


## RESEARCH ARTICLE

## Open Access



# Inter-rater reliability, sensitivity to change and responsiveness of the orthopaedic Wolf-Motor-Function-Test as functional capacity measure before and after rehabilitation in patients with proximal humeral fractures

Corinna Nerz<sup>1\*</sup> , Lars Schwickert<sup>1</sup>, Sabine Schölch<sup>1</sup>, Katharina Gordt<sup>2</sup>, Philip-Christian Nolte<sup>3</sup>, Inga Kröger<sup>4</sup>, Peter Augat<sup>4,5</sup> and Clemens Becker<sup>1</sup>

## Abstract

**Background:** The incidence of proximal humeral fractures (PHF) increased by more than 30% over the last decade, which is accompanied by an increased number of operations. However, the evidence on operative vs. non-operative treatment and post-operative treatments is limited and mostly based on expert opinion. It is mandatory to objectively assess functional capacity to compare different treatments. Clinical tools should be valid, reliable and sensitive to change assessing functional capacity after PHFs. This study aimed to analyse inter-rater reliability of the videotaped Wolf-Motor-Function-Test-Orthopaedic (WMFT-O) and the association between the clinical WMFT-O and the Disability of the Arm, Shoulder and Hand (DASH) and to determine the sensitivity to change of the WMFT-O and the DASH to measure functional capacity before and after rehabilitation in PHF patients.

**Methods:** Fifty-six patients ( $61.7 \pm 14.7$  years) after surgical treatment of PHF were assessed using the WMFT-O at two different time points. To determine inter-rater reliability, the videotaped WMFT-O was evaluated through three blinded raters. Inter-rater agreement was determined by Fleiss' Kappa statistics. Pearson correlation coefficients were calculated to assess the association between the clinical WMFT-O and the video rating as well as the DASH. Sensitivity to change and responsiveness were analysed for the WMFT-O and the DASH in a subsample of forty patients ( $53.8 \pm 1.4$  years) who were assessed before and after a three week robotic-assisted training intervention.

**Results:** Inter-rater agreement was indicated by Fleiss' Kappa values ranging from 0.33–0.66 for functional capacity and from 0.27–0.54 for quality of movement. The correlation between the clinical WMFT-O and the video rating was higher than 0.77. The correlation between the clinical WMFT-O and the DASH was weak. Sensitivity to change was high for the WMFT-O and the DASH and responsiveness was given. In comparison to the DASH, the sensitivity to change of the WMFT-O was higher.

(Continued on next page)

\* Correspondence: [corinna.nerz@rbk.de](mailto:corinna.nerz@rbk.de)

<sup>1</sup>Department for Clinical Gerontology, Robert-Bosch-Hospital, Auerbachstr. 110, 70376 Stuttgart, Germany

Full list of author information is available at the end of the article



(Continued from previous page)

**Conclusion:** The overall results indicate that the WMFT-O is a reliable, sensitive and responsive instrument to measure more objectively functional change over time in rehabilitation after PHF. Furthermore, it has been shown that video assessment is eligible for studies to ensure a full blinding of raters.

**Trial registration:** Clinicaltrials.gov, [NCT03100201](https://clinicaltrials.gov/ct2/show/study/NCT03100201). Registered on 28 March 2017. The trial was retrospectively registered.

**Keywords:** Wolf motor function test, Reliability, Sensitivity to change, Responsiveness, Orthopaedic assessment, Shoulder function, Proximal humeral fracture

## Background

The individual and societal burden of musculoskeletal injuries, in particular of bone fractures remains to be underestimated. Proximal humeral fractures (PHF) are among the leading causes of functional impairment in patients after trauma resulting in limitations in basic, instrumental and advanced activities of daily living. PHFs and wrist fractures are recognized as the most common fractures of the upper extremities accounting for more than 20% of hospital admissions caused by a fracture [1]. In patients over 40 years of age, the proportion of PHFs increases to 76% [2]. Since 2000 in Germany the incidence of PHFs has risen from 178 to 246/100.000 inhabitants/year [3]. In addition, an analysis from Bauer and colleagues showed that one third of the patients are still integrated into the work process [4]. Due to the demographic change, a further increase in the number of PHFs is expected [5, 6]. This will lead to a significant increase in PHFs requiring operative or non-operative treatment and post-trauma hospitalization and rehabilitation.

