Evaluation of intratester and intertester reliability of the Constant-Murley shoulder assessment

Marianne H. H. Rocourt, Msc, PT,[a] Lorenz Radlinger, PhD,[a] Fabian Kalberer, MD,[b] Shahab Sanavi, PT,[a] Nicole S. Schmid, PT,[a] Michael Leunig, MD,[b] and Ralph Hertel, MD,[b] *Berne, Switzerland*

This study evaluated the reliability of the Constant-Murley Score. Two testers assessed 63 patients twice on the same day using the original publication by Constant and Murley. The intratester reliability of the total score was high and the differences between the tests were small; 2 of 14 items were unreliable. The intertester reliability was high, but there were significant median differences of the total score; 12 of 14 items were unreliable. We believe that the differences between the testers were due to the brief explanations of test components in the original publication. The reliability of the Constant-Murley Score could possibly be improved by a better standardization of the assessment procedure. (J Shoulder Elbow Surg 2008;17:364-369.)

During the past decade, clinicians and researchers have increasingly used outcome assessments to determine the effectiveness of therapeutic procedures. For a comprehensive and comparable assessment of shoulder function, the European Society of Shoulder and Elbow Surgery (ESSE) promoted the Constant-Murley Score (CMS).[6] Although Constant and Murley described the method as reliable and reproducible, we hypothesized that the standardized description of the method in the original publication was not precise enough to warrant sufficient reproducibility. The authors described an intertester error of 3% (range, 0%-8%) in 100 abnormal shoulders examined by 3 different testers. The method of analysis was insufficiently described, however, and did not state whether the intertester reliability of the various items was also assessed.

From the [a]Head Office of Nursing, Medical Technical and Medical Therapeutic Administration, Physiotherapy Development and Research, Inselspital Berne; and the [b]Department of Orthopaedic Surgery, University Hospital Berne.

Reprint requests: Marianne Rocourt, Head Office of Nursing, Medical Technical and Medical Therapeutic Administration, Physiotherapy Development and Research, Inselspital Bern, PKT 2 A 534, CH-3010 Berne, Switzerland (E-mail: *marianne.rocourt@insel.ch*).

1058-2746/2008/$34.00
doi:10.1016/j.jse.2007.06.024

Conboy et al[5] analyzed the reliability of the CMS by using the rate of intertester error in 25 patients. The 3 testers differed significantly in their total score averages, and of the CMS subscales, the scores for "activity of daily living" demonstrated a significant difference. The intratester analysis of variance, based on 3 patients, was not statistically significant. Overall, the authors concluded that the score was not sufficiently reliable for use in the clinical follow-up of patients.

Only the strength component of the CMS has been methodologically standardized, and the use of a standard test position with different test instruments (spring balance vs Isobex dynamometer [Bern, Switzerland]) has been studied. It has been shown that a simple spring balance can give similar values to those from the more sophisticated Isobex.

Even though the reliability of the CMS is not precisely defined, the score is widely used and accepted throughout the European shoulder and elbow surgeon's community as the gold standard for the assessment of shoulder function,[4,5,8,9,11,12,14-16] and new assessments are often validated by comparing them with the CMS.[8-12,16,17]

The purpose of this study was to reassess the intertester and intratester reliability of the CMS, using the test criteria outlined in the original publication by Constant and Murley.

MATERIALS AND METHODS

Patients

The study included 63 consecutive patients (28 women, 35 men) scheduled for routine CMS assessment with various shoulder dysfunctions between May and September 2002. Their average age was 50.7 years (SD, 14.1; range, 21-76 years). In 39 patients the right shoulder was affected, and in 24 patients the left shoulder was affected. Eight patients were assessed before surgery, 29 patients at 3 months postoperatively, and 26 patients at 1 year after shoulder surgery. The ethics committee of Canton Berne authorized this study, and informed consent was obtained from all patients.

Before patient recruitment, it was important to know which testing protocol was applied in the various international shoulder units. Test protocol forms and descriptions were requested from different European hospitals. Seven protocols were received from units in England, France,

Germany, the Netherlands, and Switzerland and checked for consistency of their content, translation, and standardization, with the hope of obtaining a consistent description of the testing procedure. However, the differences among these CMS protocols varied widely; therefore, we decided to use the original description published by Constant and Murley[6] (Tables I and II).

