Validity, reliability and responsiveness of a goniometer watch to measure pure forearm rotation

Link to publication record in Ulster University Research Portal

Document Version
Author Accepted version

Validity, reliability and responsiveness of a goniometer watch to measure pure forearm rotation

Validity, Reliability and Responsiveness of a Goniometer Watch to Measure Pure Forearm Rotation

Journal:	*Hand Therapy*
Manuscript ID	HTH-23-0032.R2
Manuscript Type:	Original article
Keywords:	Range of motion, articular, Wrist fractures, Therapeutics, Supination, Outcome measures
Abstract:	Introduction Innovative instruments have been designed to assess forearm rotation, an anatomically challenging motion to measure. This study assessed the validity, reliability and responsiveness of a novel goniometer watch (GoWatch) to measure pure forearm rotation. The modified finger goniometer (MFG) was the gold standard reference. Methods Forty participants with restricted forearm rotation were recruited. Two raters measured supination and pronation using the GoWatch and MFG before and after a hand therapy session. Repeated-measures ANOVA assessed for systematic bias with an apriori residual error of 10º deemed as acceptable. Secondary analysis used intraclass coefficients (ICCs) to categorise interrater reliability. Responsiveness of the GoWatch was calculated using cohen's d. Results The GoWatch demonstrated acceptable agreement with the MFG with mean difference for supination 1.19º and pronation 0.20º. Interrater reliability was also within acceptable limits with mean difference GoWatch supination 4.43ºand pronation 2.23º. Interrater reliability for GoWatch supination and pronation were categorized as excellent (ICC =0.94) and good (ICC=0.85) respectively. Systematic bias was observed in the instrument by rater interaction with rater 2 consistently underestimating GoWatch measures (p<0.05). GoWatch supination showed small to medium responsiveness (Rater 1: d=0.14; Rater 2: d=0.29) and pronation very small to medium responsiveness (Rater 1: d=0.29; Rater 2: d= 0.05). Discussion The GoWatch is a viable and user-friendly alternative to measure forearm rotation with demonstrable validity, interrater reliability and responsiveness. Further research is required to ensure systematic bias is not endemic when used across multiple raters.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

VALIDITY, RELIABILITY AND RESPONSIVENESS OF A GONIOMETER WATCH TO MEASURE PURE FOREARM ROTATION

ABSTRACT

Introduction

Innovative instruments have been designed to assess forearm rotation, an anatomically challenging motion to measure. This study assessed the concurrent validity, interrater reliability and responsiveness of a novel goniometer watch (GoWatch) to measure pure forearm rotation. The modified finger goniometer (MFG) was the criterion reference.

Methods

Forty participants with restricted forearm rotation were recruited. Two raters measured supination and pronation using the GoWatch and MFG before and after a hand therapy session. Repeated-measures ANOVA assessed for systematic bias with an *apriori* residual error of 5° deemed as acceptable. Secondary analysis used intraclass coefficients (ICCs) to categorise interrater reliability. Responsiveness of the GoWatch was calculated using cohen's *d*.

Results

The GoWatch demonstrated acceptable agreement with the MFG with mean difference for supination 1.19° and pronation 0.20°. Interrater reliability was also within acceptable limits with mean difference GoWatch supination 4.43° and pronation 2.23°. Interrater reliability for GoWatch supination and pronation were categorized as excellent (ICC =0.94) and good (ICC=0.85) respectively. Systematic bias was observed in the instrument by rater interaction with rater 2 consistently underestimating GoWatch measures ($p<0.05$). GoWatch supination showed small to medium responsiveness (Rater 1: d=0.14; Rater 2: d=0.29) and pronation very small to medium responsiveness (Rater 1: d=0.29; Rater 2: d= 0.05).

Conclusion

The GoWatch is a viable and user-friendly alternative to measure forearm rotation with demonstrable validity, interrater reliability and responsiveness. Further research is required to ensure systematic bias is not endemic when used across multiple raters.

INTRODUCTION

Forearm rotation requires unrestricted mobility of the proximal radioulnar, distal radioulnar and humeroradial joints making it possible to orientate the palm up or down in functional tasks(1). When injury occurs to these structures, function can be impaired (2). Supination in particular can be affected due to restrictions imposed by casts and splints, protective posturing and as a result of a natural tendency to perform most activities of daily living with the forearm in pronation(3). One example of an injury that can present with restrictions in forearm rotation are distal radius fractures. The incidence of this injury is rising globally(4) and as such patients with distal radius fractures are commonly assessed and treated by healthcare practitioners.

