with shoulder pain reported any sick leave during the 6 month follow-up time.[17]

There are different ways to assess the impact of shoulder disorders. Traditionally, it has been evaluated locally, observing the range of motion, strength or pain, whereas today, the research aims to determine the overall impact on the body by focussing on the person's functioning in daily life activities and how their psychological wellbeing is affected.[3] This subjective information given by the patient is obtained by so called Patient-Reported Outcome (PRO) measures. These PRO measures generally focus on the assessment of physical function, psychosocial issues, or simply, quality of life, and try to capture the possible effect of a disease or an intervention by incorporating the experience and perception of the patient.[4,38] Numerous generic or disease-specific PRO measures exist,[11] some with a similar purpose, content and applicability issues, but yet slight differences might exist; thus they need to be balanced against each other

regarding their strengths and weaknesses. For example, some of the shoulder-specific PRO measures have been designed for the whole upper extremity, and others independently of the underlying condition- (e.g. shoulder instability), whereas some are shoulder disease- (e.g. rotator cuff disease, osteoarthritis) or population- (e.g. wheelchair users) specific.[44,23,9] So it is a hard work to select the right PRO measure for a certain purpose among all those available.

PRO measurement requires reliable and valid measures. Outcome measures must be adequately selected regarding the individual study purpose, setting and the available resources. Direct comparison among them regarding some of their performance characteristics, like measurement model, metric properties and administration issues, can facilitate this task. Some efforts have been undertaken to classify or evaluate shoulder-specific PRO measures regarding those characteristics,[35,2,36,27,3,25,14,31] but yet no one examined neither the whole spectrum of those characteristics nor have undertaken a direct comparison among shoulder-specific PRO measures.

The *Evaluating Measures of Patient-Reported Outcomes* (EMPRO) tool was developed to facilitate a standardized, comprehensive, and comparative evaluation of PRO measures.[39] It combines three fundamental requirements: a) well described and established quality attributes for assessment; b) expert reviewers to conduct the assessment, and c) scores which allow direct comparisons among outcome measures. EMPRO is based on an exhaustive series of recommendations regarding the ideal attributes of PRO measures,[37] and has been shown to be valid[39] and useful (REF empro Prostate Cancer & empro Heart Failure).

The aim of this study was to obtain a standardized expert evaluation of the available evidence on development process, metric properties and administration issues of multi-item shoulder-specific PRO measures that are applicable to a wide spectrum of shoulder disorders. Our results should help clinicians and researchers to select the most appropriate shoulder-specific PRO measure used in patients with functional limitations due to shoulder disabilities and those applicable to a wide range of shoulder disorders.

Methods:

Identification of shoulder-specific PRO measures and their relevant information

We carried out a systematic literature review in the PubMed database (March 2011) to obtain all the available published evidence. We combined keywords using MeSH terms and free-text entries: *(Shoulder* or *Shoulder Joint* or *Shoulder Pain* or *Rotator Cuff)* and *(Quality of Life* or *Questionnaires* or *Disability Evaluation* or *Cross-Cultural Comparison).* Articles were eligible for inclusion if they contained information regarding the development process, the metric properties or administration issues of multi-item shoulder-specific PRO measures. We excluded articles about PRO measures designed for: musculoskeletal conditions in general, upper extremity as a whole, specific shoulder conditions (like osteoarthritis or instability), specific populations (like wheelchair users or athletes), and systemic diseases (like breast or oral cancer). We furthermore excluded research protocols, congress abstracts, and secondary research articles.

In a three-step process, titles, abstracts and full-text articles were independently reviewed by two investigators. A third investigator was determined to mediate and resolve possible discrepancies found in each of the steps. Additionally, we examined manually the bibliographic reference lists of the articles selected for full review in order to complete the search.

Evaluating Measures of Patient-Reported Outcomes (EMPRO)

EMPRO[39] was designed to measure the quality of PRO measures. It is composed of eight attributes and 39 items, and assesses how well the development process of the outcome measure was and how it is described ("conceptual and measurement model"),

how well it performs in terms of metric properties ("reliability", "validity", "responsiveness to change", and "interpretability"), as well as administrative issues ("burden", "alternative modes of administration", and "cross-cultural and linguistic adaptations").