Two independent testers, an orthopaedic surgeon (tester A) and a physiotherapist (tester B), were recruited from our institution. Both were given the original CMS protocol to read but received no further instructions on how to perform the assessment. Both testers assessed each patient independently twice on the same day (test 1, test 2). Patients were tested in groups, and assessments were performed in random order. Testers were blinded to patient diagnosis.

A fixed-spring balance calibrated in kilograms was used to measure the isometric maximal voluntary contraction (MVC power). An additional scale was used to convert the measurement result from parametric kilograms (Système International d'Unités) into nonparametric British pounds. Each tester then had to chose the equivalent points.

A video camera was used to observe the testing procedure with discretion, and a further observer recorded the performance of the tester step-by-step by means of a separate protocol.

Statistical analysis

Owing to the ordinal data level in the CMS and the presence of some outliers and extreme values, we used nonparametric procedures to calculate descriptive statistics (median, minimum, maximum, 25th and 75th percentiles), intertester and intratester correlations (Spearman rank correlation), and median differences (Wilcoxon signed rank test).

All calculations were done for the total score and separately for the 14 items within and between the two testers. Results of correlations were interpreted using coefficients of reliability. A limit of $r \geq 0.700$ for intertester and intratester reliability was accepted to be adequate for group measurements, according to Bös[2,3] and Lienert et al.[13] Statistical analysis was done with SPSS 11.5.1 software (SPSS Inc, Chicago, IL).

RESULTS

The intratester reliability of the total score correlated higher than the intertester reliability.

Intratester reliability

Total score. Tester A and B had a significant high intratester correlation of the total score (A: $\rho = 0.963$, $P < .001$; B: $\rho = 0.940$, $P < .001$; Table III). The median differences between the test 1 and test 2 scored randomly was only 3 to 4 points for each tester (A, $P = .347$; B, $P = .077$). The difference between the 2 assessments of tester B almost achieved significance (Figures 1 and 2; Table III). Tester A had the same total score in test 1 and test 2 in 8 of 63 cases (12.7%) and tester B in 12 of 63 cases (19.0%).

Table I Scoring for individual variables according to Constant and Murley (1987)[6]

Variable	Score
Pain	15
Activities of daily living	20
Range of motion	40
Power	25
Total	100

Single items. Also within the 14 items, the intratester reliability was higher than the intertester reliability. The intratester reliability was insufficient in 2 of 14 items. The item "full elevation from top of head" for tester B had a correlation of less than 0.7 ($\rho = 0.613$; $P < .001$). Scores for the test item of work showed a significant median difference ($P = .046$) between test 1 and 2 of tester B (Table III).

Intertester reliability

Total score. The intertester correlation coefficients for the total score were high ($\rho < 0.902$, $P < .001$). Nevertheless, there was a significant difference in the allocated total score between the testers. Tester A scored about 10 points higher than tester B ($P < .001$). About 50% of the differences were between 10 and 25 points (Figures 1 and 3, Table III). The testers A and B allocated the same score in 1 of 63 cases (1.6%).

Single items. Insufficient intertester reliability was found in 12 of the 14 items of the CMS (correlation $\rho < 0.700$, $P < .05$ or median difference of $P < .05$, or both). The only 2 items observed to have satisfactory intertester reliability were unaffected sleep and abduction (correlation $\rho \geq 0.700$, $P < .05$ or median difference of $P > .05$, or both; Table III).

To determine differences in methodology of the various international procedures, we compared protocols obtained from 7 other European shoulder units. A wide range of test methodologies was noted. The most important aspects are discussed.

The CMS awards pain 15 points. In some centers, pain was measured by asking if patients have none, mild, moderate, or severe pain, the answer to which corresponded with CMS categories 0, 5, 10, or 15. Others used a 15-point visual analogue scale to record the points. In some protocols, the patients were asked to describe their pain only during the last 24 hours; in others, the patients were asked to describe their general level of pain in daily life.

Also the activities of daily living were interpreted differently amongst institutions. Most protocols gave 0 to 4 points for the work item. In one CMS protocol, a new additional question was added.

