VALIDITY AND RELIABILITY OF ARABIC VERSION OF PENN SHOULDER SCORE

Shaimaa Abd El-Rahman Haiba* , Prof.Dr.Neveen Abd El Latif ; Dr.Rania Nagy Karkousha*****

*B.Sc., in physical therapy
**Professor of Physical Therapy, Basic Science Department, Faculty of Physical Therapy , Cairo University
***Assistant Professor of Physical Therapy, Basic Science Department, Faculty of Physical Therapy, Cairo University

Abstract

Background: Shoulder disorders problem is a common, disabling condition which affects large population and causes many functional impairments. So it is important to measure the pain, satisfaction and function of the shoulder joint. **Purpose:** The purpose of this study was to translate and adapt the Penn Shoulder Score into Arabic and to investigate the face validity, content validity, internal consistency reliability, the feasibility and test retest reliability of Arabic version of Penn Shoulder Score as a score to assess the pain, satisfaction and functional activity of the shoulder joint in patients with different shoulder disorders, **Subjects and methods:** One expert panel; consists of ten experts and 250 patients with different shoulder disorders participated in this study, 510 sheets (including retest sheets) were filled out in this study. Forward translation, development of preliminary initially translated version, backward translation, development of the pre-final version and testing of the pre-final version by experts then testing of the final version on patients was done. Clarity index, expert proportion of clearance, index of content validity, expert proportion of relevance, descriptive statistics, missed items index, time taken to answer the score and Cronbach`s coefficient alpha were applied for statistical analysis. **Results:** The study showed that score index of clarity equals 89.58%, scale-level clarity index Universal Agreement equals 75% and the mean of proportion of clearance (clear responses) equals 93.75%. Also Scale-Level Index of Content Validity equals 87.5%, Scale-Level Index of Content Validity / Universal Agreement equals 58.33% and the mean of proportion of relevance (relevant responses) equals 100%. The score items were filled out by 95.3% in all sheets and it needed an average of 4.8-14.8 minutes to be answered in about 100% of all sheets. Cronbach's alpha equals 0.955. **Conclusion:** The translated Arabic-Language version of the Penn Shoulder Score has a face and content validity, feasibility, internal consistency and test-retest reliability enough for research and clinical application as a score to assess the pain, satisfaction and functional activity of the shoulder joint in patients with different shoulder disorders.

Keywords: Validity - Reliability - Feasibility - Penn Shoulder Score..

Introduction

Shoulder disorders are the third most prevalent musculoskeletal condition after spine and knee pain [(1)]. 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 [(2)].

A number of shoulder-specific questionnaires have been developed. However, the usefulness of these tools depends on their reliability, validity, and responsiveness, as established by the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) [(3)].

Penn Shoulder Score was developed in 1999 to asses subjects with shoulder dysfunction, consisting of a 100-point scale that includes three domains: pain, satisfaction, and function. The pain and satisfaction subscales have, respectively, three items and one item assessed using a 10 Numeric Rating Scale, where 0 corresponds to no pain and not satisfied, while 10 corresponds to the worst pain possible and very satisfied. The domain of function subscale contains twenty items, graded with a 4-point Likert scale, ranging from 0, which means "can not do at all" to 3 "without difficulty", with a maximum score of 60 points. The PSS score ranges from 0 to 100 points, with the maximum score indicating no pain, high satisfaction, and good function. The Penn Shoulder Score is a valid, reliable, and responsive self-report questionnaire used to assess patients with various shoulder disorders. It is a patient-reported outcome measure (PROMs) for the shoulder generally evaluates a patient's pain, current satisfaction level with their shoulder, and the joint's overall functionality [(4)].

Purpose of the Study:

This study was conducted to:

Translate and adapt the Penn Shoulder Score into Arabic and to investigate the face validity, content validity, internal consistency reliability, the feasibility and test retest reliability of Arabic version of penn shoulder score as a score to assess the pain, satisfaction and functional activity of the shoulder joint in patients with different shoulder disorders.