To date, there is no robust evidence-based consensus on rehabilitation after PHF regarding standardisation of content, duration, intensity or frequency [7–9]. One essential requirement to perform controlled studies on surgical and rehabilitation interventions is the availability of objective, reliable and valid assessments. If possible, these assessment tools should be blinded to treatment allocation. To assess functional capacity and task performance after PHF, at least two types of measures are required: patient-reported outcomes using questionnaires to assess activities of daily living and a supervised clinical-based assessment to measure functional capacity of the patients [10, 11]. The Disability of the Arm, Shoulder and Hand (DASH) questionnaire is the most commonly used questionnaire for assessing activities of daily living after shoulder and arm injuries [10]. A clinically administered assessment of the functional capacity including the quality of movement of PHF patients is the Wolf-Motor-Function-Test-Orthopaedic (WMFT-O) which has previously been assessed regarding re-test reliability, inter-rater reliability, and internal consistency [12]. One further property required of an outcome measurement is the sensitivity to change. It is understood as the ability to describe changes occurring during a treatment or observational

period. The particular meaning of this property is described by the fact that positive changes in a given period represent the classic therapeutic goal [13]. Beyond assessing functional change in clinical state over time with sufficient sensitivity to change [14–16] clinical-based assessment tools also need a high responsiveness to decide if a change over time is clinically meaningful [17, 18]. There are no studies that examined the sensitivity to change of the WMFT or the WMFT-O which shows the need to consider also the change over time of the functional capacity and the quality of movement of the WMFT-O.

The aims of this study were 1st to test the inter-rater reliability of the videotaped WMFT-O, 2nd to describe the correlation of the functional capacity assessed by the WMFT-O and the activities of daily living from a patient perspective assessed by the DASH questionnaire, and 3rd to describe the sensitivity to change and the responsiveness of the WMFT-O and the DASH in a group of patients with PHFs.

## Methods

### Patients

For testing inter-rater reliability of the videotaped WMFT-O two patient populations were assessed. The first sample was a group of sixteen patients with an age range from 75 to 90 years with surgical treatment after PHF [12]. Due to the funding guidelines of the sponsor of the study (“Deutsche Gesetzliche Unfallversicherung (DGUV)”) only persons up to 69 could be included into the intervention study. This decision is based on the study approach to be able to quickly re-integrate patients after PHF into everyday work [19]. The second group consisted of 40 patients with an age range from 34 to 69 years with surgical treatment after PHF participating in a randomised controlled trial to measure the effectiveness of robot-assisted training added to conventional rehabilitation [19]. The proximal humeral fractures of both patient groups were surgically fixed by plate osteosynthesis, screw fixation, endoprotheses or humeral nails. The second patient group was recruited at three different clinical sites in Germany and patients were randomised into an intervention group and a control group. They were assessed before randomisation (baseline) and after completing an intervention period of

3 weeks (reassessment). At baseline, cognition was assessed by the Short Orientation-Memory-Concentration Test [20] as well as visual acuity, gait speed (10-m walk [21]), level of pain in the affected arm, ability to work, disability of the arm, shoulder and hand (DASH [22]), range of motion of the affected arm (goniometer measurement [23]), and motor function of the affected arm and shoulder (WMFT-O; [12]). Clinical reassessment directly after the intervention assessed disability of the arm, shoulder and hand (DASH) as well as range of motion and functional capacity (WMFT-O). The WMFT-O was videotaped at both time points. To analyse the sensitivity to change and the responsiveness of the WMFT-O the second subsample with 40 patients was analysed. Both studies were approved by the ethical committee of the University of Tübingen and are in agreement with the Declaration of Helsinki. All participants gave written informed consent.

#### Outcome measures

The **WMFT-O** is an adapted version of the basic WMFT [24] with good clinical inter- and intra-rater reliability. The modified version for shoulder injuries was developed by our group [12]. The WMFT was developed to measure functional improvement between the beginning and the end of therapy. The WMFT-O is supervised and includes 20 arm motion tasks used in daily living that are evaluated in terms of functional capacity and quality of movement. The single items are progressing from coarse movements in the elbow and shoulder area to more complex and dexterous tasks in the fingers and the hand area. The endpoint is a total score calculated from the ratings of the functional capacity and the quality of movement (5 is the best value, 0 the worst). The total WMFT-O score ranges from 0 to 100 points with lower scores indicating greater disability (Additional file 1: Video tutorial Rating Scale of the WMFT-O: <https://youtu.be/WQUSl2XQJMY> and Additional file 2: Video tutorial performance instructions of the WMFT-O: [https://youtu.be/K6g3Z\\_ibNa8](https://youtu.be/K6g3Z_ibNa8)). First a clinical rating was performed by pre-trained assessors. As a next step each videotaped WMFT-O (baseline and reassessment,  $n = 112$ ) was assessed by three out of five pre-trained raters (4 physiotherapists, 1 occupational therapist).

The first module of the **DASH questionnaire** was used to measure self-perceived activity and limitation of shoulder, arm and hand function [22]. The DASH is considered as a valid and reliable non-supervised test [25–28]. The questionnaire consist of 30 questions to assess the restrictions related to the function and activity of the shoulder, arm and hand in daily living, as well as self-esteem and potentially existing symptoms of the shoulder, arm and hand, such as pain or prickle. The

endpoint is a total score calculated from the ratings of the individual responses (1 is the best value, 5 the worst). The total DASH score ranges from 0 to 100 points with higher scores indicating greater disability.