All EMPRO attributes and items are accompanied by a short description to explain on what the expert should focus on, and to facilitate the understanding of the intended meaning of each item in the evaluation process in order to guarantee standardization. Agreement with each item can be made on a 4-point Likert scale, from 4 (strongly agree) to 1 (strongly disagree). Experts can check the "no information" box, in case of insufficient information. Five items allow replying with "not applicable". Experts are asked to provide detailed comments to justify their ratings on each item. These comments were considered in the interpretation of the EMPRO scores to better reflect the scores meaning and prevent from misinterpretation.

Standardized expert evaluation

Each shoulder-specific PRO measures was assigned to 2 different experts. Experts were identified and invited because of their expertise and experience in PRO measurement (6 belonged to the EMPRO tool development working group and 16 had previously been accredited as EMPRO experts by undergoing a training course). In order to minimize the potential for bias, experts were neither authors nor had been involved in the development, evaluation or adaptation process of any of these evaluated instruments.

The EMPRO evaluation process consisted of two consecutive rounds. In the first round, every expert evaluated the assigned shoulder-specific PRO measure independently by

reviewing the provided full-text articles that were identified in the systematic literature review and applied the EMPRO tool.[39] In the second round, each expert was provided with the rating results of the other expert of the instrument both had evaluated. In case of discrepancies, they were invited to resolve those through discussion in order to reach a consensus. A third reviewer was available if needed to solve discrepancies.

Statistical analysis:

The attribute-specific scores were obtained by calculating the response mean of the applicable items when at least 50% of the items were rated. Items for which the response option "no information" had been selected a score of 1 (lowest possible score) was assigned. The scores were then linearly transformed to a range of 0 (worst possible score) to 100 (best possible score). Separate subscores for the "reliability" and "burden" attributes were calculated as those attributes are composed of two components, "internal consistency" and "reproducibility", and "respondent" and "administrative", respectively. For the reliability attribute, the highest subscore was then chosen to represent the total score for that attribute. In addition to the attribute-specific scores, we calculated an overall score that consisted of the mean of five metric related attributes: "conceptual and measurement model", "reliability", "validity", "responsiveness to change" and "interpretability". If any of these attribute scores is missing because not enough information was available, a cero was assigned. The overall score was only calculated when at least three of these five attributes had a rating. EMPRO scores were considered reasonably acceptable (REF HF & PC) if they reached at least 50 points (half the maximum score). Analysis was done with SPSS statistics version 12 and graphics were designed with Microsoft Excel 2003.

Results:

We identified 2325 articles in our systematic literature search (Figure 1). After the title review we excluded 1726 articles because they were not topic related. Abstracts were reviewed, and a further 222 articles were excluded: 111 did not contain any PRO measure; 40 only used generic PRO measures; 33 because they were secondary research literature; 30 included disease-specific outcome measures other than shoulder disorders; and 8 were lacking of information on development process, metric properties or administration issues. We identified 377 articles with information concerning 52 different instruments. After applying defined exclusion criteria, 263 articles related to 41 outcome measures were excluded, mostly because they were only applicable to patients with a specific-shoulder condition (11), they were not patient-reported (9) or not shoulder-specific (5). Instead, by revising the bibliographic lists of identified articles we included 8 additional articles that entered the inclusion criteria. Finally, 108 articles provided information about the development process, metric properties or administration issues of 11 shoulder-specific PRO measures at the end of the review process.

Eleven shoulder-specific PRO measures together with their instrument-specific information were identified and evaluated with EMPRO (Table 1). The number of published articles identified to be included varied from 2 to 27. The instruments were developed between 1987 and 2003 in order to be applicable to a variety of shoulder disorders. Seven out of eleven outcome measures are unidimensional; the others include 2 to 7 subdimensions. Their content includes mainly pain and function, assessed by the evaluation of daily life activities. The broader focused outcome measures additionally may include psychosocial issues (appetite or social contacts) or satisfaction. Answer

options are based on dichotomous scales (Yes/No answer options), Likert, numeric or visual analogue scales. The number of items included varies from 5 to 30. The time to complete takes between less than 3 minutes to less than 10 minutes and the period of assessment ranges from the last 24 hours to the last month.

The detailed EMPRO results are presented in Table 2 and summarized graphically in Figure 2. Final EMPRO scores were achieved by consensus rating between the two experts for every outcome measure; the third reviewer for discrepancy resolution was not needed at any time. The overall summary scores oscillated between 77.4 and 26.7 points. Thereby, six out of eleven shoulder-specific PRO measures presented scores above the threshold of 50 points, thus presenting acceptable overall results: the American Shoulder and Elbow Surgeons shoulder assessment – patient self-evaluation section (ASES-p), the Simple Shoulder Test (SST), the Oxford Shoulder Score (OSS), the Flexilevel Scale of Shoulder Function (FLEX-SF), the Shoulder Pain and Disability Index (SPADI), and the Dutch Shoulder Disability Questionnaire (SDQ-NL). The Appendix List shows the articles used in the EMPRO evaluation.