Table II Protocol form designed for this study with all of the original items of Constant-Murley Score*

Item No.	Variable	CMS	Right Score	Left Score
1	Pain			
	None	15		
	Mild	10		
	Moderate	5		
	Severe	0		
	Activities of daily living			
2	Full work	4		
3	Full recreation/sport	4		
4	Unaffected sleep	2		
5	Positioning			
	Up to waist	2		
	Up to xiphoid	4		
	Up to neck	6		
	Up to top of head	8		
	Above head	10		
6	Forward elevation (°)			
	0-30	0		
	31-60	2		
	61-90	4		
	91-120	6		
	121-150	8		
	151-180	10		
7	Lateral elevation (°)			
	0-30	0		
	31-60	2		
	61-90	4		
	91-120	6		
	121-150	8		
	151-180	10		
	External rotation			
8	Hand behind head with elbow held forward	2		
9	Hand behind head with elbow held back	2		
10	Hand on top of head with elbow held forward	2		
11	Hand on top of head with elbow held back	2		
12	Full elevation from on top of head	2		
13	Internal rotation			
	Dorsum of hand to lateral thigh	0		
	Dorsum of hand to buttock	2		
	Dorsum of hand to lumbosacral junction	4		
	Dorsum of hand to waist (3^{rd} lumbar vertebra)	6		
	Dorsum of hand to 12^{th} dorsal vertebra	8		
	Dorsum of hand to interscapular region (DV7)	10		
14	Power (25 lbs = 25 points)			
	Total			

CMS, Constant-Murley Score.
In activities of daily living and in external rotation, we defined each single point as an item (8 items: Item No. 2, 3, 4, 8, 9, 10, 11, 12).
*We defined pain, positioning, forward elevation, lateral elevation, internal rotation, and strength as an item (6 items: Item No. 1, 5, 6, 7, 13, 14).

The protocols differed in asking about ability to work. One of them asked about the "normal" work situation, another about "occupational" and "leisure" work, and a further asked about "professional" work. One protocol measured the work ability as determined by the medical doctor and not by the patient. Furthermore, it did not define whether the work ability was related to the affected shoulder or to other pathology.

The CMS awarded 10 points for positioning, and the spectrum of questions ranged from asking in which position the patient could use his arm comfortably to how high the patient could lift his arm. In no protocol was the method of assessment of range of motion described with sufficient precision. Techniques to measure the internal rotation differed significantly, including using the lateral edge of the hand or using the thumb as a "pointer."

We also found different descriptions of testing positions to perform the strength test. For example, some centers seemed to tolerate an abduction of less than 90° to test the strength, whereas others considered that strength was 0 when the 90° position could not be reached. Another center measured the strength in kilograms instead of pounds and prorated 0.5 kg (= 500 g) to 1 point instead of 1 lb (= 453.56 g) to 1 point, which is a 10% error.

In our study, the descriptive analyses of a video and a separate testing protocol demonstrated some of these mentioned differences when the CMS was assessed.

- Similar to the international differences, our 2 testers chose different ways of performing the CMS. One tester asked very detailed questions about all pain circumstances (medications, pain during activity, pain during rest) using the ordinal scale 0 to 15, whereas the other tester simply asked if the patient had none, mild, moderate, or severe pain.
- The description for positioning item led to further misinterpretations. Tester A asked his patient to bring his hand on the top of his head, and tester B asked the patient to raise the straight arm in front of his body.
- Both testers visually estimated the angle instead of measuring it with a goniometer.
- The strength test was performed in a sitting position for tester A and in a standing position for tester B. There were significant strength differences in all tests between our testers, but not between test 1 and test 2 within each tester (Table III).

DISCUSSION

The CMS is a widely used instrument to assess the functional status of the shoulder. It was proposed by

Table III Intratester and intertester reliability of all 14 items and the total score of the Constant-Murley Score