#### Data analysis

All statistics and outcome analyses were performed using RStudio software (Version 1.1.383). The inter-rater agreement of the video rating was determined by Fleiss' Kappa statistics with corresponding confidence intervals (CI). Pearson correlation coefficients were calculated to assess the association between the WMFT-O clinical rating and the video rating as well as between the WMFT-O clinical rating and the DASH questionnaire.

*Sensitivity to change* was defined as the ability of an instrument to respond to changes in the measured construct, regardless of whether the change is relevant or meaningful to the decider [14, 18]. In order to analyse the sensitivity to change of the WMFT-O, the change scores and the standardized effect size as well as the standardized response mean [29] were calculated. Change scores imply the delta of the results between the baseline- and reassessment. This score represents the extent of which a patient changes in performance in the corresponding test. The Wilcoxon test was calculated to compare the baseline and the reassessment score ( $p > 0.01$ ). The standardized effect size was calculated by dividing the change score by the standard deviation of the baseline score [30]. The standardized response mean was calculated by dividing the change score by the standard deviation of that change score [31]. Values  $< 0.01$  for the standardized effect size are considered as very small,  $< 0.2$  as small,  $< 0.5$  as medium,  $< 0.8$  as large,  $< 1.2$  as very large and  $< 2.0$  as huge [32]. Husted and colleagues [29] interpreted the values of the standardized effect size and the standardized response mean considering the same benchmarks ( $< 0.2$  for trivial,  $0.2$  to  $< 0.5$  for small,  $0.5$  to  $< 0.8$  for moderate, and  $0.8$  or greater for large). Cohen's threshold values are  $> 0.8$  equates to large,  $> 0.5$  to medium and  $> 0.2$  to small effect sizes and intended for intervention studies, but are sometimes used to apply sensitivity to change of questionnaires [31].

*Responsiveness* was defined as the ability of an instrument to measure a meaningful or important change in a clinical state [14, 18] and is usually reported through the minimal important difference [33, 34]. For being considered as important, a change score of a measure should equal or exceed its minimal important difference estimate. Minimal important difference values were calculated using two commonly used effect size estimates in the literature:  $0.3^*$  standard deviation of the baseline score and  $0.5^*$  standard deviation of the baseline score [34, 35].

## Results

The study sample consisted of two groups. The first subsample included 16 older patients with a mean age of 81.4 years (range 75–90 years). The second subsample includes 40 younger patients with a mean age of 53.8 years (range 39–69 years). The subject characteristics are listed in Table 1.

### Inter-rater reliability of the videotaped WMFT-O

Table 2 lists inter-rater reliability across all items of the videotaped WMFT-O. The Fleiss' Kappa values ranged between 0.35 and 0.67 (CI = 0.27 to 0.73) for the functional capacity and between 0.27 and 0.54 (CI = 0.21 to 0.60) for the quality of movement. Grip strength (task 15) was not considered since the task could not be analysed from the video recordings.

### Correlation between the WMFT-O clinical and video rating and the DASH

Pearson correlation coefficients ( $r$ ) of the WMFT-O clinical rating and the video rating were high for all three raters and highly significant (functional capacity and quality of movement  $r \geq 0.82$ ,  $p < 0.001$ ) (Fig. 1). After reducing missing or occluded video recordings, the Pearson correlation coefficient for the functional capacity could be calculated for 49 subjects and for 48 subjects for the quality of movement. Due to the uniform results of the three different raters, in Fig. 1 only the ratings of the first rater are shown.

The correlation between the WMFT-O clinical rating and the DASH questionnaire was significantly weak at baseline (functional capacity  $r = -0.27$  and quality of movement  $r = -0.32$ ,  $p < 0.05$ ) (Fig. 2). The Pearson correlation coefficient for the DASH could be calculated for 56 subjects.

### Sensitivity to change and responsiveness

The sensitivity to change and the responsiveness indices of the WMFT-O and DASH are listed in Table 3. For the WMFT-O the clinical rating was carried out locally in the individual study centres by one trained rater and the DASH was completed by the subject himself on site. The baseline and the reassessment scores for the WMFT-O functional capacity and quality of movement as well as the DASH differ significantly from each other ( $p < 0.01$ ). The WMFT-O and the DASH demonstrated large standardized effect sizes, ranging between 0.8 and 0.9 (Table 3). Large standardized response means were also obtained for the WMFT-O functional capacity, quality of movement and the DASH. The standardized response mean for the DASH was slightly less sensitive to change when compared to the WMFT-O functional capacity and quality of movement. The minimal important differences for the WMFT-O functional capacity ranged between 5.0 (0.3\* standard deviation of the baseline score) and 8.3 (0.5\* standard deviation of the baseline score) and for the WMFT-O quality of movement between 6.3 (0.3\* standard deviation of the baseline score) and 10.5 (0.5\* standard deviation of the baseline score) and are comparable to the minimal important differences of the DASH (Table 3).