.

Subject, materials and methods

Study Design

This study is a prospective observational study; it follows the recommendations of **Beaton et al. (5)** for testing the face and content validity, feasibility, and internal consistency and test retest reliability of a translated instrument.

Participants

One expert panel; consists of ten experts and 250 patients with different shoulder disorders participated in this study.

nt ProceduresMeasureme

In this study, the penn shoulder score was used, the score consists of 3 domains: pain (3 items), satisfaction (1 item) and function (20 items). All items are self-assessed.

Test Procedures

The penn shoulder score was translated and adapted into the Arabic language following the process postulated by **Beaton et al. [(5)]**.

Process of translation and adaptation of instruments

Step 1: Initial translation The first step was to translate the questionnaire from the original language (English) into the target language (Arabic). Two bilingual translators, both native Arabic speakers (one in the medical field and one outside the medical field) provided 2 independent translations (T1andT2) of the original questionnaire.

Step 2: Merged translation The 2 initial translations (T1 and T2) were merged into a single one (T1-2). During this second step, all discrepancies between the 2 initial translations were discussed and resolved by the researcher and research committee of basic science for physical therapy, faculty of physical therapy.

Step 3: Back translation into the original language Without having read the original questionnaire, another bilingual translator, this time 2 native English speakers (one in the medical field and one outside the medical field), translated the merged version of the questionnaire, back into its original language (English) to obtain a new English version (BT1 and BT2). This step ensured that the translation faithful to the original questionnaire and had the same concepts.

Step 4: Expert committee An expert committee was setup. They met to produce, from the original version of the questionnaire, the first translations and the back translation, a pre-final version of the questionnaire with semantic, idiomatic, experiential, and conceptual equivalence regarding instructions, items, response format, wording sentence structure, meaning and relevance.

Step 5: Pre-final version testing Testing the pre-final Arabic version for face and content validity by two expert panels.

The expert panel (ten experts) was asked to evaluate each item of the tool for clarity (face validity) and provide suggestions to improve its clarity; dichotomous questions (clear\unclear) are used regarding instructions (1), items (24) and response words (5) with a total of 30 answers needed from each expert .

The scoring system was modified to be illustrated more clearly.

Also the expert panel (ten experts) was asked to evaluate each item of the pre-final Arabic version of the score for content equivalence (content-related validity) using the following scale: 1= not relevant; 2= unable to assess relevance; 3= relevant but needs minor alteration; 4= very relevant and succinct. And also give suggestions to improve its relevance. (1 and 2 considered not relevant, 3 and 4 considered relevant).

Pre-final version testing The pre-final version was tested on 30 patients with different shoulder disorders. After having filled in the questionnaire, they were questioned about their understanding of the different items and about the answers they provided, and they found no difficulties to answer the questionnaire and no misunderstanding or confusion about any item.

Step 6: After the pre-final version passes expert face and content validity test and the pilot study, it was named the final version.

Step 7: The final version was conducted on 250 patients with different shoulder disorders:

Patients filled out 250 data collection sheets which were used to collect demographic data (name, age, sex, weight,

height, body mass index and penn shoulder score).

Feasibility was evaluated by the assessment of the frequency of missing answers per item and administration time.

Two hundred and fifty patients with different shoulder disorders completed the data collection sheet again after one week.

Data Analysis

The demographic data of the patientsincluding age(years), weight (kg), height(m) and body mass index(kg/m2)were represented as the mean and standard deviation (SD) values .The data were explored for normality by checking data distribution. Calculation the mean, median and SD values were calculated**.** SPSS computer program (version 23) was used for data analysis.

Face validity was tested by clarity index and expert proportion of clearance. Content validity was tested by content validity index (CVI) and expert proportion of relevance. Descriptive statistics of data collected from patients including age (years), weight (kg), height(m) and body mass index and from sheets results were made using mean, median, SD, mode, minimum (min) and maximum (max). Feasibility index was calculated using missed item index and time taken to fill the questionnaire. Internal consistency reliability was measured using Cronbach's coefficient alpha. Test retest reliability was measured using Intra-class Correlation Coeffiecient (ICC).