Item	Reliability	Intratester*		Intertester†			
		A1/A2	B1/B2	A1/B1	A1/B2	A2/B1	A2/B2
Pain	ρ*	0.971	0.839	0.838	0.761	0.866	0.796
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.089	.153	.002	.099	.007	.283
Work	ρ	0.956	0.970	0.683	0.687	0.702	0.722
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	1.000	.046	.829	.516	.754	.496
Recreation, sport	ρ	0.904	0.867	0.711	0.742	0.751	0.805
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.384	.477	.207	.065	.122	.035
Sleep	ρ	0.854	0.946	0.888	0.894	0.781	0.805
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.564	.414	.405	.763	.837	.819
Positioning	ρ	0.917	0.961	0.562	0.553	0.587	0.582
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.317	.206	<.001	<.001	<.001	<.001
Forward elevation	ρ	0.916	0.900	0.841	0.802	0.827	0.786
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.527	.083	.022	.275	.060	.519
Lateral elevation	ρ	0.902	0.911	0.818	0.847	0.809	0.830
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.083	.317	.806	.467	.152	.513
External rotation 1	ρ	0.903	0.770	0.596	0.576	0.586	0.633
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.739	.705	<.001	<.001	<.001	<.001
External rotation 2	ρ	0.809	0.721	0.541	0.626	0.460	0.557
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.059	1.000	.004	.002	.001	<.001
External rotation 3	ρ	0.792	0.806	0.606	0.544	0.547	0.547
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.366	1.000	<.001	<.001	<.001	<.001
External rotation 4	ρ	0.848	0.753	0.576	0.696	0.559	0.627
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.063	.705	.015	.004	.003	.001
External rotation 5	ρ	0.701	0.613	0.587	0.779	0.533	0.656
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.086	.564	.010	.005	<.001	<.001
Internal rotation	ρ	0.856	0.901	0.875	0.854	0.875	0.891
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.186	.197	.004	<.001	.059	.004
Power	ρ	0.830	0.936	0.716	0.759	0.792	0.836
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	P (Wilcoxon)	.818	.262	<.001	<.001	<.001	<.001
Total	ρ	0.963	0.940	0.914	0.940	0.913	0.940
	P (ρ)	<.001	<.001	<.001	<.001	<.001	<.001
	diff (mean)	3.480	4.190	10.970	9.020	11.560	9.750
	diff (minimum)	0	0	0	0	0	0
	diff (maximum	15	53	55	32	58	29
	P (Wilcoxon)	.341	.077	<.001	<.001	<.001	<.001

ρ, Nonparametric correlation; *P* (ρ), significance of correlation; *P* (Wilcoxon): significance of nonparametric Wilcoxon statistics; *diff (mean)*, mean of absolute point differences between tests; *diff (min)*, minimum of absolute point differences between tests; *diff (max)*, maximum of absolute point differences between tests.
*A1/A2, tester A test 1/tester A test 2; B1/B2, tester B test 1/tester B test 2
†A1/B1, tester A test 1/tester B test 1; A1/B2, tester A test 1/tester B test 2; A2/B1, tester B test 2/tester B test 1; A2/B2, tester B test 2/tester B test 2.

the ESSSE as an outcome measure to compare shoulder function before and after treatment. To be effective, such an index must be based on well-designed reliability studies. The large discrepancies of the reported results between the various centers led to the hypothesis that allocation of the points could be

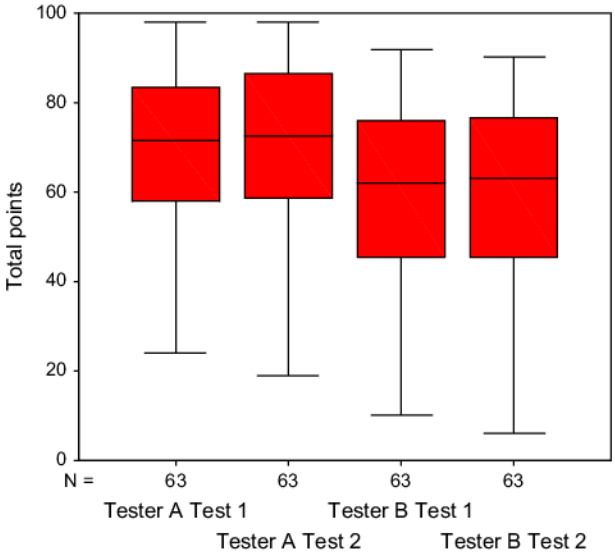

Figure 1 Box-and-whiskers plot of the total points of tester A (test 1 and 2) and tester B (test 1 and 2). The *horizontal line* in the middle of each box indicates the median (50th percentile), the *top* and *bottom borders* of the box mark the 75th and 25th percentiles, and the *whiskers* mark the 90th and 10th percentiles.

