Original Article

Interchangeability of Diverse Analog Scales Used Within the Constant–Murley Score

Freek Hollman, M.D., Wanda M. de Raadt, M.Sc., Nienke Wolterbeek, Ph.D., Lodewijk W. van Rhijn, Ph.D., M.D., and Kiem G. Auw Yang, Ph.D., M.D.

Purpose: To assess the interchangeability of various existing answering scales within the subjective part of the Constant–Murley Score (CMS) and to determine the effect of the different answering scales on the inter- and intraobserver reliability. **Methods:** In this prospective, single-center, cross-sectional trial, patients with shoulder problems were included from June to September 2018. Subjects recruited were 18 years or older, presented various shoulder complaints, e.g., diagnosis of osteoarthritis, subacromial pain syndrome, rotator cuff or biceps tendon problems, or frozen shoulder. An extended version of the CMS was prepared including the same questions multiple times but with varying answer scales. Six versions were made with random order of the questions. The answering scales were a verbal and paper based visual analog scale (VAS), smiley face scale, Numeric Rating Scale (NRS), and categories. Internal consistency of the various CMS, Spearman correlation coefficients, intraobserver, and interobserver agreement was assessed (ICC). **Results:** In total, 93 patients were included. The total CMS using the paper-based VAS, smiley face score, and NRS were 46.9 ± 19.4, 45.2 ± 18.5, and 45.0 ± 18.7. Correlations of the total scores of the different versions varied from 0.98 to 0.99. CMS-category versus CMS-smiley face score and CMS-category versus CMS-NRS pain were significantly different ($P = .02$ and $P = .01$). Good internal consistency (0.76-0.79) and acceptable inter- and intraobserver reliability were found (ICC: 0.89-0.97, 0.98-0.99; $P < .001$). **Conclusions:** The different answering scales for the subjective subscales within the CMS for pain, work, and recreational activity were not interchangeable on item level and significantly influenced the total CMS score. Differences were below the smallest detectable change and interpreted as not clinically relevant. Particularly on item level, data from different studies cannot be pooled and compared when different answering scales are being used. The inter- and intraobserver reliability were excellent. **Level of Evidence:** Level I, prospective cross-sectional study.

The Constant–Murley Score (CMS) is a shoulder-specific questionnaire that is used to evaluate shoulder pathology and treatment outcome.1 The CMS is a combined patient-report and objective (health professional administered) shoulder outcome measure. In 2008, the CMS was officially updated by the original author.[2] A validated adjustment was made by replacing

the categories for pain, work, and recreational activities by a visual analog scale (VAS). In addition, a score modification, adjusting for age and sex, was proposed. Currently, different versions of the CMS are used, namely the original version, the updated version, and various mixtures of both. In particular, the subjective part is subject to variation. For these items, different measurement scales are used, e.g., different Numeric Rating Scales (NRS), VAS, or the original ordinal structured rating system of the CMS. The reliability and validity of the CMS has been studied extensively, with acceptable inter- and intraobserver reliability values reported. However, because of the variations in application and scoring, the CMS may not be interchangeable across studies.[3-5] Therefore, making accurate comparisons between CMS studies is difficult.

The purposes of this study were to assess the interchangeability of various existing answering scales within the subjective part of the CMS and to determine the effect of the different answering scales on the inter- and intraobserver reliability. It was hypothesized that the different answering scales of pain, work and

recreational activity of the CMS would be interchangeable and would not influence the total score. We further hypothesized that the inter- and intraobserver measurement of the different answering scales would be reliable.

Methods

This study was a prospective, single-center, cross-sectional study. Ethical approval (W18.042) was obtained, and subjects were fully informed about the study. All participants signed informed consent. Eligible patients with shoulder problems visiting the orthopaedic outpatient clinic of the St. Antonius Hospital during the period of June to September of 2018 were included. Subjects recruited were 18 years or older, presented various shoulder complaints, e.g., diagnosis of osteoarthritis, subacromial pain syndrome, rotator cuff or biceps tendon problems, or frozen shoulder. Patients with shoulder instability (luxation) or patients who recently (<6 months) underwent surgery on the ipsilateral side (e.g., rotator cuff repair, shoulder prosthesis, osteosynthesis) were excluded. Subjects with cognitive impairment or insufficient comprehension of the Dutch language also were excluded.

