Contents lists available at ScienceDirect

Musculoskeletal Science and Practice

journal homepage: www.elsevier.com/locate/msksp

Technical and measurement report

Test-retest reliability of the ASES-p shoulder scale

Kalliopi Vrotsou[a,b,c,*], Ricardo Cuéllar[b,d], Félix Silió[e], Daniel Garay[f], Gorka Busto[g], Antonio Escobar[c,h]

[a] *Unidad de Investigación AP-OSIS Gipuzkoa, Osakidetza, San Sebastián, Spain*
[b] *Instituto de Investigación Sanitaria Biodonostia, San Sebastián, Spain*
[c] *Red de Investigación en Servicios de Salud en Enfermedades Crónicas (REDISSEC), Kronikgune, Barakaldo, Spain*
[d] *Servicio de Traumatología y Cirugía Ortopédica, Hospital Universitario Donostia, San Sebastián, Spain*
[e] *Servicio de Traumatología y Cirugía Ortopédica, Hospital Universitario Basurto, Bilbao, Spain*
[f] *Servicio de Traumatología y Cirugía Ortopédica, Hospital Galdakao-Usansolo, Galdakao, Spain*
[g] *Servicio de Traumatología y Cirugía Ortopédica, Hospital Mendaro, Mendaro, Spain*
[h] *Unidad de Investigación, Hospital Universitario Basurto, Bilbao, Spain*

ARTICLE INFO

Keywords:
ASES-p
Patient-reported-outcomes
Score
Shoulder
Test-retest reliability

Introduction: Shoulder disorders are common musculoskeletal problems. The self-assessed ASES questionnaire (ASES-p) is one of the most widely used tools for evaluating shoulder function. Its 11 items are divided in a *function* (10 items) and *pain* (1 item) dimension, assigned between 0 and 50 points each. Their sum is the scale's total score, with higher values indicating better health status. The current work explores the test-retest reliability of the Spanish version of the ASES-p score values.

Materials and methods: The scale was administered twice to a sample of subjects with various shoulder pathologies, via telephone interviews performed at 3–7 days apart. Exact agreement was calculated on an item and score basis. Score variability was assessed with the 95% limits of agreement method (LoA).

Results: N = 161 subjects were initially contacted, and a total of 82 stable health status subjects provided valid test-retest replies. "Do usual sport" was the only item with missing data. Exact agreement oscillated between 67 and 89% per item. The 95% LoA ranged between −5.9 and 6.9 points for function; −13.2 to 11.9 for pain and −10.3 to 10.1 for the total ASES-p score.

Conclusions: Test-retest reliability in stable patients was considered acceptable for the *function* and total scores, but not for *pain*. This may reflect usual pain behaviour, but it also implies that the pain evaluation should be further studied. The ASES-p pain subscore should not be used as the single measure for monitoring shoulder pain. Revisiting the "do usual sports" item may increase the scale's applicability.

1. Introduction

Shoulder pathologies are frequent musculoskeletal disorders, and their prevalence increases with age (Luime et al., 2004; Vicente-Herrero et al., 2009). They affect quality of life and are related to important economic and social costs (Virta et al., 2012; Paananen et al., 2011). Currently over 20 patient reported outcomes (PRO) scales for shoulder evaluation exist. The patient self-report section of the American Shoulder and Elbow Surgeons (ASES-p) is among the commonest used such scales (Roe et al., 2013; Schmidt et al., 2014). The ASES questionnaire was developed in 1994 in the United States (Richards et al., 1994; Michener et al., 2002). The complete instrument has several sections, but the scale's score is derived based exclusively on 11 self-report items.