The “conceptual and measurement model” scores ranged from 81 to 14.3, whereby ASES-p (81 points), OSS; FLEX-SF and SDQ-NL (each 66.7 points) reached the highest scores. Instead four shoulder-specific PRO measures scored below 50 and for the Penn Shoulder Score (PSS) we could not find sufficient information to calculate this attribute. Eight of the outcome measures were judged to be reliable, with “reliability” scores ranging from 83.3 (SPADI) to 50 (Shoulder Rating Questionnaire - SRQ). The SDQ-NL and the Subjective Shoulder Rating System - SSRS) scored low (41.6 points), and for the United Kingdom Shoulder Disability Questionnaire (SDQ-UK) we could not

find sufficient information to calculate a "reliability" score. "Validity" scores in general were quite high. The SDQ-NL reached the highest rating (93.4), followed by the ASES-p, the FLEX-SF and the SST (all ≥ 80 points). Also the OSS and the SPADI showed to be valid instruments (75 and 66.6 points, respectively). The Subjective Shoulder Rating System (SSRS), as well as the Shoulder Rating Questionnaire (SRQ) scored below the threshold. For the PENN we could not find sufficient information to calculate a score. The "responsiveness to change" attribute scores were also high and ranged from 100 (SST and SDQ-NL) to 33.3 (FLEX-SF). The FLEX-SF received its worst result for this attribute; in contrast, the SDQ-UK scored surprisingly high here (88.9 points). Seven out of the eleven instruments presented information to evaluate its "interpretability", but only four presented acceptable information: the ASES-p and the OSS (66.7 points), as well as the SST and the FLEX-SF (55.6 points).

In the "burden" attribute (Table 2), the SDQ-NL reached the maximum score (100 points), whereas the ASES-p, OSS, PSS, SDQ-NL, SSRS and SST also presented acceptable EMPRO scores (91.7-66.7 points), meaning that they either present a low respondent or administrative burden. The attribute "alternative forms of administration" was only applicable for the FLEX-SF and the SPADI, which developed, respectively, a computer adaptive test version [7] and a telephone-interview version [43]. For the other evaluated shoulder-specific PRO measures only the original self-administered version exists. Finally, the attribute "cross-cultural & linguistic adaptation" (3 items) was not evaluated here because our study did not aim to evaluate the specific quality of country-specific instrument versions. Nevertheless, articles reporting on the instruments' cross-cultural and linguistic validation (e.g. Arabic,[45] Italian,[29] German,[13] Portuguese,[15] and

233 Turkish[5] ASES-p versions), as well as the metric properties of these new versions were

234 considered in our EMPRO evaluation, but not evaluated separately.

235

Discussion:

In this study we assessed the quality of multi-item shoulder-specific PRO measures that are designed for patients with a wide spectrum of shoulder disorders by evaluating conceptual, metric and administrative characteristics. Twenty-two experts in PRO measurement assessed the 11 identified outcome measures and the best rated following EMPRO standard criteria were the ASES-p, SST, and OSS. Acceptable results were also found for 3 other questionnaires, the FLEX-SF, SPADI, and SDQ-NL. All these 6 instruments are relatively short and easy to administer, but some of them failed in providing good or sufficient information on specific attributes which are detailed in the following.

The ASES-p obtained the best overall score (around 80 points) followed by SST and OSS (both around 70 points). The ASES-p was always among the top 3 outcome measures in the 5 attributes that were used for the overall score calculation; except for the "responsiveness" attribute, where it obtained the forth place due to little information about stable group comparison. The ASES-p scored continuously above 70 points, except for "interpretability" (66.7 points). It uses minimal clinical important difference (MCID) for score interpretation, with a MCID estimated to be of 6.5 points.[26] The SST scored among the top 3 in "reliability", "responsiveness to change", and "interpretability". In contrast, it scored low (52.4 points) in the "conceptual and measurement model" attribute, because insufficient information about its development process, involvement of the target population, and measurement level was found. For its interpretation an anchor-based strategy is proposed by linking its scores with different levels of disease severity.[12]

The OSS was among the top 3 in "conceptual and measurement model" and in "interpretability", and it also reached good results for "validity" and "responsiveness". Its "reliability" was below 60 points because some aspects of methods (such as data collection or time interval for the test-retest evaluation) could be either improved or better described. As these 3 instruments are similar in content, number of items, and administration time, the choice among them could be made upon the their dimensionality or answer options: ASES-p is bidimensional and permits obtaining separate scores for pain and function using Likert scales as response options; SST and OSS are unidimensional with dichotomous and Likert response options, respectively.