Results

The results includes descriptive statistics of patient general characteristics as shown in table (1), descriptive statistics of sheet general characteristics as shown in table (2), clarity index, expert proportion of clearance of the final Arabic version to show the face validity, index of content validity and expert proportion of relevance to show the content validity, internal consistency reliability and test retest reliability.

Table (1): Descriptive statistics of patient general characteristics

Study group	Age (years)	Weight (kg)	Height (cm)	BMI (kg/m2)
Missing	0	0	0	0
Valid	250	250	250	250
Mean	36.51	76.49	167.16	27.35
±SD	11.16	15.06	8.65	4.71
Median	34	75	167	27
Minimum	19	47	150	18.7
Maximum	80	120	190	39.6

Table (2). Descriptive statistics of sheets general characteristics

	Who filled the score	Affected side	Test retest	Time
Missed	0	0	0	0
Valid	250	Rt=144 Lt =106	250 test 250 re test	Min=6 Max=34 Mean=14.7±4.8
Total	250	250	250	250

Clarity index of the final Arabic version

The scale index of clarity equaled 89.58% and scale-level clarity index universal agreement (UA) equaled 75% as shown in table (3).

Table (3): Item index of clarity of the final version

No	Item	Number of Rater's Agreements (clear response)	Item index of Clarity (IC)
	Pain		
1.	Question 1	10	100%
2.	Question 2	10	100%
3.	Question 3	10	100%
	Satisfaction		
4.	Question 1	10	100%
	Function		
5.	Question 1	10	100%
6.	Question 2	10	100%
7.	Question 3	10	100%
8.	Question 4	10	100%
9.	Question 5	10	100%
10.	Question 6	10	100%
11.	Question 7	10	100%
12.	Question 8	10	100%
13.	Question 9	10	100%
14.	Question 10	10	100%
15.	Question 11	10	100%
16.	Question 12	7	70%
17.	Question 13	7	70%
18.	Question 14	10	100%
19.	Question 15	7	70%
20.	Question 16	7	70%

21.	Question 17	10	100%
22.	Question 18	10	100%
23.	Question 19	8	80%
24.	Question 20	9	90%
Mean index of clarity for all items		89.58%	94%

The mean of proportion of clearance (clear responses) equaled 93% as shown in Table (4).

Table (4): Expert proportion of clearance of the final Arabic version

Expert number	Number of agreement (clear responses)	Proportion of clearance
1	20	83.33%
2	23	95.8%
3	23	95.8%
4	20	83.33%
5	24	100%
6	24	100%
7	20	83.33%
8	24	100%
9	24	100%
10	23	95.8%
Mean	22.5	93.75%

Index of content validity of the final Arabic version

The scale index of content validity (S-CVI) equaled 87.5% and scale index of content validity/universal agreement(S-CVI/UA) equaled 58.33% as shown in Table (5).

Table (5): Item index of content validity of the final Arabic version

No	Item	Number of raters that agree (relevant responses)	I-CVI
	Pain		
1.	Question 1	10	100%
2.	Question 2	10	100%
3.	Question 3	10	100%
	Satisfaction		
4.	Question 1	14	60%
	Function		
5.	Question 1	10	100%
6.	Question 2	10	100%

7.	Question 3	10	100%
8.	Question 4	10	100%
9.	Question 5	10	100%
10.	Question 6	10	100%
11.	Question 7	10	100%
12.	Question 8	10	100%
13.	Question 9	10	100%
14.	Question 10	11	90%
15.	Question 11	10	90%
16.	Question 12	15	50%
17.	Question 13	14	60%
18.	Question 14	10	100%
19.	Question 15	15	50%
20.	Question 16	15	50%
21.	Question 17	11	90%
22.	Question 18	10	100%
23.	Question 19	14	60%
24.	Question 20	11	90%
Mean index of relevance for all items		87.50%	87%

Expert proportion of relevance of the final Arabic version

The mean of proportion of relevance (relevant responses) equaled 100% as shown in Table (6).