The CMS is a 100-point scoring system that is divided into 4 subscales: pain (15 points), general daily activities (20 points), range of motion (ROM; 40 points), and strength (25 points). In the original version of CMS, the pain score was evaluated in 4 categories: none (15), mild (10), moderate (5), and severe (0). This was replaced in the modified version by a VAS with a sliding cursor.[2] Since the VAS with the sliding cursor that was used by Constant et al. is not widely available, it was replaced by a VAS on paper and by a smiley face score with a sliding cursor.[2] For the paper VAS, patients were asked to put a mark on a 15-cm line to indicate their pain score.[6] In addition, the NRS pain score (verbal score of 0-10 points, whole numbers) was used. Scores for activities of daily living (work and sports) were recorded using categories, the paper VAS and smiley face scores. To prevent bias due to the question sequence, 6 versions of the CMS were prepared with random ordering of the answering scales for the measures of pain, work, and recreational activity. Therefore, each version contained 4 different answering scales for pain, 3 answering scales for daily work, and 3 answering scales for recreational activities. The 6 versions were randomly allocated to the patients. A working protocol for assessing the CMS was developed and discussed by the observers before the inclusion of the first patient.

During the visit to the clinic, the same compiled CMS was completed twice for both shoulders. For intraobserver measurement, it was completed first by the researcher before the consultation with the orthopaedic surgeon and second after the consultation. The

Table 1. Descriptive Statistics for Demographic and Clinical Data (n = 92)

	Value (N)
Mean age, y, $\pm$ SD	58 $\pm$ 10.4
Sex, male/female	37 (40%)/55 (60%)
Dominant side, right/left	81 11
Diagnosis*	
Arthritis	8
SAPS	16
Cuff problems	41
Biceps problems	12
Frozen shoulder	6
Other	32
Most affected shoulder, right/left	58 (63%)/34 (37%)
Dominant shoulder affected	55 (60%)
Both shoulders affected	52 (57%)
Months with shoulder complaints	
$\leq$6 mo	31 (34%)
7-11 mo	14 (15%)
1-2 y	22 (24%)
>2 y	25 (27%)

SAPS, subacromial pain syndrome.
*There were 15 patients with multiple shoulder problems.

interobserver reliability was measurement by completion of the CMS during the consultation by the orthopaedic surgeon involved in this study and by the researcher before or after the consultation. To reduce bias, the evaluations of the observers were performed independently in separate rooms.

Statistical Analysis

Results were analyzed using SPSS statistical software (IBM SPSS Statistics for Windows, Version 24.0; IBM Corp., Armonk, NY). Studies determining measurement properties require at least 50 patients.[7] Internal consistency (Cronbach's alpha coefficient) of the CMS was calculated using the various answering scales of pain, work, and recreational activity.[8] Spearman correlation coefficients were calculated to examine the convergent validity between the total scores. In addition, paired t tests were used to determine whether the mean differences between the various pain, work, and recreational scores could be considered insignificant. To assess systematic errors and agreement, Bland–Altman plots were compiled using the mean difference and the limits of agreement (mean difference $\pm$ 1.96 * standard deviation of the difference). Intraclass and interclass correlation coefficients (ICCs) were calculated for intra- and interobserver reliability (2-way mixed and random effects model, single measurements, and consistency). Assessment of both shoulders (affected and non-affected) were used for intra- and inter-reliability only.

Results

A total of 93 patients were assessed for eligibility. One patient was excluded due to not fully understanding the
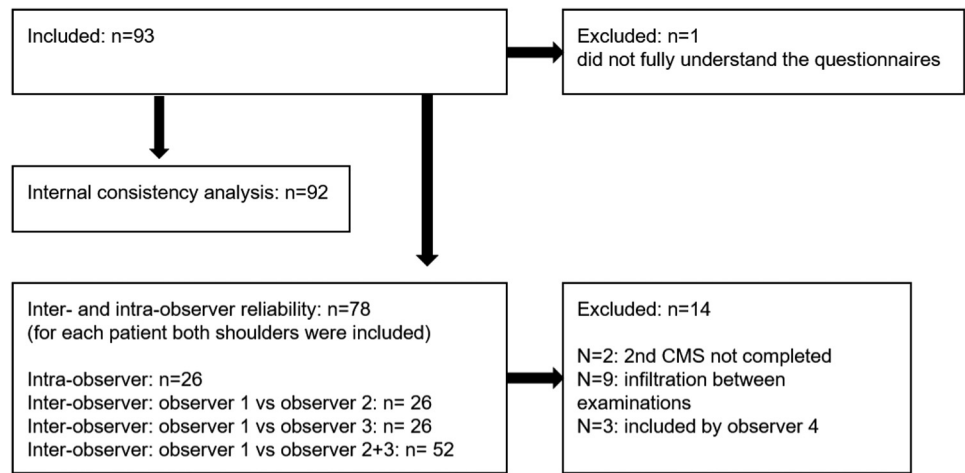
Fig 1. Schematic diagram for the populations that were included in this study. (CMS, Constant–Murley Score.)