The FLEX-SF, SPADI, and SDQ-NL were drawn at the forth, fifth, and sixth place, respectively, in our overall score ranking with around 60 points. These three instruments presented acceptable results in all (except one) attribute-specific scores: FLEX-SF failed on "responsiveness", SPADI on "interpretability", and SDQ-NL on "reliability". Regarding the FLEX-SF, [6] its major particularity comes from its structure on 3 different testlets designed to minimize the respondent burden. Each testlet –easy, medium, and hard– consists of 15 items that can then be flexibly administered offering each patient only adequate questions, although the initial screening question could require a higher administrative burden. Additionally, a computer adaptive test version[7] has been developed and evaluated to facilitate data administration in large studies (even if it requires greater resources such as hard- and software). Nevertheless, it is necessary to mention the low expert ratings on the "responsiveness" attribute despite the fact that high standardized coefficients were shown. This was due to the fact that it was not clear which methods were used in the longitudinal design to obtain them.

The SPADI[34] is a commonly used instrument which clearly required further research for "interpretability". The SPADI's answer options initially consisted of visual analogue scales but were later transformed to numerical scales with the purpose of making it suitable for telephone administration, which was also judged to be reliable and valid.[43] The SDQ-NL requires further "reliability" testing. However, it could be a very good option for measuring change over time in longitudinal studies or clinical surveillance, not only because of its excellent "responsiveness", but also because of its low "respondent burden" (average time needed to complete <3 minutes and easy Yes/No answer options).[41]

Our study has some limitations. Firstly, the basis of the EMPRO evaluation is the information retrieved from a systematic literature review conducted only in the PubMed database. Although PubMed is the leading database in health sciences, we may have failed to identify all the eligible shoulder-specific PRO measures or all the published articles with their specific information on development process, metric properties, and administration issues. However, our sensitive search strategy, and also the additional hand search of identified articles, may have minimized this problem. Secondly, as the EMPRO assessment is based on the published evidence, it is affected by the quantity and the quality of this available information. A lack of evidence on a few items or attributes penalizes the EMPRO scores, because these were then rated with the worst score. Nevertheless, to avoid a strong penalization, the EMPRO attribute score was not obtained if more than half of the information was missing. Missing information on the interpretability attribute penalized the overall EMPRO score for most of the evaluated instruments, and pointed out the necessity of developing interpretability strategies as a facilitator for the extension of these measures beyond the research setting. Thirdly, the

EMPRO ratings may have been biased by the individual expertise of the evaluators, although the pair of reviewers that independently rated one outcome measure, followed by a consensus round, may have attenuated this concern. Finally, country-specific instrument versions were not evaluated separately in our study as our objective was to conduct a overall EMPRO evaluation of all the available information, and the evaluation of every country-specific version was not feasible.

To our knowledge this is the first study that provides a standardized and reliable expert-based evaluation of the available shoulder-specific PRO measures used in patients with different disorders. The basis of our assessment is the available published information that was retrieved in a systematic literature review. Each outcome measure was independently reviewed by two experts who reached final ratings by consensus. Our findings can be of interest in clinical practice as well as in research to help selecting the right shoulder-specific PRO measure for a certain purpose, facilitating decision making for individual patient care, or improving patient-doctor communication by understanding how the patient feels and acts in daily life.

Conclusion:

In conclusion, the evidence supports a preferential use of the ASES-p, SST, and OSS, which have been shown to be highly reliable, valid, and responsive instruments, with an acceptable conceptual and measurement model, interpretability, and low administrative burden. The use of the FLEX-SF, SPADI, and SDQ-NL can be recommended as they also presented acceptable properties in most of the attributes. Choosing among these instruments will mainly depend on particular study requirements. For use in longitudinal studies or clinical trials, where responsiveness to change and reproducibility are the maximum priority, SST would be recommended. In clinical practice, for patient surveillance SDQ-NL might be preferred to minimize respondent and administrative burden, but further information on its reliability is needed. To discriminate among patients or groups in one point evaluation, ASES-p or OSS could be the most reliable and valid option. Our results may facilitate the decision making process regarding the right instrument selection, its use, and interpretation for a certain study purpose or setting.