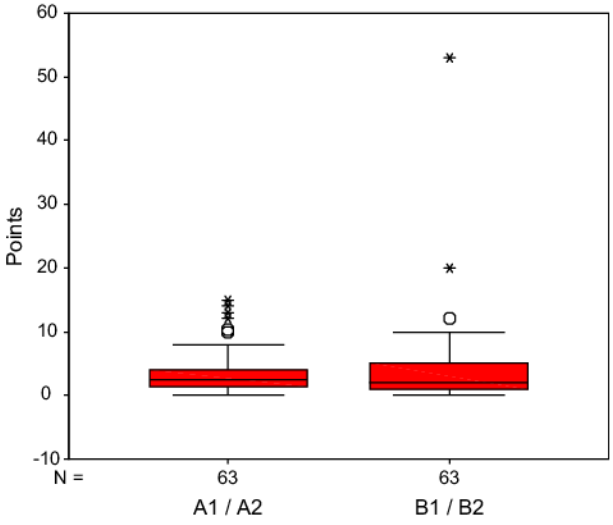

Figure 2 Intratester reliability is shown in a box-and-whiskers plot of the absolute differences in points of the total score within the tester. The *horizontal line* in the middle of each box indicates the median (50th percentile), the *top* and *bottom borders* mark the largest (75th percentile) and smallest (25th percentile) value and that is not outlier, and the *whiskers* mark the 90th and 10th percentiles; the *circle* indicates an outlier of more than 1.5 box lengths from the 75th percentile; the * indicates extremes more than 3 box lengths from the 75th percentile. *A1/A2,* Differences between test 1 and 2 of tester A; *B1/B2,* differences between test 1 and 2 of tester B.

inconsistent and that methodologic incertitudes could be the main reason for this.[7]

Our results showed high *intratester* reliability, and we found only 2 of 14 items in which reliability was unsatisfactory. The *intertester* reliability for the total score seemed to be acceptable. Nevertheless, we found significant median differences of the total score between the 2 testers. Those differences were about 10 points, and 50% of the values ranged between 10 and 25 points. Because the CMS uses a maximum of 100 points, a difference of 10 points or more is relevant. To determine whether the total score was solidly based on the various items, we analyzed the reliability of the items and found most of them to have unacceptable intertester reliability. This showed that testing the reliability of the total score is insufficient to truly assess its value.

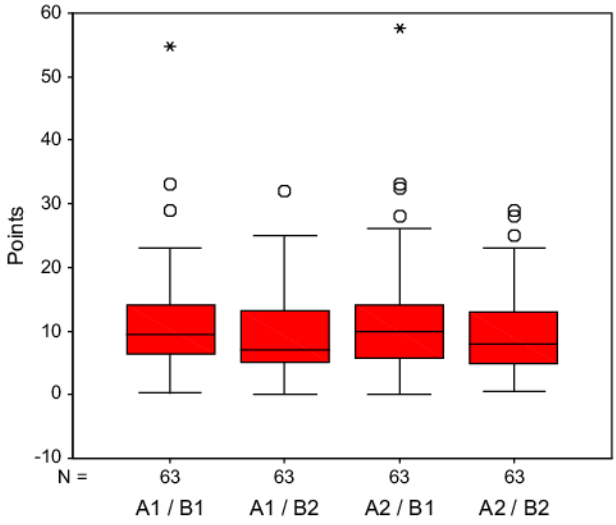

Figure 3 Intertester reliability is shown in a box-and-whiskers plot of the absolute differences in points of the total score between the testers. The *horizontal line* in the middle of each box indicates the median (50th percentile), the *top* and *bottom borders* mark the largest (75th percentile) and smallest (25th percentile) value and that is not outlier, and the *whiskers* mark the 90th and 10th percentiles; the *circle* indicates an outlier of more than 1.5 box lengths from the 75th percentile; the * indicates extremes more than 3 box lengths from the 75th percentile. *A1/B1,* Differences between tester A test 1/tester B test 1; *A1/B2,* differences between tester A test 1/tester B test 2; *A2/B1,* differences between tester B test 2/tester B test 1; *A2/B2,* differences between tester B test 2/tester B test 2.

The 2 testers were consistent within themselves, but each used a different value scale. Conboy et al [5] provided similar data. Constant and Murley (1987) published an average intertester error of the total score of 3 points (range, 0-8 points). This was lower than our intertester error of 10 points (range, 0-25 points). No further comparable data on the reliability of the CMS are available.

The analyses of the video recordings showed many differences in handling, application, and evaluation of the CMS items. This occurred even though the testers were using the same original protocol form.

Protocols used by various European shoulder units showed considerable differences in comparison with the original publication by Constant and Murley and

also among themselves. Major differences were found in the standardization of the test, such as questioning patients, quoting points, and instruction of test movements. Bankes et al[1] provided a comprehensive description for strength testing, but this methodology was obviously not transferred into clinical reality because we found different descriptions of testing positions and instructions in these protocols.