Over the years the ASES-p has been adapted in several languages (Celik et al., 2013; Yahia et al., 2011; Knaut et al., 2010; Padua et al., 2010; Piitulainen et al., 2014). It was found to have the best overall evaluation in a standardized comparison of several shoulder PRO scales (Schmidt et al., 2014) and was recently validated into Spanish. The present report explores the test-retest reliability of the scale scores,

* Corresponding author. Unidad de Investigación de AP-OSIS Gipuzkoa, Instituto Biodonostia, Pº Dr. Beguiristain s/n, 20014, San Sebastian, Spain.
E-mail addresses: kalliopi.vrotsoukanari@osakidetza.eus (K. Vrotsou), ricardo.cuellargutierrez@osakidetza.eus (R. Cuéllar), felixmanuel.silioochandiano@osakidetza.eus (F. Silió), daniel.garayrodriguez@osakidetza.eus (D. Garay), gorka.bustoavis@osakidetza.eus (G. Busto), antonio.escobarmartinez@osakidetza.eus (A. Escobar).

https://doi.org/10.1016/j.msksp.2019.02.004
Received 26 November 2018; Received in revised form 11 February 2019; Accepted 20 February 2019
2468-7812/

completing previously presented results (Vrotsou et al., 2016).

2. Materials and methods

All subjects participating in the first phase of the study (Vrotsou et al., 2016) were contacted again via the phone by a trained interviewer. Briefly, these individuals had been previously recruited in our study for having a shoulder pathology. Recruitment took place in five public hospitals, located in the Basque Country (North of Spain). Patients were diagnosed by the collaborating orthopaedic specialists (Vrotsou et al., 2016). Those accepting to participate in the current study phase were administered the ASES-p scale twice, telephonically. The second interview (retest) was performed 3–7 days after the first (test). During the retest, participants were asked whether they had experienced any health changes since the test. Only those with an unchanged health status were considered in the respective analyses. Subjects reporting worse/better shoulder condition or other health issues were excluded. Approval was granted by the local ethics committee (Comité Ético de Investigación Clínica del Area Sanitaria de Gipuzkoa, 21/11/2012). All participants had signed an informed consent upon recruitment.

2.1. The ASES-p

The ASES-p scale is composed of 11 self-report items representing two dimensions, function (10 items) and pain (1 item). The function items are replied on a 4-point Likert scale ranging from “0 = unable to do” to “3 = not difficult”. Current pain level is measured on a 10 cm visual analog scale (VAS) on which the 0 and 10 ends indicate “absence of pain” and “maximum pain”, respectively. Each dimension's replies are then transformed to a 0–50 subscore. For the function dimension this is calculated as: ((5/3) x sum of 10 function item replies), and for the pain dimension as: (10 - VAS x 5). In either case, 0 points correspond to the worst and 50 to the best evaluation. The scale's total score is the sum of the two subscores, with a maximum value of 100 points representing a shoulder in perfect conditions.

2.2. Statistical analysis

Continuous variables are described with means and standard deviations (SD). Categorical variables are described with frequencies and percentages. Agreement was studied on item and total score basis. Item agreement was assessed in terms of identical test-retest replies. Score reliability was tested with the 95% limits of agreement method (LoA). All test-retest pairwise differences were derived; the mean difference ($\bar{d}$) and its standard deviation (SD_d) were estimated. The 95% LoA indicate that 95% of the differences are expected to lie in the interval $\bar{d} \pm 1.96 \text{ x } SD_d$. Score variability is considered acceptable when the obtained interval is not clinically important. The graphical presentation of these results is done with dispersion diagrams depicting the differences (y-axis) versus the mean values of the two administrations (x-axis) (Bland and Altman, 1986). Analyses were performed with the SAS software (version 9.3; SAS Institute, Cary, NC). The plots were drawn with SPSS (IBM Corp. Released, 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.). No imputations were performed in this study, and only valid replies were analysed.

3. Results

Between April and September 2016 all 161 previously described subjects (Vrotsou et al., 2016) were contacted. None of them had deceased and 27 refused participating at this stage. At the retest 40 individuals dropped-out or could not be located, while 12 reported a considerable health change at the second phone-call. A total of 82 valid test-retest replies were eventually obtained.

Participant mean age was 64.6 (SD: 11.8) years, 50% were females and the majority suffered by a subacromial pathology. Tendinopathy, instability and other shoulder conditions were less frequent (Table 1).

Table 1
Participant characteristics.