Acknowledgements

We thank all the experts for their participation in this standardized assessment of shoulder-specific PRO measures. Namely we thank: Michael Herdman, Angels Pont, Oriol Cunillera, Yolanda Pardo, Luis Rajmil, Mireya García-Duran, Juan Ignacio Arraras, Aida Ribera, Virginia Becerra, Sonia Rojas and Gabriel Medin. All of them have contributed to a better understanding of this topic by summarizing and evaluating the available information.

Reference List

1. Allander E. Prevalence, incidence, and remission rates of some common rheumatic diseases or syndromes. Scand.J.Rheumatol. 1974;3:145-53.

2. Angst F, Pap G, Mannion AF et al. Comprehensive assessment of clinical outcome and quality of life after total shoulder arthroplasty: usefulness and validity of subjective outcome measures. Arthritis Rheum. 2004;51:819-28.

3. 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.

4. Black N. Patient reported outcome measures could help transform healthcare. BMJ 2013;346:f167.

5. Celik D, Atalar AC, Demirhan M, Dirican A. Translation, cultural adaptation, validity and reliability of the Turkish ASES questionnaire. Knee.Surg.Sports Traumatol.Arthrosc. 2012.

6. Cook KF, Roddey TS, Gartsman GM, Olson SL. Development and psychometric evaluation of the Flexilevel Scale of Shoulder Function. Med.Care 2003;41:823-35.

7. Cook KF, Roddey TS, O'Malley KJ, Gartsman GM. Development of a Flexilevel Scale for use with computer-adaptive testing for assessing shoulder function. J.Shoulder.Elbow.Surg. 2005;14:90S-4S.

8. 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.

9. Curtis KA, Roach KE, Applegate EB et al. Reliability and validity of the Wheelchair User's Shoulder Pain Index (WUSPI). Paraplegia 1995;33:595-601.

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

11. Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ 2002;324:1417.

12. Godfrey J, Hamman R, Lowenstein S, Briggs K, Kocher M. Reliability, validity, and responsiveness of the simple shoulder test: psychometric properties by age and injury type. J.Shoulder.Elbow.Surg. 2007;16:260-7.

13. John M, Angst F, Awiszus F, King GJ, MacDermid JC, Simmen BR. The American Shoulder and Elbow Surgeons Elbow Questionnaire: cross-cultural adaptation into German and evaluation of its psychometric properties. J.Hand Ther. 2010;23:301-13.

14. Kirkley A, Griffin S, Dainty K. Scoring systems for the functional assessment of the shoulder. Arthroscopy 2003;19:1109-20.

15. Knaut LA, Moser AD, Melo SA, Richards RR. Translation and cultural adaptation to the portuguese language of the American Shoulder and Elbow Surgeons Standardized Shoulder assessment form (ASES) for evaluation of shoulder function. Rev.Bras.Reumatol. 2010;50:176-89.

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

17. Kuijpers T, van der Windt DA, van der Heijden GJ, Twisk JW, Vergouwe Y, Bouter LM. A prediction rule for shoulder pain related sick leave: a prospective cohort study. BMC.Musculoskelet.Disord. 2006;7:97.

18. 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.

19. Largacha M, Parsons IM, Campbell B, Titelman RM, Smith KL, Matsen F, III. Deficits in shoulder function and general health associated with sixteen common shoulder diagnoses: a study of 2674 patients. J.Shoulder.Elbow.Surg. 2006;15:30-9.

20. Leggin BG, Lannotti J. Shoulder outcome measurement. In: Lannotti J, Williams G, editors. Disorders of the Shoulder: Diagnosis and Management. Philadelphia, PA: Lippincott, Williams & Wilkins; 1999. p. 1024-40.

21. Lippitt S, Harryman D, Matsen F. A practical tool for evaluating function: The simple shoulder test. In: Matsen F, F F, Hawkins R, editors. The Shoulder: A balance of mobility and stability. Rosemont, IL: The American Academy of Orthopaedic Surgeons; 1993. p. 501-18.

22. Luime JJ, Koes BW, Hendriksen IJ et al. Prevalence and incidence of shoulder pain in the general population; a systematic review. Scand.J.Rheumatol. 2004;33:73-81.

23. McClure, P and Michener, L. Measures of adult shoulder function: The American Shoulder and Elbow Surgeons Standardized Shoulder Form Patient Self-Report Section (ASES), Disabilities of the Arm, Shoulder, and Hand (DASH), Shoulder

Disability Questionnaire, Shoulder Pain and Disability Index (SPADI), and Simple Shoulder Test. S49, 50-58. 2003.