questionnaires. Of the 92 patients, 37 (40%) were men and the mean age was 58 years (range 36-80 years) (Table 1). Two patients did not complete the second CMS, and 9 patients were infiltrated between the 2 examinations. Since this intervention could affect the outcome of the CMS, these patients were excluded from the intra- and interobserver analysis. Of these 81 patients, there were 3 patients assessed by observer 4. Because of the low number of patients, it was decided to exclude observer 4 from the intra- and interobserver analyses, which did not affect the outcome (Fig 1).

CMS Total

The mean (±SD) score of the first assessment for the affected shoulder of the original CMS with pain, work, and recreational scores given in categories was 44.2 ± 18.8. The total scores of the CMS using the paper VAS, the smiley face scores, and NRS pain were 46.9 ± 19.4, 45.2 ± 18.5, and 45.0 ±18.7, respectively (Table 2). Mean differences for the paper VAS versus the other scores were significantly different (Table 3). In addition, CMS-category versus CMS-smiley face score and CMS-category versus CMS-NRS pain were significantly different ($P = .02$; $P = .01$). No floor or ceiling effects in total scores were found.

To determine whether it is possible to convert the continuous outcome scale back to the original category scale, the outcome scores of the categories were compared in box plots with the outcome scores of the continuous data of the paper VAS, smiley face, and NRS pain. Table 4 shows that the mean values measured on a continuous scale are increased with increased category values. However, the minimum and maximum score show that converting to categorical values cannot be reliable performed, as they overlap multiple categories. Only with regard to the NRS pain scores and the smiley face scores the lowest category (0) and highest (15) category are discriminative (Table 4).

Internal Consistency

Cronbach's alpha outcomes showed a good consistency for the various answering scales (range 0.76-0.79) (Table 2). The correlations of total CMS scores using the answering scales varied from 0.98 to 0.99 and were all significant, $P < .001$ (Table 5). The Bland–Altman plots indicated high levels of agreement and no systematic errors between answering scales. The Bland–Altman plot of the original CMS and the CMS using the paper VAS is presented in Fig 2.

Intra- and Interobserver Reliability

For intraobserver reliability, there were 26 patients assessed twice by the researcher (observer 1). In this part of the analysis, both shoulders were included, resulting in 52 shoulders. No significant mean difference was found between the first and second examination and high ICC values between 0.98 and 0.99 ($P < .001$). For interobserver reliability, observers 2 and 3 included 26 patients each. All these patients also were assessed by observer 1, resulting in 52 patients with 104 shoulders for interobserver reliability (Fig 1). The interobserver correlation coefficient for the total scores

Table 2. Descriptive Statistics of the Total CMS Outcomes for the Affected Side (n = 92)

Measurement	Mean	SD	Minimum	Maximum	Cronbach Alpha
CMS-category	44.2	18.8	10.0	88.0	0.78
CMS-paper VAS	46.9	19.4	10.0	95.4	0.79
CMS-smiley face score	45.2	18.5	10.0	89.8	0.76
CMS-NRS pain*	45.0	18.7	10.0	88.5	0.78

CMS, Constant–Murley Score; NRS, Numeric Rating Scale; VAS, visual analog scale.

*In this version, the answering scale for pain was a NRS scale, whereas for daily work and recreational activities, the category scale was used.

Table 3. Paired *t*-Tests of Mean Differences (±SD) of Total CMS Scores of Various Answering Scales (n = 92)

	CMS-Category	CMS-Paper VAS	CMS-Smiley Face Score
CMS-paper VAS	2.7 (3.5); *P* < .001	—	—
CMS-smiley face score	1.0 (4.0); *P* = .02	1.7 (4.0); *P* < .001	—
CMS-NRS pain*	0.8 (2.9); *P* = .01	1.9 (3.3); *P* < .001	0.2 (3.8); *P* = .6

CMS, Constant−Murley Score; NRS, Numeric Rating Scale; VAS, visual analog scale.