Table (6): Expert proportion of relevance of the final Arabic version

Expert number	Number of agreement (relevant responses)	Proportion of relevance
1	24	100%
2	24	100%
3	24	100%
4	24	100%
5	24	100%
6	24	100%
7	24	100%
8	24	100%
9	24	100%
10	24	100%
Mean	24	100%

Feasibility measures

A feasibility measure is related to the easy application of the score. Retest sheets were not enrolled in data.

Missed item index

Invalid sheets were 0 and valid sheets were 250.The scale items were filled out by 95.3% in all sheets. Missed data index represent not answered data in relation to the tool data as shown in Table (7).

Time needed to measure the questions

Invalid sheets were 0 and valid sheets were 250.The score needed in average about 4.8-14.8 minutes to be answered as shown in Table (8).

Table 7. Missed index data

No	Item	Missed data (not answered)	Percentage of missed data
	Pain		
1.	Question 1	0	0 %
2.	Question 2	0	0%
3.	Question 3	0	0%
	Satisfaction	0	0%
4.	Question 1	0	0%
	Function	0	0%
5.	Question 1	0	0%
6.	Question 2	0	0%
7.	Question 3	0	0%
8.	Question 4	0	0%
9.	Question 5	0	0%
10.	Question 6	0	0%
11.	Question 7	0	0%
12.	Question 8	0	0%
13.	Question 9	0	0%
14.	Question 10	0	0%
15.	Question 11	0	0%
16.	Question 12	0	0%
17.	Question 13	0	0%
18.	Question 14	0	0%
19.	Question 15	0	0%
20.	Question 16	0	0%
21.	Question 17	70	28%
22.	Question 18	60	24%
23.	Question 19	122	48.8%
24.	Question 20	28	11.2%

Table (8): Descriptive statistics of time of 250 sheets

Study group (n=250)	Time in Minutes
Mean	14.78
Median	14
±S. D.	4.8
Minimum	6
Maximum	34

Internal Consistency

The internal consistency was measured by Cronbach's alpha. Results revealed that the internal consistency of observer scale of the Penn Shoulder Score was high level with Cronbach's alpha = 0.955. Farther more table (9) showed Cronbach's alpha if the item removed with no significant difference from the total scale Cronbach's alpha which confirm a very high level of internal consistency of the Penn Shoulder Score.

Table (9): Internal consistency of the Penn Shoulder Score by Cronbach's Alpha:

No	Item	Cronbach's Alpha if Item Deleted	Cronbach's Alpha of scale as total
	Pain		0.955
1.	Question 1	0.957	
2.	Question 2	0.953	
3.	Question 3	0.957	
	Satisfaction		
4.	Question 1	0.955	
	Function		
5.	Question 1	0.953	
6.	Question 2	0.953	
7.	Question 3	0.954	
8.	Question 4	0.953	
9.	Question 5	0.954	
10.	Question 6	0.953	
11.	Question 7	0.954	
12.	Question 8	0.954	
13.	Question 9	0.956	
14.	Question 10	0.952	
15.	Question 11	0.953	
16.	Question 12	0.953	
17.	Question 13	0.952	
18.	Question 14	0.953	
19.	Question 15	0.953	
20.	Question 16	0.952	
21.	Question 17	0.954	

22.	Question 18	0.953	
23.	Question 19	0.954	
24.	Question 20	0.953	

Test retest reliability

As shown in Tables (10 and 11) the total scores of questionnaire at the 1st and 2nd occasions by the same tester (intra-rater reliability). The total value of total score of questionnaire mean ±SD was (61.26±24.26) for the first reading of the main tester and (64.16±23.9) for the second reading for the same tester after one week. The intra-rater reliability using the Intra-class Correlation Coefficient (ICC) showed that there was a high reliability of total score of questionnaire (with ICC=0.97 and P-value=0.0001*).