Outcome measures are vital for appropriate goal setting and to ascertain patient progress and subsequently must be reliable, valid and responsive(5). It is recommended when using the International Classification of Functioning, Disability, and Health (ICF) as a framework that assessment should incorporate *Body Functions and Structures*, *Participation, Activities* and the *Environment*(6). In the context of hand therapy, one example of *Body Functions and Structures* assessment is goniometry to objectively measure range of movement (ROM). Guidelines to standardise methods of goniometer measurement(7) recommend measuring forearm rotation using a two-arm goniometer with one arm placed on the wrist and the other aligned parallel to the humerus(7). This standard method measures pure forearm rotation and has excellent test-retest reliability and inter-rater reliability(8). However the circumference of the distal wrist is oval-like in shape posing issues with accuracy in placing the level surface of the goniometer arm. Additionally, the assessor is required to assume that the second arm is truly vertical using vision only.

2

An alternative approach is referred to as functional rotation, taking into consideration hand orientation and therefore may be more meaningful to everyday function. Two techniques to measure functional rotation have been described(9, 10). One involves the patient gripping a pencil while the therapist aligns one arm of the goniometer with the pencil(9). The second describes a tubular handle attached perpendicular to the horizontal arm of a standard goniometer and a plumb line attached to its axis(10). The patient holds the tubular handle to define the plane of the palm so when the forearm is rotated the weighted plumb line establishes the vertical plane. These measures have reported high intra-rater and inter-rater reliability(10-12), though several disadvantages to both approaches are noted. When measuring functional rotation, the patient must be able to grasp the handle or pencil to define the palmar plane, however commonly observed in patients with distal radius fractures is the inability to make a fist in the early phase of rehabilitation. Further, compensatory movements at the wrist and the fourth and fifth metacarpals need to be considered along with proximal positions.

A design by Szekeres *et al.* demonstrated how measurement of forearm rotation can be improved by attaching a plumb line to a finger goniometer with flat arms(13) (Figure 1). The weight on the goniometer uses gravity to achieve a true vertical position. The authors propose that the flat arms of the finger goniometer allow the therapist to secure it more firmly to the flatter surface of the dorsal aspect of the wrist during measurement, gaining a more stable position. This design was further tested and demonstrated the same inter-rater reliability as the standard approach in measuring supination and slightly higher reliability in measuring pronation(14). However this method requires continual manipulation by the clinician during measurement, limiting observation of the upper extremity as the therapist is unable to move back from the patient to observe for any compensatory postures.

In this study we tested an alternative design to measure pure forearm rotation which considers the issues raised with exisitng measuring approaches. Unlike the recent method proposed by Szekeres *et al.*(13), this method does not require continual manipulation by the clinician, allowing

the therapist to observe for any compensatory postures. This is achieved by applying consistent pressure to the wrist using a "snap" bracelet with a mounted goniometer watch. A "snap" bracelet is a bi-stable object that can be manipulated into two different configurations(15): it can transit from a straight shape with a groove along its entire width to a coiled shape where the groove disappears.

The GoWatch was designed in-house using a novelty toy watch purchased from a commercial outlet. The strap of the watch was a "snap" bracelet with a silicone cover that also had a circular mount where the small toy watch face was inserted. The GoWatch dial face was produced using a high resolution 3D-printer so that its base could fit into the existing mount.

Previous attempts to design goniometers that attach to the wrist(16, 17) were impractical as they involved the assembly of a large apparatus. Further, the design in this study uses a ball bearing to act as a gravity-assisted dial, improving linear motion by reducing friction between the goniometer components. Therefore this design may theoretically offer further improvements to the recent method described and potentially be more user-friendly.

The purpose of this study was to test the concurrent validity, inter-rater reliability and responsiveness of this new goniometer design in measuring pure forearm rotation with the modified finger goniometer as the criterion instrument for reference.

METHODS

Design

This was a prospective repeated measures study examining the validity, inter-rater reliability and responsiveness of a new goniometer device to measure pure forearm rotation. The Guidelines for Reporting Reliability and Agreement Studies (GRRAS)(18), COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) Risk of Bias checklist were used to ensure quality procedures and reporting(19).