*In this version, the answering scale for pain was a NRS scale, whereas for daily work and recreational activities the category scale was used.

were excellent (range 0.89-0.97) for the various CMS answering scales (*P* < .001).

Discussion

The results of this study indicated that different answering scales for the CMS subscales pain, work, and recreational activity were not interchangeable. Changing the answering scale had a minor, although significant, effect on the CMS total score. Nevertheless, the inter- and intraobserver reliability and correlations using the different answering scales were excellent. In addition, the Cronbach's alpha correlation coefficients of the CMS using the various answering scores were between 0.76 and 0.79. These small differences indicate that changing the answering scale did not influence the internal consistency of the CMS.

The original CMS measured pain in categories, which was changed to a VAS score with a sliding cursor for pain and subdomains of activities of daily living in the updated version.[1,2] The VAS score with the sliding cursor was replaced by a 15-cm paper VAS by several authors.[6,9] Although this new method was introduced as an improvement, this new outcome scale was not validated.[10] In this study, changing the answering scale affected the CMS total score 0.8 to

2.7 points. This was found to be significant; however, it is a small difference on a 0-100 scale. Furthermore, previously the minimal detectable change of the CMS has been determined to be 10 to 12 points; therefore, this significant effect of less than 3 points can be considered as not clinically relevant.[3] Despite the clinically irrelevant effect on the CMS total score, on item level the different answering scales cannot be used interchangeably. As presented in Table 4, the pain scores measured with the paper VAS, the smiley face score, and NRS scale could not be converted to the consecutive categories. The lack of interchangeably between answering scales has been studied previously in patients with chronic pain disorders.[11] In that study, pain measured with a paper-based continuous horizontal VAS (0-100) was compared with a 5-category verbal rating score. As in this study, the scores could not be converted to each other and could not be used interchangeable. In addition, they found that pain was scored greater when using a VAS scale compared with the categorized verbal rating score. This also in accordance with our findings. Yet, a VAS rating scale might be more appropriate in detecting changes over time compared with categorical scales.[12,13] This might be caused by a certain threshold for choosing a category compared with putting a line on a nonhatched VAS line. Results showed that with increasing pain, the difference between these 2 scores decreased. This might indicate that the VAS score might be more accurate, especially for detecting lower levels of pain.

Table 4. Comparison of Pain Scores Completed in Categories Versus Other Pain Scores of the Affected Shoulder (n = 92)

Categories (Points)	N		Mean	SD	Minimum	Maximum
No pain (0 points)	25	Paper VAS	3.4	2.8	0.0	12.3
		Smiley face	2.3	2.8	0.0	9.9
		NRS pain	2.9	1.7	0.0	6.0
Mild (5 points)	46	Paper VAS	7.0	2.5	1.0	11.8
		Smiley face	6.1	2.7	0.0	11.4
		NRS pain	5.6	2.4	1.5	12.0
Moderate (10 points)	17	Paper VAS	11.1	1.8	8.9	14.8
		Smiley face	8.7	2.9	3.6	11.4
		NRS pain	8.5	3.6	1.5	15.0
Severe (15 points)	4	Paper VAS	14.8	0.4	14.2	15.0
		Smiley face	15.0	0.0	15.0	15.0
		NRS pain	14.6	0.8	13.5	15.0

CMS, Constant−Murley Score; NRS, Numeric Rating Scale; VAS, visual analog scale.

Table 5. Spearman Correlation Coefficients (r) of Total CMS Scores Using the Various Answering Scales (n = 92)

	CMS-Category	CMS-Paper VAS	CMS-Smiley Face Score
CMS-paper VAS	0.98	—	—
CMS-smiley face score	0.98	0.98	—
CMS-NRS pain	0.99	0.99	0.98

NOTE. In this version, the answering scale for pain was a NRS scale, whereas for daily work and recreational activities the category scale was used.

CMS, Constant−Murley Score; NRS, Numeric Rating Scale; VAS, visual analog scale.