## Discussion

Our study on patients with fractures of the proximal humerus demonstrated high sensitivity to change and good responsiveness of the orthopaedic Wolf-Motor-Function-Test indicating its usefulness as a functional capacity assessment tool in operative and rehabilitation studies.

This study also found moderate inter-rater reliability for the videotaped version of the WMFT-O (Table 2). According to Landis and Koch [36] the calculated Fleiss'

**Table 1** Participant demographics

Characteristic	1st Subsample	2nd Subsample	Total sample
Subjects (n)	16	40	56
Gender			
Female (%)	13 (81.2%)	25 (62.5%)	38 (67.9%)
Age (years)			
Mean (SD)	81.4 (1.1)	53.8 (1.4)	61.7 (14.7)
Minimum	75	39	39
Maximum	90	69	90
Affected Arm			
Right (%)	10 (62.5%)	17 (42.5%)	27 (48.2%)
Arm Dominance <sup>a</sup>			
Left (%)	1 (6.3%)	5 (12.5%)	6 (10.9%)
Right (%)	15 (93.7)	34 (85.0%)	49 (89.1%)
Dominant arm affectedd (%)	9 (56.3%)	17 (42.5%)	26 (47.3%)

n Number, SD Standard deviation; <sup>a</sup>Dominant arm could not be determined in one subject.

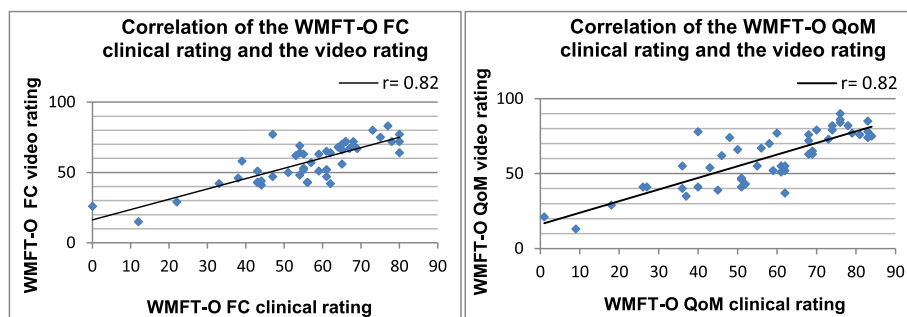
**Table 2** Inter-rater reliability across all items of the video WMFT-O;  $n = 56$  ( $p < 0.005$ )

Task	Functional Capacity	Quality of Movement
	Inter-rater reliability Fleiss' Kappa (95% CI)	Inter-rater reliability Fleiss' Kappa (95% CI)
1. Forearm to table lateral	0.41 (0.35–0.48)	0.30 (0.24–0.37)
2a/b. Forearm to box 15/30 cm lateral	0.47 (0.41–0.53)	0.39 (0.34–0.44)
3. Forearm to box lateral with weight	0.64 (0.57–0.71)	0.54 (0.48–0.60)
4. Extend elbow lateral	0.49 (0.44–0.55)	0.46 (0.41–0.51)
5. Extend elbow lateral with weight	0.56 (0.50–0.62)	0.43 (0.38–0.49)
6. Hand to table frontal	0.60 (0.53–0.67)	0.38 (0.32–0.45)
7a/b. Hand to box 15/30 cm frontal	0.55 (0.49–0.61)	0.51 (0.46–0.57)
8. Hand to box frontal with weight	0.67 (0.60–0.73)	0.53 (0.48–0.59)
9. Reach and retrieve frontal	0.49 (0.43–0.55)	0.44 (0.37–0.50)
10. Lift can frontal	0.59 (0.52–0.65)	0.50 (0.44–0.56)
11. Lift pencil frontal	0.60 (0.53–0.67)	0.45 (0.38–0.51)
12. Lift paper clip frontal	0.51 (0.45–0.58)	0.39 (0.32–0.45)
13. Stack checkers frontal	0.50 (0.44–0.57)	0.34 (0.27–0.40)
14. Flip cards frontal <sup>a</sup>	0.49 (0.43–0.56)	0.37 (0.30–0.43)
16. Turn key in lock frontal	0.35 (0.27–0.42)	0.27 (0.21–0.33)
17. Fold towel frontal	0.49 (0.42–0.55)	0.42 (0.36–0.47)
18. Lift basket frontal	0.47 (0.41–0.53)	0.43 (0.37–0.48)

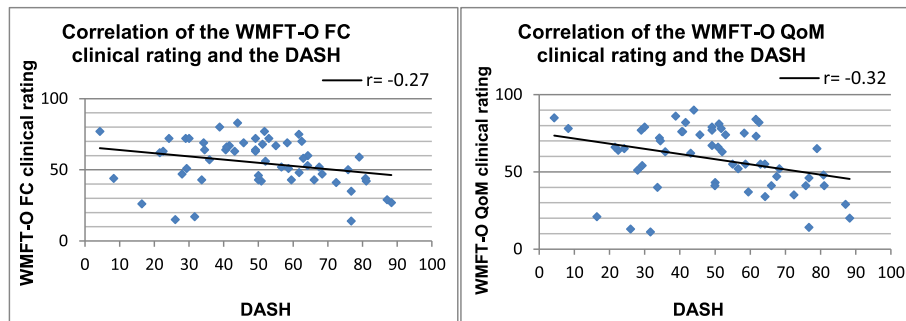
CI Confidence Interval, <sup>a</sup> task 15 (grip strength) was omitted.