Variables	N = 82
Age in years; mean (SD)	64.6 (11.8)
Gender; n (%)	
Female	41 (50)
Male	41 (50)
Educational level; n (%)	
Primary or less	36 (44)
Secondary	36 (44)
University or higher	9 (12)
Diagnosis; n (%)	
Subacromial path. with/without RC rupture	62 (75)
Tendinopathy: tendinitis/tendinosis	12 (15)
Instability	3 (4)
Other	5 (6)
Received treatment	
Surgery	33 (40)
Infiltration	21 (26)
Rehabilitation	19 (23)
Medication only	9 (11)

SD: standard deviation; n(%): frequency (percentage); RC: rotator cuff.

Table 2
ASES-p range of item replies and test-retest agreement.

item	Description	Min (%)	Max (%)	Exact item agreement %
Pain				
	How bad is you pain today	0 (68)	10 (1)	73
Function				
1	Put on a coat	1 (7)	3 (73)	83
2	Sleep on your painful or affected side	0 (11)	3 (61)	80
3	Wash back/do up bra in back	0 (7)	3 (65)	78
4	Manage toileting	1 (10)	3 (78)	82
5	Comb hair	0 (4)	3 (74)	88
6	Reach a high self	0 (6)	3 (54)	77
7	Lift 10lbs above shoulder	0 (16)	3 (46)	67
8	Throw a ball overhand	0 (12)	3 (44)	68
9	Do usual work	0 (5)	3 (66)	89
10[a]	Do usual sport	0 (41)	3 (50)	87
Score				
ASES-p total (n = 46)		22.5 (2)	100 (28)	33
Pain (n = 82)		0 (1)	50 (68)	73
Function (n = 46)		5 (2)	50 (28)	39

Pain item replies range from 0 to 10 (min-max). All other items are replied on a 0: unable to do to 3: not difficult scale. Min and Max: item minimum and maximum obtained replies during the test interview. % indicates the percentage of the respective values in relation to obtained replies. Exact item agreement: the percentage of identical test-retest replies.

[a] Based on n = 46; rest of the items are based on n = 82.

Information on individual item test-retest replies and dimension agreement is presented in Table 2.

Most participants (68%) reported absence of shoulder pain. At the same time, over 50% of the sample had no difficulty in performing the majority of the ASES-p actions. Items 7 “lift 10lbs above the shoulder” and 8 “throw a ball overhand” presented more difficulty for the sample. Item 10 “do usual sports” was the only item with missing data, obtaining a total of 46 replies. As many as 41% of the responders were unable to do their usual sport.

The mean (SD) scale scores were as follows. ASES-p total: 84.0 (21.3) points; function subscore: 40.3 (11.2) and pain: 42.1 (13.8) points. The test-retest data resulted in a high percentage of identical replies (i.e. > 70%) for 9 of the 11 scale items. In this case items 7 and 8 stood out for presenting less agreement than the rest.

As far as the score differences between the two study moments were

Table 3
Test-retest differences and 95% limits of agreement of the ASES-p scores.

Score	Difference	SD	95% LoA
ASES-p total (n = 46)	−0.09	5.10	−10.3 to 10.1
Pain (n = 82)	−0.64	6.26	−13.2 to 11.9
Function (n = 46)	0.51	3.20	−5.9 to 6.9

Difference: pre minus post values. LoA: limits of agreement.

concerned, those were almost 0 for the ASES-p total and < 1 point in absolute values, for the two subscores (Table 3).

The 95% LoA were ± 10.2 points for the total score; ± 6.4 for the function and ± 12.5 for the pain subscores (Fig. 1). These variations represent 10%, 13% and 25% of the three score ranges respectively.

4. Discussion

This study explored the test-retest reliability of the ASES-p scale scores in a sample of subjects with pathological shoulders, complementing previously presented validation findings (Vrotsou et al., 2016). The current results appear to direct the attention toward two issues. The first is how the pain level is assessed and the weight that each dimension has on the scale's total score. The second is the “do usual sports” item and the applicability of the scale in patients who do not practice sports.