24. Mehlum IS, Kjuus H, Veiersted KB, Wergeland E. Self-reported work-related health problems from the Oslo Health Study. Occup.Med.(Lond) 2006;56:371-9.

25. Michener LA, Leggin BG. A review of self-report scales for the assessment of functional limitation and disability of the shoulder. J.Hand Ther. 2001;14:68-76.

26. Michener LA, McClure PW, Sennett BJ. American Shoulder and Elbow Surgeons Standardized Shoulder Assessment Form, patient self-report section: reliability, validity, and responsiveness. J.Shoulder.Elbow.Surg. 2002;11:587-94.

27. Oh JH, Jo KH, Kim WS, Gong HS, Han SG, Kim YH. Comparative evaluation of the measurement properties of various shoulder outcome instruments. Am.J.Sports Med. 2009;37:1161-8.

28. Paananen M, Taimela S, Auvinen J, Tammelin T, Zitting P, Karppinen J. Impact of self-reported musculoskeletal pain on health-related quality of life among young adults. Pain Med. 2011;12:9-17.

29. Padua R, Padua L, Ceccarelli E, Bondi R, Alviti F, Castagna A. Italian version of ASES questionnaire for shoulder assessment: cross-cultural adaptation and validation. Musculoskelet.Surg. 2010;94 Suppl 1:S85-S90.

30. Patte, D. DIrectIons for the use of the index severity for painful and/or chromcally disabled shoulders. 36-41. 1987. Paris, The first open congress of the European Society of Surgery of the Shoulder and Elbow [SECEC].

440 31. Placzek JD, Lukens SC, Badalanmenti S et al. Shoulder outcome measures: a
441 comparison of 6 functional tests. Am.J.Sports Med. 2004;32:1270-7.

442 32. Pope DP, Croft PR, Pritchard CM, Silman AJ. Prevalence of shoulder pain in the
443 community: the influence of case definition. Ann.Rheum.Dis. 1997;56:308-12.

444 33. Richards RR, An KN, Bigliani LU et al. A standardized method for the assessment
445 of shoulder function. J.Shoulder.Elbow.Surg. 1994;3:347-52.

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

448 35. Roe Y, Soberg HL, Bautz-Holter E, Ostensjo S. A systematic review of measures
449 of shoulder pain and functioning using the International Classification of
450 Functioning, Disability and Health (ICF). BMC.Musculoskelet.Disord.
451 2013;14:73.

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

454 37. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health
455 status and quality-of-life instruments: attributes and review criteria. Qual Life Res
456 2002;11:193-205.

457 38. Testa MA, Simonson DC. Assesment of quality-of-life outcomes. N.Engl.J.Med.
458 1996;334:835-40.

459 39. Valderas JM, Ferrer M, Mendivil J et al. Development of EMPRO: A Tool for the
460 Standardized Assessment of Patient-Reported Outcome Measures. Value Health
461 2008;11:700-8.

40. van der Heijden GJ. Shoulder disorders: a state-of-the-art review. Baillieres Best.Pract.Res.Clin.Rheumatol. 1999;13:287-309.

41. 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.

42. Virta L, Joranger P, Brox JI, Eriksson R. Costs of shoulder pain and resource use in primary health care: a cost-of-illness study in Sweden. BMC.Musculoskelet.Disord. 2012;13:17.

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

44. Wright RW, Baumgarten KM. Shoulder outcomes measures. J.Am.Acad.Orthop.Surg. 2010;18:436-44.

45. Yahia A, Guermazi M, Khmekhem M, Ghroubi S, Ayedi K, Elleuch MH. Translation into Arabic and validation of the ASES index in assessment of shoulder disabilities. Ann.Phys.Rehabil.Med. 2011.