Kappa values for the functional capacity and quality of movement can be interpreted as a fair to substantial agreement. The observed agreements were weaker than in a previous clinical study using a non-videtaped assessment of WMFT-O in similar patients [12]. Another publication investigating the reliability of the neurological based WMFT-N version [37] also had higher interrater agreement. The lower agreement of the videotaped measurement could have several reasons. Possibly the training session of the video raters were not enough. A further explanation for the lower agreement could also be an inadequate positioning of the chosen video camera and a different camera position might lead to better results. The performance of the tasks on such

videos (e.g. in task 16- “Turn key in lock frontal”, Fleiss’ Kappa for functional capacity =0.33 and quality of movement =0.27; Table 2) was not easy to identify which resulted in more diverse ratings between the raters. Moreover, an inadequate description of the correct end-position in task 1 (“Forearm to table lateral”, Fleiss’ Kappa functional for capacity =0.41 and quality of movement =0.30; Table 2) lead to different opinions about fulfilling or not fulfilling the task. A possible solution could be reducing the prescribed tasks to improve rater agreement. The tasks that were poorly rated could potentially be omitted and a future short version of the WMFT-O could be provided. According to Landis and Koch [36], an inter-rater reliability of 0.01 to 0.20 is considered as a



**Fig. 1** Correlation of the WMFT-O functional capacity and quality of movement clinical rating and video rating (FC Functional capacity, QoM Quality of movement; Pearson correlation coefficient ( $r$ );  $p < 0.001$ )



**Fig. 2** Correlation of the WMFT-O functional capacity and quality of movement clinical rating and DASH (FC Functional capacity, QoM Quality of movement; Pearson correlation coefficient ( $r$ );  $p < 0.05$ )

slight agreement and 0.21 to 0.40 as a fair agreement. If these values were used as a basis to decide which items of the WMFT-O could be deleted in a short version, this would delete task 1 and 16 due to the low inter-rater reliability of the functional capacity and quality of movement as well as task 2ab, 12, 13 and 14 due to the low inter-rater reliability of the quality of movement (Table 2).

A third aspect is the calculation of different statistical measures. Inter-rater agreement for two different raters, as calculated in Oberle and colleagues 2018 [12] should be determined by weighted Cohen’s Kappa ( $K_w$ ) statistics [38]. For three or more raters, Fleiss’ Kappas is recommended as the method of choice. The Fleiss’ Kappa assumes that the examiners were randomly selected from a group of available examiners. Cohen’s Kappa, on the other hand, assumes that examiners have been specifically selected and trained. Therefore, the probability of agreement in the Fleiss’ Kappa and the Cohen’s Kappa is estimated in different ways. In some cases, Fleiss’ Kappa in general may produce lower values even if the agreement is actually high as described before [39]. That in turn may be a possible explanation why the inter-rater reliability values of the video-based WMFT-O calculated by Fleiss’ Kappa (Table 2) were lower than the inter-rater reliability values of the clinically WMFT-O as reported by Oberle and colleagues [12] and calculated by Cohen’s Kappa.

We found a very strong correlation between the WMFT-O clinical baseline rating and the video baseline ratings for the functional capacity and a strong correlation

for the quality of movement according to the standards of Evans [40].

To assess treatment effects of interventions for musculoskeletal conditions, functional capacity and personal activity need to be evaluated. Therefore, the WMFT-O has to be augmented by other methods to evaluate levels of disability, activity and participation. A widely applied method is the DASH. It is expected that the correlation between activity levels and functional capacity is often less than anticipated. The correlation between the clinical-based measures (WMFT-O clinical rating of the functional capacity and the quality of movement) and the patient-reported questionnaires (DASH) at baseline was indeed weak [40]. One aspect is that patients do not return to their activity levels due to psychological problems such as insufficient self-efficacy. Other aspects are methodological problems. The DASH is not explicitly designed for the affected arm. This means that in instances where tasks are mostly carried out with the dominant hand (e.g. turning a key in a lock) the restriction in the non-dominant arm, shoulder and hand are not necessarily being captured through the DASH. This is only the case if the affected hand is also the dominant hand. This misjudgement could be avoided if care is taken that when answering the questions of the DASH, the assessment of the restriction always relates to the performance of the activity with the affected shoulder, arm or hand. If it is not possible to carry out the activity of daily living with the affected shoulder, arm or hand the patient must be able to imagine the execution of the