The obtained results suggested that the total scale score and function subscore are repeatable in consecutive administrations, but the same cannot be concluded for the pain subscore. The score variation observed in the pain dimension seems to contradict the elevated percentage of identical test-retest replies of the corresponding item. Nonetheless, the nature of this particular question could explain the findings. The pain item evaluates pain level today. Shoulder pain can vary from one day to the next, even for patients whose health status has not otherwise changed (Minns Lowe et al., 2014). While, the wider response options of this particular item (i.e. 0–10 VAS) may have also introduced further variation. As a result, the pain subscores, obtained under theoretically identical circumstances between two consecutive moments, could differ by as many as 12.5 points. But given the 0–50 dimension score range, the orthopaedic specialists do not consider 25 and 37.5 points, say, as representing similar pain levels. Thus, from a clinical perspective, the pain dimension LoA obtained here were too wide. The obtained results imply that the way the ASES-p pain is assessed, its assigned score or the weight each dimension has on the total scale should be further studied.

To the best of our knowledge, one previous ASES-p publication has studied test-retest reliability via 95%LoA. That estimation, reported only for the scale's total score (i.e. −8.8 to +10.2), was similar to ours (Yahia et al., 2011). Other studies have mostly explored the test-retest reliability via intraclass correlation coefficients (ICC) (Michener et al., 2002; Celik et al., 2013; Padua et al., 2010; Piitulainen et al., 2014; Goldhahn et al., 2008; Kocher et al., 2005). It should be mentioned that the ICC estimations may differ depending on whether “absolute agreement” or “consistency” is of interest (McGraw and Wong, 1996). But unfortunately, most authors do not usually specify which type they report (Celik et al., 2013; Padua et al., 2010; Piitulainen et al., 2014; Goldhahn et al., 2008; Kocher et al., 2005). In spite of that, it is worth mentioning that with one exception (Celik et al., 2013) the pain dimension has generally presented lower ICC values compared to the total and function subscores (Michener et al., 2002; Piitulainen et al., 2014; Goldhahn et al., 2008; Kocher et al., 2005). Nonetheless, information on the actual test-retest differences, as exposed in the present paper, would have been more informative and would have allowed for a better insight on the scale's score variability in other populations (Bland and Altman, 1986).

As far as the “do usual sports” item was concerned, over half of the participants did not reply to it, reporting not doing any sports. As a result, both the function subscore and the scale total score were based on a reduced sample. In our setting, people over 65 years of age are not usually involved in sport activities (Departamento de Salud del Gobierno Vasco, 2007). The received replies corroborate the lack of this habit. They also highlight a further issue; that the ASES-p scale in its current form may not be useful for patients who do not practice sports. This would result in missing data which, combined with the lack of an imputation algorithm, would lead to the impossibility of estimating a function subscore and subsequently a total scale score for many of the older patients. Creating an adaptive “do usual sports” item, by incorporating different physical activity levels, or developing an imputation algorithm that would allow for missing data, could be interesting fields for future research.

Other than that, individual item agreement was generally high, except for “lift 10lbs above the shoulder” and “throw a ball overhand”. Both items are related to actions imposing an elevated stress to the articulation of the shoulder (Reinold and Gill, 2010). Some participants commended avoiding these actions, but the interviewer was not instructed to specifically inquire on that. Based on their long clinical experience with shoulder pathologies, the orthopaedic specialists co-authoring the present manuscript confirm that this would most likely be the case, for the majority of their patients. While the doctors themselves specifically advice patients to avoid any action that may lead to further articulation injuries. We could thus hypothesize that the reduced agreement may be attributed to the fact that the subjects rated actions that avoid doing, at least with the pathologic shoulder. It seems reasonable to observe more variation in these items and less variation in more basic daily life activities, like put on a coat or comb the hair. “Lift