Table (10): Comparison of scores of test with retest

	Total score of questionnaire	
	1st reading	2nd reading
Mean	61.26	64.16
±SD	±24.26	±23.9
ICC	0.97	
P-value	<0.001**	
Significance level	Significant	

Table (11): Intra-class Correlation Coefficient (ICC) for test and retest Intra rater reliability of total score of questionnaire:

	Test	Re test
Mean	51.85	53.9
±SD	21.06	20.92
Median	57	60
Minimum	0	4.22
Maximum	84.67	84.67

Discussion

The present study was designed to translate the English version of Penn Shoulder Score - to evaluate the shoulder joint`s pain, satisfaction and overall function in patients with different shoulder disorders - into Arabic version, adapt and test its face validity, content validity, internal consistency reliability, feasibility and test retest reliability. One expert panels (consists of ten experts) and 250 patients with different shoulder disorders participated in this study. 510 sheets (including retest sheets) were filled out in this study. This study was conducted in PT center in Tanta city.

The original score was forward translated into two Arabic versions then preliminary initial translated version was developed then it was backward translated into two English versions then pre-final version was developed then it was tested by experts for face and content validity, then a pilot study was made on 30 patients to insure its clearance, finally, it was tested by 250 patients for feasibility, internal consistency reliability and test retest reliability.

The Arabic version of PSS has high face validity as scale index of clarity equals 89.58%, scale-level clarity index UA equals 75% and the mean of proportion of clearance (clear responses) equals 93.75%. Also it has high content validity as S-CVI equals 87.5%, S-CVI/UA equals 58.33% and the mean of proportion of relevance (relevant responses) equals 100%. The results of the current study come in agreement with **Polit and Beck [(6)]** who stated that as scale to be judged as having an excellent content validity, it would be composed of items with item indexes of content validity (I-CVI) that meet the following criteria (I-CVI of 1.00 with 3-5 experts and a minimum I-CVI of 0.78 for 6-10 experts) and it would have S-CVI of 0.90 or higher. Also this came in agreement with Waltz et al. (2005) who stated that S-CVI/Ave of 0.90 or higher is the minimum acceptable indices are revised and re-evaluated. Also the study came in agreement with Sangoseni et al. (2013) who proposed a S-CVI of ≥0.78 as significant level for inclusion of an item into the study.

The Arabic version of the Penn Shoulder Score has excellent feasibility because the score items were filled out by 95.31% in the sheets but the 17th, 18th, 19th and 20th of the function items had a missing rate (28%, 24%, 48.8%, 11.2%) respectively. Also the scale needed an average of 4.8-14.78 minutes to be answered in about 100% of all sheets. The results of the current study determined according to **Van et al. [(7)]** who stated that missing rate on the item level was considered acceptable if no single item had a missing rate exceeding 10%, and completion time was considered acceptable if 95% of sheets were completed in less than 15 minutes.

The Arabic version of Penn Shoulder Score has excellent internal consistency and excellent test retest reliability as Cronbach`s alpha equals 0.995.

These results come similar to that of the original score that showed the PSS demonstrated excellent test-retest reliability as the pain subscale of the PSS demonstrated excellent reliability (ICC = 0.88), satisfaction subscale demonstrated excellent test-retest reliability (ICC = 0.93), while the function subsection demonstrated excellent test-retest reliability (ICC = 0.93). Internal consistency of the PSS (= .93) indicates that the items within the scale measure the same construct [(4)].

These results strengthened by the conducted study for cross cultural adaptation and validation of Penn Shoulder Score-Brazilian version which was conducted on 62 patients. Consistent with the original PSS, The PSS-Brazil displayed acceptable internal consistency, with a Cronbach

alpha of .92. Test-retest reliability was excellent, with an intraclass correlation coefficient of 0.95; the standard error of measurement and minimal detectable change were 12.8 and 14.4 points, respectively.