4

Participants

Participants with upper limb trauma or disease that had impeded forearm rotation were consecutively recruited if they met the inclusion criteria. The availability of the chief investigator and both raters was necessary if participation was to occur on the day of recruitment otherwise the participant could be involved at a future therapy session.

Ethics

This study was approved by the Office of Research Ethics Committees, Northern Ireland (ORECNI) and the Medicines and Healthcare products Regulatory Authority (MHRA). MHRA recommended that the goniometer watch was labelled "exclusively for clinical investigations" as it was a new medical device. All participants provided written, informed consent.

Recruitment

Potential participants were identified by surgeons and therapists through the Trust's fracture clinic, physiotherapy and occupational therapy departments. Each eligible patient had a "cooling off" period to consider if they wished to participate in the study. Consent and participation in the study did not occur until participants saw the goniometer watch, gauged its acceptability and read the study information sheet. Alternatively, patients could choose to participate at their next planned hospital review. Nobody who met the study criteria and was invited declined to participate in the study.

Sample Size

Bland and Altman recommend that agreement is necessary to establish if the precision of a new measuring device is acceptable compared to the gold standard. They highlighted this using a sample size of only 17 participants(20). A sample of 30 is commonly employed as a benchmark to rely on the central limit theorem, as statistical research has found that with a sample size of 30

the sampling distribution of the mean is approximately normal(21). Therefore 40 participants

were proposed as an adequate and cost-effective sample size.

Inclusion Criteria:

The study aimed to recruit people aged 18 years and over with a history of upper limb trauma or

disease affecting forearm rotation and that active ROM of the forearm was permitted by the

treating physician.

Exclusion Criteria:

Any patient with an open wound around the distal portion of the wrist, hyperaesthesia or

allodynia to the wrist or a cognitive impairment or learning difficulty that prevented

understanding of verbal instructions or informed consent were excluded from the study. Also

patients were excluded if ROM was contra-indicated due to phase of healing.

Materials

- Modified Finger Goniometer (MFG) (Device 1):

The weighted finger goniometer was placed on the dorsal aspect of the wrist, the long arm

crossing both Lister's tubercle and the ulnar head with the weight over the ulnar side of the wrist.

It is flipped to measure either supination or pronation (Figure 1).

- Goniometer Watch (Device 2):

A 360° goniometer watch (GoWatch) with measurements in 2° increments was used (Figure 1).

Zero degrees are indicated at the bottom of the watch face and incorporates gravity to influence

motion of the ball-bearing. The GoWatch is attached to the radial border of the wrist just

proximal to the ulnar styloid using the snap bracelet mechanism. The strap is attached so the 0°

marking on the watch face is aligned with the centre of the anatomical snuffbox. The arm is

positioned to the side of the trunk in the midline position and elbow flexed to 90°. The forearm

moves either into pronation or supination and a record is made of the degrees reached on the

GoWatch face.

Procedure

A pilot study was conducted with ten participants to familiarise therapists with the process and

address any safety concerns. No safety concerns were highlighted. Some guidance was provided

to both raters by the Chief Investigator on the application of both measuring devices during the

pilot. Two occupational therapists each with over ten years' experience were designated raters.

Two trials of each goniometry method were performed by each therapist before and after a

session of hand therapy. The first trial data were used to compare measures between raters

(inter-rater reliability) and also measures between devices (concurrent validity). The second trial

was used to examine the responsiveness of the devices.

Randomisation and Blinding

A box of 90 cards, indicating which instrument, therapist, and direction of rotation would be

measured first was used. The second therapist followed the same sequence of instrument and

direction. An online random sequence generator (www.random.org) determined the sequence of

cards. After each measure, the rater gave their result verbally to the Chief Investigator who

recorded it. Each rater did not have access to the measurement sheets to assist blinding to the

first set of results. Each rater was also blinded to the other's results during the course of the

study. Participants were not blinded to the results as they would hear the results verbally after

each measure.

Participants were assessed over one session with approximately 30-60 minutes between trials. A

therapy session including various active and passive movements and/or heat therapy occurred

7

during the interval. Participants sat on a plinth with their feet flat on the ground and were instructed to keep their elbow to the side of their trunk. Therapists observed that the participant's shoulders were level, the elbow positioned at 90° and the forearm was in mid-position before movement in either direction occurred. Participants were then instructed to move their forearm arm into maximum pronation or supination.