Constant and Murley[6] mentioned the individual items in tables but did not include a closer description of the motion, neither did they explain the reason for their selection of the items. No comment was given regarding activity or limitations of the shoulder in daily living, work, and sports. The protocol did not include an assessment of sensorimotor abilities or skills.

Constant and Murley used key words instead of asking precise questions (Table I, Table II). Some of the European protocols tried to standardize their tests by using specified questions. It would be interesting to know their level of reliability overall. We came to the conclusion that the original publication of the CMS is a too-short description of a complex test procedure. The user was not provided with sufficiently precise information about the exact test application and its interpretation. For that reason, users tend to make individual adaptations on the CMS. The consequences are major differences in allocating points to the items.

Our results, the observation of our testers, and the comparison of the various CMS protocols showed that the major problem of the CMS application was the imprecise and different standardization of the items of the test procedure; therefore, the CMS might be improved by a better standardization of the items. Because 12 of 14 items had a low intertester reliability, it must be suggested to redesign the score.[13] Most items had a sufficient intratester reliability; therefore, we believe that a new standardization could improve the intertester reliability of these items. At this time, the use of the score in its current version is only acceptable when the pretreatment and posttreatment score is determined by the same tester.

Alteration of the CMS protocol to a more standardized format would necessitate further reliability studies. Even translation of the CMS for the use in different European countries would require reassessment of its validity and reliability.

At this time caution is mandatory when comparing CMS results. When the total score is used to evaluate the outcome after an intervention or a treatment, differences between testers should be known, particularly when long-term studies assessed by different testers are presented. Comparing the outcome between different shoulder units using CMS is questionable and probably inadequate.

Although the reliability of a test procedure is essential, its validity is even more important. In the strict sense of the term, the validity is the most relevant criterion; the question is: does the CMS really measure the relevant variables in the correct way? We consider that in this sense, the validity of the CMS has yet to be assessed. We conclude that in the light of the arguments presented here, it appears to be problematic to validate other assessments by comparing them with the CMS.

REFERENCES

1. Bankes MJK, Crossman JE, Emery RJH. A standard method of shoulder strength measurement for the constant score with a spring balance. J.Shoulder Elbow Surg 1998;7(2):116-21.
2. Bös K. Handbuch sportmotorischer Tests. Göttingen: Hogrefe Verlag für Psychologie; 1987.
3. Bös K. Handbuch sportmotorischer Tests. Göttingen: Hogrefe Verlag für Psychologie; 2001.
4. Bosch U, Skutek M, Fremery RW, Tscherne H. Outcome after primary and secondary hemiarthroplasty in eldery parients with fractures of the proximal humerus. J.Shoulder Elbow Surg 1998; 7(5):479-84.
5. Conboy VB, Morris RW, Kiss J, Carr AJ. An evaluation of the Constant-Murley shoulder assessment. J Joint Bone Surg Br 1996;78:229-32.
6. Constant CR, Murley AHG. A clinical method of functional assessment of the shoulder. Clin Orthop Rel Res 1987;214:160-4.
7. Croft P, Pope D, Zanca M, O'Neill T, Silman A. Measurement of the shoulder related disability: results of a validation study. Ann Rheumat Dis 1994;53:525-8.
8. Dawson J, Fitzpatrick R, Carr A. The development and validation of a questionnaire. J Bone Joint Surg Br 1999;81:420-6.
9. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. J. Bone Joint Surg Br 1996;78: 593-600.
10. Fries JF, Spitz PW, Young YD. The dimensions of health outcomes: the Health Assessment Questionnaire, Disability and Pain scales. J Rheumatol 1982;9:789-93.
11. Kirkley A, Griffin S, Mc Lintock H, Linda NG. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability (WOSI). Am J Sports Med 1998;26:764-72.
12. Kohn D, Geyer M. The subjective shoulder rating system. Arch Orthop Trauma Surg 1997;116:324-8.
13. Lienert GA, Raatz U. Testaufbau und Testanalyse. Weinheim: Beltz Psychologie Verlagsunion; 1994.
14. 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.
15. Skutek M, Fremery 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.
16. Soldatis JJ, Moseley JB, Etminan M. Shoulder symptoms in healthy athletes: a comparison of outcome scoring systems. J. Shoulder Elbow Surg 1997;6(3):265-71.
17. Wülker N, Kohn D, Grimm C. Bewertung der Schulterfunktion mit unterschiedlichen Scores. Orthop Praxis 1991;12:750-4.