**Table 3** Sensitivity to change and responsiveness of the clinical WMFT-O and DASH;  $n = 40$

Parameter	Baseline Mean (SD)	Reassessment Mean (SD)	CS (SD)	SES	SRM	MID (0.35SD <sub>b</sub> )	MID (0.5SD <sub>b</sub> )
WMFT-O							
FC	54.3 (16.5)	69.2 (12.8)	14.9 (10.1)	0.9	1.5	5.0	8.3
QOM	57.7 (20.9)	76.1 (14.8)	18.4 (14.1)	0.9	1.3	6.3	10.5
DASH	51.9 (20.0)	36.7 (16.6)	-15.2 (12.6)	0.8	1.2	6.0	10.0

SD Standard deviation, CS Change score, SES Standardized effect size, SRM Standardized response mean, MID Minimally important difference, SD<sub>b</sub> Standard deviation Baseline, WMFT-O Wolf-Motor-Function-Test-Orthopaedic, FC Functional capacity, QoM Quality of Movement.

activity of daily living and the possible restrictions as best as possible and then answer the question.

One study from Wu and colleagues [41] developed the streamlined WMFT which includes the performance rating of 6 timed tasks for neurological patients. They found a low effect size for the streamlined WMFT and the original WMFT [41]. In comparison to this study the WMFT-O had a large effect size for both the functional capacity and the quality of movement and can thus be regarded as being sensitive to change over time according to the published standards [29, 32]. This result was confirmed by the absolute values of the standardized response mean, which can be considered higher than the standardized effect sizes for both the functional capacity and the quality of movement. This is a relevant finding in terms of clinical use to provide a more objective and sensitive measure for assessing functional capacity and quality of movement of the upper extremities for patients in orthopaedic rehabilitation and may improve current assessments which currently are mostly subjective. Compared to the WMFT-O, the sensitivity to change of the DASH was lower. These findings indicate that the WMFT-O is a more sensitive outcome measure for assessing functional change over time through rehabilitation in patients with PHF. MacDermid and colleagues [42] and Westphal and colleagues [11] determined the sensitivity to change of the DASH for patients after wrist fractures. They found higher values for the effect size of the DASH in patients with wrist fractures after 0–3 months. After 3–12 months they observed to be somewhat lower [11]. This indicates that the standardized effect sizes and standardized response mean depend on the point of time of the baseline- and reassessment. In the first 12 weeks after baseline, a large treatment effect can be expected. The next three quarters might lead to further improvements but the effect sizes will be smaller. In our study, the baseline assessment was conducted approximately one month after surgery. In order to find out whether the WMFT-O can also be used to assess long-term therapy results, a future study should be carried out including longer therapy intervals assessing upper extremity functional capacity measured with the WMFT-O after three month. A similar relation was found for the DASH questionnaire. In conclusion, both the WMFT-O and the DASH are responsive and complimentary assessment instruments in order to measure the functional change in patients with PHF.

## Conclusion

The WMFT-O is a responsive instrument to objectively measure patient functional change in younger and older patients with PHF. It showed a somewhat higher sensitivity to change in younger and older patients with PHF

compared to the DASH. For assessing treatment success of interventions and rehabilitation functional capacity could be measured by the WMFT-O but it should be augmented by another method (for example the DASH) to enable the evaluation of levels of disability, activity and participation. We recommend minor modifications of the camera positions and of the descriptions of some movement tasks for the videotaped WMFT-O. After a detailed training of the rater and taking into account these changes we also recommend the application of a video based assessment as this would allow robust blinding of assessors which often is not the case due to contamination of therapists and assessment staff. Furthermore a short version of the WMFT-O would be desirable.

## Additional files

**Additional file 1:** Video tutorial Rating Scale of the WMFT-O. Video tutorial with detailed instructions how to rate the single items of the WMFT-O (functional capacity and quality of movement) in English. (MP4 6957 kb)

**Additional file 2:** Video tutorial performance instructions of the WMFT-O. Video tutorial with detailed instructions how to perform the single items of the WMFT-O in English. (MP4 21016 kb)

## Abbreviations

CI: Confidence Intervals; CS: Change Score; DASH: Disability of the Arm, Shoulder and Hand; FC: Functional Capacity; MID: Minimally Important Difference; PHF: Proximal Humeral Fractures; QoM: Quality of Movement; r: Pearson correlation coefficients; SD: Standard Deviation; SD<sub>b</sub>: Standard Deviation baseline; SES: Standardized Effect Size; SRM: Standardized Response Mean; WMFT-O: Wolf-Motor-Function-Test-Orthopaedic; Kw: Cohen's Kappa

## Acknowledgments

The authors gratefully acknowledge C. Endress (BSc), E. Ermer (BSc) and R. Leonhardt (BSc) for their support in rating the videotaped WMFT-O.