Statistical analysis

Concurrent Validity and Inter-rater Reliability

Bland Altman plots were populated on Microsoft Excel (2016) to assess agreement between each device (concurrent validity) and each rater (interrater reliability) with *a priori* limits of agreement set at 5° and that the confidence interval (CI) of the mean difference included the line of equality (i.e. no difference in mean scores). A 5 to 10° margin of error has been deemed acceptable for intra and interrater reliability of wrist range of movement (22). The authors deemed a threshold of 5° difference as an acceptable benchmark for interrater reliability. Szekeres *et al.*(14) do not provide details on the level of agreement between the two-arm goniometer and the MFG however provide the mean difference between raters using Bland Altman which reported pronation as 5° [95%CI(-6, 16)] and supination as 3° [95%CI(-9, 14)]. In the absence of any *a priori* values in the literature for concurrent validity, the authors also determined within 5° difference as an acceptable level of agreement.

Concurrent validity is the extent two instruments agree in measuring the same construct. One instrument is typically an established tool used as a reference when assessing the accuracy of a new measurement instrument(23). Inter-rater reliability is the extent to which the measures by two or more raters agree(24). Bland Altman enables visualization of accuracy of these properties.

Bias

Repeated-measures ANOVA using Statistical Package for Social Sciences (v.25 SPSS) measured

the extent to which values recorded the similarity and differences between raters and

instruments. The primary benefit of using repeated-measures ANOVA instead of Intraclass

Correlation Coefficients (ICC) is that it can provide a more precise estimate of agreement(20)

and more detailed information about the difference or relationship between raters and

instruments, separating bias from unexplained residual error. Significant bias (systematic

differences across variables) was set at $p<0.05$. The square root of within-subject mean square

errors was used to calculate unexplained error.

Responsiveness

Responsiveness of each device was assessed by measuring the effect size between trial 1 and 2

using Cohen's d to interpret if the effect size with small, medium or large with (d) equal to 0.2,

0.5 and 0.8 respectively(25). Responsiveness is the extent to which a device can detect change

over time(26). No previous analysis of responsiveness of the MFG has been conducted.

Interrater Reliability (Secondary Analysis)

ICCs were performed as a secondary analysis of inter-rater reliability to compare with key results

reported in the literature. Inter-rater reliability using ICCs (type 2.2) was calculated at 95% CI. A

two-way random effects model with the mean scores ($k = 2$) was used to ascertain absolute

agreement between raters for both techniques(27). An ICC score greater than 0.9 is deemed

excellent while scores less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9 are poor,

moderate and good respectively(28). ICC scores were reported as average measures. Standard

error of measurement and 95% CI (SEM_{95}) were also calculated to determine the precision of

each device and Shapiro-Wilk test ($p>0.05$) was used to ascertain normality of data sets used.

RESULTS

9

From November 2019 until December 2022 40 participants were recruited (Table 1). One upper extremity was assessed for each participant and all had restrictions in forearm rotation due to injury with most participants being female. Descriptive statistics on both measures are provided in Table 2.

Concurrent Validity

The mean difference between the MFG and GoWatch values for supination and pronation in trial 1 followed a normal distribution as determined by the Shapiro-Wilk test(p>0.05). Bland Altman plots illustrate the mean difference between both instruments for supination was 1.19º [95%CI(-1.77, 4.15)] and for pronation was 0.20º [95%CI(-3.10, 2.70)] both of which were within the 95% CI and the 5º acceptable range indicating agreement (Figure 2).

Inter-rater Reliability

Agreement between raters was also assessed using Bland Altman plots. Normality of the mean difference scores between raters was determined using the Shapiro-Wilk tests however values for the mean difference in MFG pronation values were non-normal (p<0.05) and Log10 transformations could not be performed due to zero and negative values.

The mean difference between rater 1 and 2 measuring supination with the GoWatch was 4.43º [95%CI(.52, 8.33)] and for pronation was 2.23º [95%CI(-2.22, 6.65)] whereby the line of equality fell within the 95% CI and also within the 5º acceptable range indicating agreement. The mean difference between rater 1 and 2 measuring supination with the MFG was -1.30º (CI95: -5.01, 2.41) whereby the line of equality fell within the 95% CI, also within the 5º acceptable range and indicating agreement (Figure 3).

Secondary analysis using ICCs was then performed. MFG data sets did not show linearity along the QQ plot line which was confirmed by the Shapiro-Wilks test (p<0.05). GoWatch supination in trial 1 was 0.94 [95%CI(.86-.97)] and pronation was 0.85 [95%CI(.71-.92)] indicating excellent

10

and good interrater reliability respectively. The SEM_{95} for the GoWatch in trial 1 was 6.44º for supination and 7.33º for pronation, both outside the acceptable limits (<5º).