A high correlation was obtained between the PSS and the Shoulder Pain and Disability Index (0.96) and the Disabilities of the Arm, Shoulder and Hand questionnaire (0.86) **(8)**.

Also results come similar to that obtained by **Hazar et al. (9)** who conducted a study for translation, cultural adaptation, reliability, and validity of the Turkish version of the Penn Shoulder Score. It was conducted on 97 patients and the results showed that, it was determined that PSS-T is compatible with the Turkish language, and it is reliable and valid. In finding out about the validity of the scale, Constant Score, ASES, and WORC scales, all of which were proven to have reliable and valid Turkish versions, were used. Developers of the original version, found a Cronbach alpha value of 0.93 for internal consistency, and De Souza et al. stated it as 0.92 for Brazilian version. Cronbach alpha value of PSS-T was found as 0.81. PSS-T internal consistency value shows similarity with the original and Brazilian versions. This result indicates that PSS-T has high internal consistency. In PSS original version, test–retest ICC value was determined as 0.94 for the entire scale (subscales: pain: 0.88, satisfaction: 0.93, and function: 0.93). Brazilian version test–retest ICC value for the entire scale was 0.92 (subscales: pain: 0.85, satisfaction: 0.64, function: 0.94). In PSS-T, test–retest ICC value was determined as 0.90 for the entire scale (subscales: pain: 0.83, satisfaction: 0.78, and function: 0.90). These results show that PSS-T has high test–retest reliability. PSS-T was shown to have a very good correlation with Constant Score (0.65), ASES (0.78), and WORC (-0.77).

Conclusion

This study has provided evidence for the use of the Arabic version of the PSS as a region-specific shoulder measure for reporting outcome of patients with various shoulder disorders. Clinicians treating patients with similar shoulder diagnoses to those in this study can apply the measurement properties – face and content validity, feasibility and internal consistency and test retest reliability - of the subscales and total score as presented in this study. The individual subscales and the PSS total score can be considered a reliable and valid measure that can be used confidently to assess outcome of both individuals and groups of patients with shoulder disorders.

Reference

(1) Tekavec E, Joud A, Rittner R et al. Population based consultation patterns in patients with shoulder pain diagnoses, BMC Musculoskeletal Disord, 2012; 213-38.

(2) Higginson I and Carr A Measuring quality of life: using quality of life measures in the clinical setting, BMJ, 2001; 322: Pp: 1297-300.

(3) Terwee C, Mokkin L, Knol D, Osteo R, Bouter L and De vet H

Rating the methodological quality in systematic review of studies on measurements properties: a scoring system for the COSMIN checklist, Qual Life Res, 2012; 21: Pp: 651-57.

(4) Leggin B, Michener L, Shaffer M, Brenneman S, Iannotti J, Williams G The Penn shoulder score: reliability and validity, J Orthop Sports Phys Ther, 2006; 36: Pp: 138–51.

(5) Beaton D, Bombardier C, Guillemin F and Ferraz M Guidelines for the process of cross-cultural adaptation of self-report measures, Spine, vol.25, no.24, 2000; 4: Pp: 3186-91.

(6) Polit D and Beck C The content validity index: Are you sure you know what`s being reported? Critique and Recommendations, Research in Nursing and Health, 2006; 29(5): 489-97.

(7) Van V, Erwin B, Poeran J, Toriji H, Steeegers E and Bonsel G Feasibility and reliability of a newly developed antenatal risk score card in routine care, 2015; 31(1): 147-54.

(8) Marcela B, Jaqueline M, Gisele H and Anamaria S Measurement properties of the Brazilian version of the Penn Shoulder Score (PSS-Brazil): reliability, validity and responsiveness, Journal of Orthopedic and Sports Physical Therapy, vol. 45, no.2. 2015; Pp: 137-43.

(9) Hazar K, Gunaydin G, Osman P and Sozlu U Translation, cultural adaptation, reliability and validity of the Turkish version of the Penn Shoulder Score, Disabil Rehabil, 2018; 40(10): 1214-19.