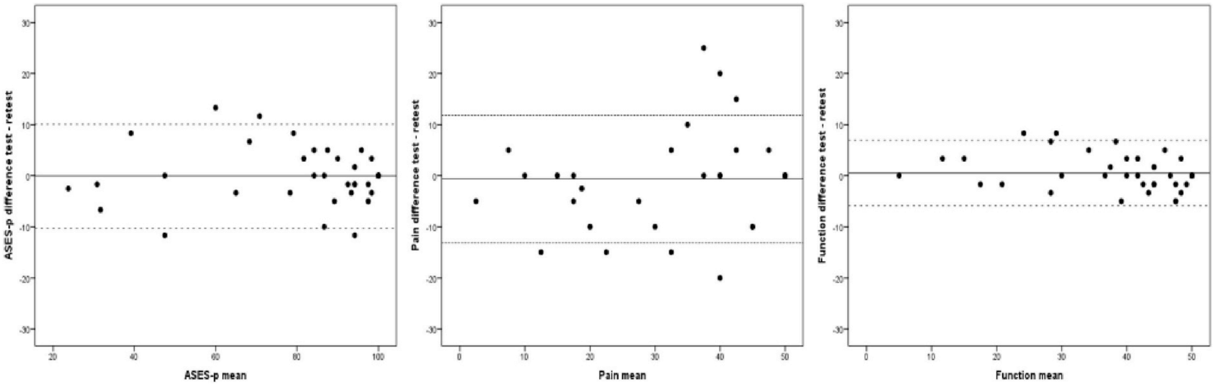

Fig. 1. 95% Limits of agreement for the total ASES-p, pain and function subscores, respectively.

10lbs above the shoulder" and "throw a ball overhand" were among the four items of least agreement in a previous validation study, performed on patients with more deteriorated shoulder functions (Goldhahn et al., 2008).

The present work has certain limitations. The data were collected via telephone interviews, meaning that the expertise of the interviewer may have affected the obtained replies. We should stress that the interviewer was experienced with such data collection and familiarized with the ASES-p. A standardized protocol was followed in all interviews, while during the retest the interviewer was blinded to the first test replies. Given the elevated number of drop-outs we believe that the phone interviews helped in avoiding further missing data. Only the ASES-p scale was administered in this phase. This was done for not overloading the participants and was sufficient for meeting the present study aim. Finally, participants reported an overall good shoulder state, which may limit the generalizability of the obtained results. Nonetheless test-retest reliability of a scale should be proved under all pathology states, with stable pathology data serving as the basis.

5. Conclusions

The clinicians using the ASES-p scale should be aware of its limitations. The pain subscore test-retest variations cannot be considered acceptable. Given its observed lack of stability, it would not be advisable to implement this dimension as the only measure neither for evaluating the actual shoulder pain level, nor for monitoring pain evolution over time. On the other hand revisiting the "do usual sports" item could increase the applicability of the scale to patients with different activity levels.

Conflicts of interest

None.

Author contributions

All authors participated in the conception and planning of the study. RC, FS, DG, GB recruited the study patients. KV, AE designed the first draft of the article. All authors contributed to this draft and accepted the final version of the article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.msksp.2019.02.004.

Funding source

This work was supported by the Health Department of the Basque Country Government (grant number: 2010111156). The funding source had no involvement in the study design, data collection, data interpretation and writing of this manuscript.