## Authors' contributions

CN, LS, CB and PA were major contributors in the writing of the manuscript. CN, LS, CB, PN and PA were involved in the methodological development of the study protocol. CN, SS, KG and IK helped with the recruitment and data collection. All authors read and approved the final manuscript.

## Funding

The study is funded and supported by the German Social Accident Insurance ("Deutsche Gesetzliche Unfallversicherung (DGUV) - Alte Heerstrasse 111, 53757 St. Augustin", project number 412.02-FR-0233). The DGUV had no influence on data assessment or the writing of the manuscript.

## Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Ethics approval and consent to participate

Both studies were approved by the ethical committee of the University of Tübingen (study number 381/2015BO1) and are in agreement with the Declaration of Helsinki. All participants gave written informed consent.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Department for Clinical Gerontology, Robert-Bosch-Hospital, Auerbachstr. 110, 70376 Stuttgart, Germany. <sup>2</sup>Network Aging Research (NAR), Heidelberg University, Heidelberg, Germany. <sup>3</sup>Department of Trauma and Orthopaedic Surgery, BG Trauma Centre Ludwigshafen, Ludwigshafen, Germany. <sup>4</sup>Institute for Biomechanics, BG Unfallklinik Murnau, Murnau, Germany. <sup>5</sup>Institute for Biomechanics, Paracelsus Medical University, Salzburg, Austria.