Table 1: Summarized characteristics of the identified shoulder disorder-specific instruments

Instrument	Articles for EMPRO	Author, publication year	Purpose of development	Shoulder disorder	Response options & comments	Time to complete, Period covered	nº items	Subscales (nº ítems)
1. **ASES-p**	**27**	Richards et al. (1994)[33]	A standardized form for the assessment of shoulder function	A variety of shoulder disorders	Visual analogue scale (pain item), 4-point Likert scales (activities of daily living). Score range 0-100 (worst to best).	<5' Not restricted to any period	11	Pain (1) Function (10)
2. **FLEX-SF**	**2**	Cook et al. (2003)[6]	To develop an adaptive scale that combines measurement precision with low response burden	A variety of shoulder disorders	Consists of 3 testlets: easy, medium, hard. Patient completes 1 of 3 testlets based on their response on an initial screening question. 6-point Likert scale. Score range 0-60 (worst to best).	-	15	-
3. **OSS**	**17**	Dawson et al. (1996)[10]	To assess the outcomes after shoulder operation	Patients with shoulder operations other than stabilization	5-point Likert scale. Score range 12-60 (best to worst) (new scoring system recommended: 0-48, worst to best)	<4', Last month	12	-
4. **PSS**	**5**	Leggin et al. (1999)[20]	To develop a region-specific shoulder outcome measure	A variety of shoulder disorders	0-3- or -10 point scale. Score range 0-100 (worst to best)	<10', (n.i.)	24	Pain (3) Function (20) Satisfaction (1)
5. **SDQ-NL**	**6**	Van der Heijden (2000)[41]	To evaluate functional disability limitation for clinical trials patients	Soft tissue shoulder disorders	Yes/No answer options. All items are pain-related. Score range 0-100 (best to worst).	3', Last 24h	16	-
6. **SDQ-UK**	**2**	Croft et al. (1994)[8]	To assess the restriction in everyday activities resulting from shoulder symptoms	Shoulder pain	Yes/No answer options. Score range 0-100 (best to worst)	(n.i.) Last 24h	22	-
7. **SPADI**	**26**	Roach et al. (1991)[34]	To measure pain and disability associated with shoulder pathology	Shoulder pain	Initially visual analogue scales. Later visual analog scales were transformed to numeric scales for telephone administration. Score range 0-100 (best to worst)	5-8' Last week	13	Pain (5) Function (8)
8. **SRQ**	**6**	L'Insalata et al. (1997)[18]	Designed to assess symptoms and function of the shoulder	A variety of shoulder disorders	5-option Likert scales, a visual analogue scale (global assessment). A non-graded question to select 2 areas in which the patient believes improvement is most important. Score range 17-100 (worst to best).	5-10' Last month	21	Global assessment (1) Pain (4) Activities of daily living (6) Work (5) Recreational & athletic activities (3) Satisfaction (1) Improvement (1)
9. **SSI**	**2**	Patte (1987)[30]	Disability outcome assessment for functioning and activities of daily living	A variety of shoulder disorders	Yes/No answer options.	7' (n.i.)	30	-
10. **SSRS**	**3**	Kohn & Geyer (1997)[16]	Disability outcome assessment for functioning and daily activities	A variety of shoulder disorders	0 to 5 or 35 point scale, Score range 0-100 (worst to best)	<3' (n.i.)	5	-
11. **SST**	**12**	Lippitt et al. (1993)[21]	A function-based outcome assessment tool	A variety of shoulder disorders	Yes/No answer options. Score range 0-12 (worst to best)	<3' (n.i.)	12	-

ASES-p: American Shoulder and Elbow Surgeons shoulder assessment – patient self-evaluation section; FLEX-SF: Flexilevel Scale of Shoulder Function; OSS: Oxford Shoulder Score; PSS: Penn Shoulder Score; SDQ-NL: Dutch Shoulder Disability Questionnaire (also known as van der Heijden shoulder disability questionnaire); SDQ-UK: United Kingdom Shoulder Disability Questionnaire (also known as Croft shoulder disability questionnaire); SPADI: Shoulder Pain and Disability Index; SRQ: Shoulder Rating Questionnaire (also known as L'Insalata Self-Administered Questionnaire - SAQ); SSI: Shoulder Severity Index ; SSRS: Subjective Shoulder Rating System; SST: Simple Shoulder Test (also known as Patte score). n.i.: no information.

Table 2: Expert ratings of each EMPRO item and attribute for every identified shoulder disorder-specific instrument