References

Bland, J.M., Altman, D.G., 1986 Feb 8. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1 (8476), 307–310.
Celik, D., Atalar, A.C., Demirhan, M., Dirican, A., 2013 Sep. Translation, cultural adaptation, validity and reliability of the Turkish ASES questionnaire. Knee Surg. Sports Traumatol. Arthrosc. 21 (9), 2184–2189.
Departamento de Salud del Gobierno Vasco, 2007. Encuesta de Salud del País Vasco. [cited 2016 Apr 22]. http://www.osakidetza_euskadi_eus/r85-ckpubl01/es/contenidos/informacion/encuesta_salud/es_escav/encuesta_salud_html.
Goldhahn, J., Angst, F., Drerup, S., Pap, G., Simmen, B.R., Mannion, A.F., 2008 Mar. Lessons learned during the cross-cultural adaptation of the American Shoulder and Elbow Surgeons shoulder form into German. J. Shoulder Elbow Surg. 17 (2), 248–254.
Knaut, L.A., Moser, A.D., Melo, S.A., Richards, R.R., 2010 Mar. 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. 50 (2), 176–189.
Kocher, M.S., Horan, M.P., Briggs, K.K., Richardson, T.R., O'Holleran, J., Hawkins, R.J., 2005 Sep. Reliability, validity, and responsiveness of the American Shoulder and Elbow Surgeons subjective shoulder scale in patients with shoulder instability, rotator cuff disease, and glenohumeral arthritis. J Bone Joint Surg Am 87 (9), 2006–2011.
Luime, J.J., Koes, B.W., Hendriksen, I.J., Burdorf, A., Verhagen, A.P., Miedema, H.S., et al., 2004. Prevalence and incidence of shoulder pain in the general population; a systematic review. Scand. J. Rheumatol. 33 (2), 73–81.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1 (30), 46.
Michener, L.A., McClure, P.W., Sennett, B.J., 2002 Nov. American shoulder and Elbow Surgeons standardized shoulder assessment form, patient self-report section: reliability, validity, and responsiveness. J. Shoulder Elbow Surg. 11 (6), 587–594.
Minns Lowe, C.J., Moser, J., Barker, K., 2014 Jul 9. Living with a symptomatic rotator cuff tear 'bad days, bad nights': a qualitative study. BMC Muscoskelet. Disord. 15, 228.
Paananen, M., Taimela, S., Auvinen, J., Tammelin, T., Zitting, P., Karppinen, J., 2011 Jan. Impact of self-reported musculoskeletal pain on health-related quality of life among young adults. Pain Med. 12 (1), 9–17.
Padua, R., Padua, L., Ceccarelli, E., Bondi, R., Alviti, F., Castagna, A., 2010 May. Italian version of ASES questionnaire for shoulder assessment: cross-cultural adaptation and validation. Musculoskelet Surg 94 (Suppl. 1), S85–S90.
Piitulainen, K., Paloneva, J., Ylinen, J., Kautiainen, H., Hakkinen, A., 2014. Reliability and validity of the Finnish version of the American shoulder and Elbow Surgeons standardized shoulder assessment form, patient self-report section. BMC Muscoskelet. Disord. 15, 272.
Reinold, M.M., Gill, T.J., 2010 Jan. Current concepts in the evaluation and treatment of the shoulder in overhead-throwing athletes, part 1: physical characteristics and clinical examination. Sport Health 2 (1), 39–50.
Richards, R.R., An, K.N., LU, Bigliani, Friedman, R.J., Gartsman, G.M., Gristina, A.G., et al., 1994 Nov. A standardized method for the assessment of shoulder function. J. Shoulder Elbow Surg. 3 (6), 347–352.
Roe, Y., Soberg, H.L., Bautz-Holter, E., Ostensjo, S., 2013. A systematic review of measures of shoulder pain and functioning using the International classification of functioning, disability and health (ICF). BMC Muscoskelet. Disord. 14, 73.
Schmidt, S., Ferrer, M., Gonzalez, M., Gonzalez, N., Valderas, J.M., Alonso, J., et al., 2014 Mar. Evaluation of shoulder-specific patient-reported outcome measures: a systematic and standardized comparison of available evidence. J. Shoulder Elbow Surg. 23 (3), 434–444.
Vicente-Herrero, M.T., Capdevilla-Garcia, L., Lopez Gonzalez, L., Ramirez-Iñiguez de la Torre, M.V., 2009. El hombro y sus patologías en medicina del trabajo. Semergen 35 (4), 197–202.
Virta, L., Joranger, P., Brox, J.I., Eriksson, R., 2012. Costs of shoulder pain and resource use in primary health care: a cost-of-illness study in Sweden. BMC Muscoskelet. Disord. 13, 17.
Vrotsou, K., Cuellar, R., Silio, F., Rodriguez, M.A., Garay, D., Busto, G., et al., 2016 Oct 18. Patient self-report section of the ASES questionnaire: a Spanish validation study using classical test theory and the Rasch model. Health Qual. Life Outcomes 14 (1), 147.
Yahia, A., Guermazi, M., Khmekhem, M., Ghroubi, S., Ayedi, K., Elleuch, M.H., 2011 Mar. Translation into Arabic and validation of the ASES index in assessment of shoulder disabilities. Ann Phys Rehabil Med 54 (2), 59–72.