Received: 12 March 2019 Accepted: 25 June 2019

Published online: 06 July 2019

**References**

- Benzinger P, Riem S, Bauer J, Jaensch A, Becker C, Büchele G, et al. Risk of institutionalization following fragility fractures in older people. *Osteoporos Int*. 2019;1–8. Epub ahead of print.
- Kara H, Bayir A, Ak A, Acar D, Akinci M, Degirmenci S. Trivial trauma induced bilateral proximal end Humerus fracture: two case reports. *J Case Rep*. 2013; 3(2):366–9.
- Statistisches Bundesamt. Krankenhausstatistik - Diagnosedaten der Patienten und Patientinnen in Krankenhäusern. Wiesbaden: Statistisches Bundesamt Deutschland; c2017 [cited 2019 Feb 25]. Available from: <https://www-genesis.destatis.de/genesis/online?sequenz=statistikTabellen&selectionname=23131>. German.
- Bauer M. Klinikinterne analyse epidemiologischer Aspekte der proximalen Humerusfraktur [dissertation]. Tübingen: Universität Tübingen; 2014. [cited 2019 Feb 18]. Available from: <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/56031>. German
- Palvanen M, Kannus P, Niemi S, Parkkari J. Update in the epidemiology of proximal humeral fractures. *Clin Orthop* Januar. 2006;442:87–92.
- Kim SH, Szabo RM, Marder RA. Epidemiology of humerus fractures in the United States: nationwide emergency department sample, 2008. *Arthritis Care Res*. 2012;64(3):407–14.
- Zech A, Hübscher M, Vogt L, Banzer W, Hänsel F, Pfeifer K. Neuromuscular training for rehabilitation of sports injuries: a systematic review. *Med Sci Sports Exerc*. 2009;41(10):1831–41.
- Franke S, Ambacher T. Die proximale Humerusfraktur. *Obere Extrem*. 2012; 7(3):137–43 German.
- Handoll HH, Ollivere BJ, Rollins KE. Interventions for treating proximal humeral fractures in adults. *Cochrane Database Syst Rev*. c2012;12. Available from: <https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000434.pub3/abstract>. [cited 2019 Feb 18].
- Vincent JL, MacDermid JC, King GJ, Grewal R. Validity and sensitivity to change of patient-reported pain and disability measures for elbow pathologies. *J Orthop Sports Phys Ther*. 2013;43(4):263–74.
- Westphal T. Reliability and responsiveness of the German version of the disabilities of the arm, shoulder and hand questionnaire (DASH). *Unfallchirurg*. 2007;110(6):548–52 German.
- Oberle C, Becker C, Schölch S, Lenz J-U, Studier-Fischer S, Augat P, et al. Inter-rater and intra-rater reliability of an adapted Wolf motor function test for older patients with shoulder injuries. *Z Für Gerontol Geriatr*. 2018;51(3): 293–300.
- Beaton DE. Understanding the relevance of measured change through studies of responsiveness. *Spine*. 2000;25(24):3192–9.
- Liang MH, Lew RA, Stucki G, Fortin PR, Daltroy L. Measuring clinically important changes with patient-oriented questionnaires. *Med Care*. 2002; 40(4 Suppl):145–51.
- Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. *J Clin Epidemiol*. 2004;57(10):1008–18.
- Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Qual Life Res Int J Qual Life Asp Treat Care Rehab*. 2003;12(4):349–62.
- VanSwearingen JM, Brach JS. Making geriatric assessment work: selecting useful measures. *Phys Ther*. 2001;81(6):1233–52.
- Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care*. 2000;38(9 Suppl):1184–90.
- Nerz C, Schwickert L, Becker C, Studier-Fischer S, Müßig JA, Augat P. Effectiveness of robot-assisted training added to conventional rehabilitation in patients with humeral fracture early after surgical treatment: protocol of a randomised, controlled, multicentre trial. *Trials*. 2017;18(1):589.
- Katzman R, Brown T, Fuld P, Peck A, Schechter R, Schimmel H. Validation of a short orientation-memory-concentration test of cognitive impairment. *Am J Psychiatry*. 1983;140(6):734–9.
- Flansbjerg U-B, Drake AM, Downham D, Patten C, Lexell J. Reliability of gait performance tests in men and women with hemiparesis after stroke. *J Rehabil Med*01-010100. 2005;37(2):75–82.
- Germann G, Harth A, Wind G, Demir E. Standardisierung und Validierung der deutschen version 2.0 des "disability of arm, shoulder, hand" (DASH)-Fragebogens zur outcome-Messung an der oberen Extremität. *Unfallchirurg*. 2003;106(1):13–9 German.
- Ryf C, Weymann A. Range of motion - AO Neutral-0 method : measurement and documentation. Stuttgart: Thieme; 1999. p. 116.
- Wolf SL, Catlin PA, Ellis M, Archer AL, Morgan B, Piacentino A. Assessing Wolf Motor function test as outcome measure for research in patients after stroke. *Stroke*. 2001;32(7):1635–9.
- Desai AS, Dramis A, Hearnden AJ. Critical appraisal of subjective outcome measures used in the assessment of shoulder disability. *Ann R Coll Surg Engl*. 2010;92(1):9–13.
- Offenbaecher M, Ewert T, Sangha O, Stucki G. Validation of a German version of the disabilities of arm, shoulder, and hand questionnaire (DASH-G). *J Rheumatol*. 2002;29(2):401–2.
- Dalton E, Lannin NA, Laver K, Ross L, Ashford S, McCluskey A, et al. Validity, reliability and ease of use of the disabilities of arm, shoulder and hand questionnaire in adults following stroke. *Disabil Rehabil*. 2017;39(24):2504–11.
- Raven EEJ, Haverkamp D, Sierveit IN, van Montfoort DO, Pöhl RG, Blankevoort L, et al. Construct validity and reliability of the disability of Arm, shoulder and hand questionnaire for upper extremity complaints in rheumatoid arthritis. *J Rheumatol*. 2008;35(12):2334–8.
- Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol*. 2000;53(5):459–68.
- Kazis L, Anderson J, Meenan R. Effect sizes for interpreting changes in health status. *Med Care*. c 1989 27(3) [cited 2019 Feb 19]. Available from: [insights.ovid.com](https://insights.ovid.com).
- Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care*. 1990;28(7):632.
- Sawilowsky S. New effect size rules of thumb. *Theor Behav Found Educ Fac Publ*. 2009;8(2):597–99. Available from: [https://digitalcommons.wayne.edu/coe\\_tbf/4](https://digitalcommons.wayne.edu/coe_tbf/4)
- Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102–9.
- Eton DT, Yost KJ. Combining distribution- and anchor-based approaches to determine minimally important differences: the FACIT experience. *Eval Health Prof*. 2005;28(2):172–91.
- Cella D, Eton DT, Fairclough DL, Bonomi P, Heyes AE, Silberman C, et al. What is a clinically meaningful change on the functional assessment of cancer therapy-lung (FACT-L) questionnaire?: results from eastern cooperative oncology group (ECOG) study 5592. *J Clin Epidemiol*. 2002; 55(3):285–95.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–74.
- Pereira ND, Michaelsen SM, Menezes IS, Ovando AC, Lima RCM, Teixeira-Salmela LF. Reliability of the Brazilian version of the Wolf Motor function test in adults with hemiparesis. *Braz J Phys Ther*. 2011;15(3):257–65.
- Cohen J. Weighted kappa. *Psychol Bull*. 1968;70(4):213–20.
- Falotico R, Quatto P. Fleiss' kappa statistic without paradoxes. *Qual Quant*. 2015;49(2):463–70.
- Evans JD. Straightforward statistics for the behavioral sciences. Pacific Grove: Brooks/Cole Publishing; 1996.
- Wu CY, Fu T, Lin KC, Feng CT, Hsieh KP, Yu HW, et al. Assessing the streamlined Wolf Motor function test as an outcome measure for stroke rehabilitation. *Neurorehabil Neural Repair*. 2011;25(2):194–9.
- MacDermid JC, Richards RS, Roth JH. Distal radius fracture: a prospective outcome study of 275 patients. *J Hand Ther*. 2001;14(2):154–69.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.