ATTRIBUTES	ASES-p	FLEX-SF	OSS	PSS	SDQ-NL	SDQ-UK	SPADI	SRQ	SSI	SSRS	SST
CONCEPT AND MEASUREMENT MODEL	81	66.7	66.7		66.7	47.6	52.4	52.4	14.3	28.6	52.4
1 concept of measurement	++++	++++	++++	++++	++++	++++	++++	++++	++++	++	++++
2 obtaining and combining items	++++	++++	++	-	++++	+++	++	++	-	++	++
3 dimensionality and scales	++++	+++	++	-	++	++	++	++	-	++	++
4 involvement of target population	-	++++	++++	-	++	+++	+	++++	-	-	++
5 scale variability	++++	-	++++	++	+++	++	+++	++	+	++	++++
6 level of measurement	+++	+++	++	-	++	+	++	++	+	++	++
7 procedures for deriving scores	++++	++	+++	+++	++++	++	++++	++	+	++	++
RELIABILITY - global score	75	66.7	58.3	55.6	41.7		83.3	50	66.7	41.7	75
internal consistency - reliability	*75*	*66.7*	*55.5*	*55.6*	*41.7*		*83.3*	*50*			58.3
8 data collection methods	++++	+++	+++	++++	+++	-	++++	++	-	-	+++
9 cronbach's alpha	++++	++++	++++	+++	++++	-	++++	+++	-	-	+++
10 IRT estimates	-	++	-	+	-	-	+++	-	-	-	+++
11 different populations	++++	n.a.	n.a.	n.a.	-	-	+++	++++	-	-	++
reproducibility - reliability	*75*	*58.3*	*58.3*	*50*			*66.6*	*50*	*66.7*	*41.7*	75
12 data collection methods	++++	++	+++	++	-	-	+++	++	+++	++	++++
13 test-retest and time interval	++++	++++	+++	+++	-	++++	++++	++++	++++	+++	++++
14 reproducibility coefficients	++++	++++	++++	++++	-	-	++++	+++	++++	+++	++++
15 IRT estimates	-	-	-	-	-	-	-	-	-	-	-
VALIDITY	86.7	83.3	75		93.3	50	66.7	25	50	40	80
16 content validity	+++	++++	+++	-	++++	++	++	++	+	++	++
17 construct/criterion validity	++++	+++	+++	+++	++++	+++	+++	++	++	++	++++
18 sample composition	++++	+++	+++	-	++++	+++	+++	+	+++	++	++++
19 prior hypothesis	+++	++++	++++	+++	+++	++	++++	++	++++	++	+++
20 rational for criterion validity	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
21 different populations	++++	n.a.	n.a.	n.a.	++++	n.a.	+++	n.a.	n.a.	+++	++++
RESPONSIVENESS TO CHANGE	77.8	33.3	77.8	44.4	100	88.9	77.8	77.8	44.4	66.7	100
22 adequacy of methods	++++	++	+++	++	++++	+++	+++	++	+++	++	++++
23 description of estimated magnitude of change	++++	++	++++	++++	++++	++++	++++	++++	+++	+++	++++
24 comparison of stable and unstable groups	++	++	+++	+	++++	++++	+++	++++	+	++++	++++
INTERPRETABILITY	66.7	55.6	66.7	33.3			22.2	11.1	0		55.6
25 rational of external criteria	+++	+++	+++	++	-	-	++	++	+	-	+++
26 description of interpretation strategies	+++	++	++	++	-	-	++	+	+	-	++

27	how data should be reported	+++	+++	++++	++	-	-	+	+	-	-	+++
OVERALL SCORE		77.4	61.1	68.9	26.7	60.3	37.3	60.5	43.3	35.1	35.4	72.6
BURDEN - score												
Burden I - respondent		55.6		88.9	11.1	100	77.8	22.2	22.2	11.1	66.7	88.9
28	skills and time needed	+++	-	+++	++	++++	++++	++	++	++	++++	++++
29	impact on respondents	++	+++	++++	+	++++	+++	++	++	+	++++	++++
30	not suitable circumstances	+++	-	++++	-	++++	+++	-	+	-	-	+++
Burden II - administrative		91.7	16.7	66.7	75	100	58.3	50	33.3	25	50	41.7
31	resources required	+++	++	++++	++++	++++	+++	++++	+	+	++++	+++
32	time required	++++	-	-	++++	++++	++++	-	++	++++	-	-
33	training and expertise needed	++++	-	++++	-	++++	+++	-	+	+	-	-
34	burden of score calculation	++++	++	+++	++++	++++	-	++++	++++	+	++++	++++
ALTERNATIVE FORMS OF ADMINISTRATION			66.7					83.3				
35	metric characteristics of alternative forms	n.a.	+++	n.a.	n.a.	n.a.	n.a.	++++	n.a.	n.a.	n.a.	n.a.
36	comparability of alternative forms	n.a.	+++	n.a.	n.a.	n.a.	n.a.	+++	n.a.	n.a.	n.a.	n.a.
Explanation: ++++ 4 (strongly agree); +++ 3; ++ 2; + 1 (strongly disagree); - no information; n.a. not applicable												