Intraobserver and Interobserver Reliability

Over the years, the CMS has been criticized due to poorly defined methods in the original version resulting in different methods of conduct and interpretations of the score by health professionals.[5,14,15] Results showed that the intra- and interobserver reliability for the various answering scales were excellent. This is in accordance with results from several other studies.[3,5,16] For the "paper VAS" version, an interobserver ICC of 0.94 was reported by Moeller et al.[3] Johansson et al.[17] found intraobserver ICC ranges from 0.90 to 0.98 and interobserver range of 0.89 to 0.97 with the original CMS in categories. In these studies, like in our study, a predefined working protocol was used.[3,5,16,17] A predefined working protocol will remove a lot of variation within the trial that does normally exist in standard clinical practice. Therefore, an overestimation of the interobserver reliability could be made. In our study, the interobserver reliability of the CMS was probably affected positively. Generally, due to the different interpretations between the examiners if the protocol is not accurately standardized in a clinical trial, performing a clinical trial without a standardized working protocol is not favorable.

The smiley face slider used in this study is mostly used for children. However, Sasaki et al.[18] used the smiley face score with adults for quality-of-life questions and reported adequate test–retest results (ICC of 0.80). In our study, the results from the smiley face score were comparable with the other scores with high intra-reliability (ICC of 0.99) and inter-reliability (ICC between 0.89 and 0.97) and therefore can be reliably used in adults.

Limitations

A limitation of this study might be the short time intervals between the assessments, i.e., before and after the consultation with the orthopaedic surgeon, which might cause recall bias and thereby overestimation of the measurement properties. Furthermore, the consultation with the orthopaedic surgeon could have influenced the second assessment. Patients might answer differently when they just heard that they, e.g., need to have surgery compared with patients who just heard that the surgeon is very satisfied with the results or physical examination.

Conclusions

The different answering scales for the subjective subscales within the CMS for pain, work, and recreational activity were not interchangeable on item level and significantly influenced the total CMS score. Differences were below the smallest detectable change and interpreted as not clinically relevant. Particularly on item level, data from different studies cannot be pooled and compared when different answering scales are being used. The inter- and intraobserver reliability were excellent.

Fig 2. The Bland–Altman plot of the original CMS and the CMS using the paper VAS. The middle line indicates the mean difference of the 2 measurements (2.62 points). The upper (9.49) and lower (−4.25) lines are the limits of agreement (mean difference ± 1.96 * SD of the difference). Every patient is represented by a circle. (CMS, Constant–Murley Score; SD, standard deviation; VAS, visual analog scale.)
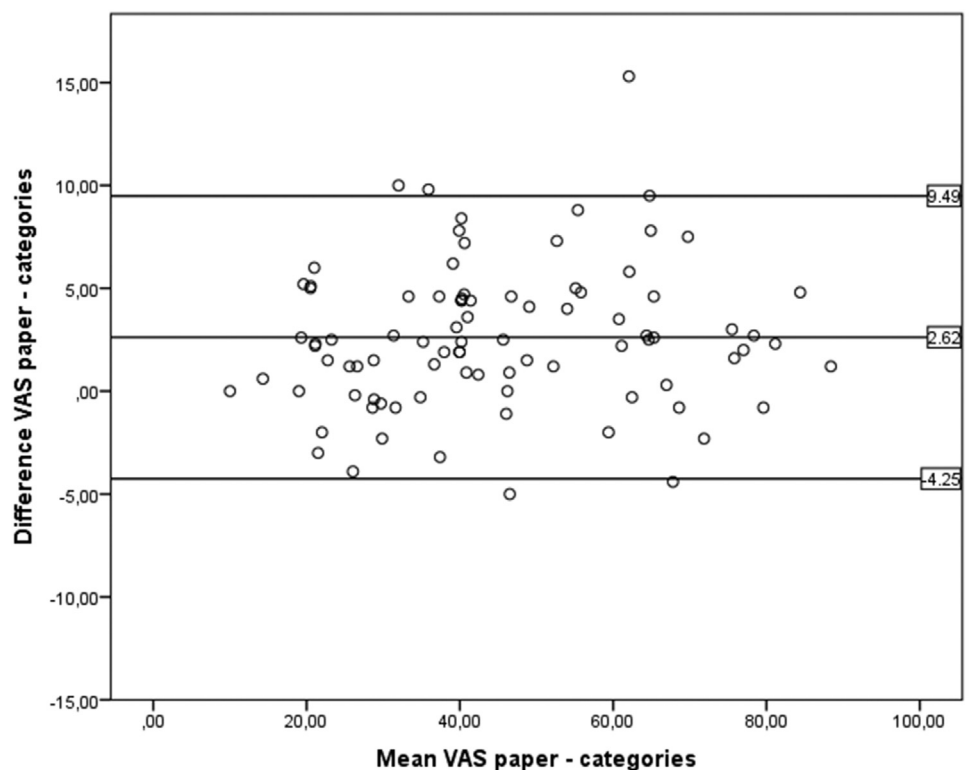
References

1. Constant CR, Murley AH. A clinical method of functional assessment of the shoulder. *Clin Orthop Rel Res* 1987;214: 160-164.
2. Constant CR, Gerber C, Emery RJH, Søjbjerg JO, Gohlke F, Boileau P. A review of the Constant score: Modifications and guidelines for its use. *J Shoulder Elbow Surg* 2008;17:355-361.
3. Moeller AD, Thorsen RR, Torabi TP, et al. The Danish version of the modified Constant-Murley Shoulder Score: Reliability, agreement, and construct validity. *J Orthop Sports Phys Ther* 2014;44:336-340.
4. Razmjou H, Bean A, MacDermid JC, van Osnabrugge V, Travers N, Holtby R. Convergent validity of the Constant-Murley outcome measure in patients with rotator cuff disease. *Physiother Can* 2008;60:72-79.
5. Rocourt MHH, Radlinger L, Kalberer F, et al. Evaluation of intratester and intertester reliability of the Constant-Murley shoulder assessment. *J Shoulder Elbow Surg* 2008;17:364-369.
6. Ban I, Troelsen A, Christiansen HD, Svendsen WS, Kristensen T. Standardised test protocol (Constant Score) for evaluation of functionality in patients with shoulder disorders. *Danish Med J* 2013;60:A4608.
7. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34-42.
8. Bland JM, Altman DG. Cronbach's alpha. *BMJ* 1997;314: 572.
9. Mahabier KC, Den Hartog D, Theyskens N, Verhofstad MHJ, Van Lieshout EMM. Reliability, validity, responsiveness, and minimal important change of the Disablties of the Arm, Shoulder and Hand and Constant-Murley scores in patients with a humeral shaft fracture. *J Shoulder Elbow Surg* 2017;26.
10. Vrotsou K, Ávila M, Machón M, et al. Constant−Murley Score: Systematic review and standardized evaluation in different shoulder pathologies. *Qual. Life Res* 2018;27: 2217-2226.
11. Lund I, Lundeberg T, Sandberg L, Budh CN, Kowalski J, Svensson E. Lack of interchangeability between visual analogue and verbal rating pain scales: A cross sectional description of pain etiology groups. *BMC Med Res Methodol* 2005;4:31:5.
12. Ferreira-Valente MA, Pais-Ribeiro JL, Jensen MP. Validity of four pain intensity rating scales. *Pain* 2011;152: 2399-2404.
13. Sriwatanakul K, Kelvie WBS, Lasagna L, Calimlim JF, Weis OF, Metha G. Studies with different types of visual analog scales for measurement of pain. *Clin Pharmacol Ther* 1983;34:234-239.
14. Conboy VB, Morris RW, Kiss J, Carr AJ. An evaluation of the Constant-Murley shoulder assessment. *J Bone Joint Surg Br* 1996;78:229-232.
15. Fialka C, Oberleitner G, Stampfl P, Brannath W, Hexel M, Vécsei V. Modification of the Constant−Murley shoulder score—introduction of the individual relative Constant score. Individual shoulder assessment. *Injury* 2005;36: 1159-1165.
16. Celik D. Turkish version of the modified Constant-Murley score and standardized test protocol: reliability and validity. *Acta Orthop Traumatol Turc* 2016;50:69-75.
17. Johansson KM, Adolfsson LE. Intraobserver and interobserver reliability for the strength test in the Constant-Murley shoulder assessment. *J Shoulder Elbow Surg* 2005;14:273-278.
18. Sasaki H, Kakee N, Morisaki N, Mori R, Ravens-Sieberer U, Bullinger M. Assessing health-related quality of life in young Japanese children with chronic conditions: Preliminary validation of the DISABKIDS smiley measure. *BMC Pediatr* 2017;17:100-116.