Systematic Bias

Repeated-measures ANOVA data from trial 1 detected significant systematic bias between all supination and pronation measures (p<0.01) with higher pronation values and a residual error of 36.21º (Mean square error = 1311.08).

Figure 4 illustrates a significant rater-by-device interaction with systematic bias between rater 1 and 2 using the GoWatch (Device 2) when analysing average measures of combined pronation and supination in trial 1 (p=0.02) with a residual error of 5.95º (Mean square error = 35.37). Rater 1 and Rater 2 closely agreed on MFG (Device 1) and GoWatch measures (interaction between error bars) however systematic bias was observed with Rater 2 consistently underestimating measures with the GoWatch compared to Rater 1. This is represented by the diagonal line between rater 1 and rater 2.

Repeated-measures ANOVA agreement between each rater, each device and each component of ROM (supination and pronation) in trial 1, detected no significant bias (p=0.18) however the residual error of 8.26º (Mean square error = 68.28) was outside the acceptable *apriori* level of 5º.

Responsiveness

Normality testing of all rater 1 MFG measures and rater 2 MFG supination measures were significant using the Shapiro-Wilk test (p<0.05) however both raters' GoWatch measures followed a normal distribution and therefore were analysed for responsiveness using Cohen's *d*. Using measures collated by rater 1 a very small effect size was observed between trial 1 and 2 GoWatch supination (*d*=0.14 95% CI [-.30, .58]) and a small to medium effect for pronation (*d*=0.29 95% CI [-.17, .72]). For rater 2 a small to medium effect size was observed between trial 1 and 2 GoWatch supination (*d*=0.29 95% CI [-.16: .72]) and a very small effect for pronation

11

(*d*= 0.05 95% CI [-.38, .49]). Repeated-measures ANOVA also confirm that there was a positive effect between trial 1 and 2 when analysing average combined measures (i.e. pronation+supination) ($p<0.01$) with a residual error of 7.66° (Mean square error = 58.74).

DISCUSSION

The GoWatch demonstrated acceptable concurrent validity and interrater reliability for both supination and pronation measures (<5°) however systematic bias was found when the rater by instrument interaction was analysed. The MFG demonstrated superior agreement between raters compared to the GoWatch. Results from the study by Szekeres *et al.*(14) further demonstrates superior agreement between raters using the MFG than the results obtained on the GoWatch in this study.

Bland Altman plots (Figure 3) illustrate that the difference between raters measuring supination using the GoWatch increased as higher ROM measures were obtained, while differences outside the *apriori* limits of agreement for GoWatch pronation and MFG supination were evenly spread. Rater 2's total measures (supination+pronation) produced on average significantly lower GoWatch measurements compared to their MFG measures in trial 1 (Figure 4). GoWatch responsiveness in detecting change between trial 1 and 2 was significant using both statistical methodologies i.e. repeated-measures ANOVA and Cohen's *d*. Each rater detected some change using the GoWatch though this varied from a very small to small / medium effect. This treatment effect between trials was hypothesized as it was intended to illicit an improvement in ROM. However the ability for the GoWatch to measure responsiveness should be considered with caution due to the systematic bias detected in this study and also that the procedure was unable to completely "blind" both raters and participants from the results. Further mean difference in scores between trial 1 and 2 for both raters measuring supination and pronation are low relative to the acceptable level of agreement (<5°) except for rater 2 measuring supination which had a mean difference of 6.45°. Further Armstrong *et al*(8) have recommended that a

12

meaningful change in rotation measures would be at least 10°. Future studies on the GoWatch

should consider measuring responsiveness over a longer duration to determine if it can detect

meaningful clinical change.

Secondary analysis of GoWatch scores using ICCs, demonstrated good interrater reliability for

measuring pronation and excellent interrater reliability for measuring supination. However this

criteria can be misleading as the use of correlation coefficients to test agreement between

measurement instruments and raters is cautioned(20). ICCs imply the GoWatch is fit for purpose

in a clinical setting however fail to identify systematic bias of the GoWatch measures. GoWatch

ICCs imply excellent interrater reliability for measuring supination and good interrater reliability

for measuring pronation yet Bland Altman plots indicate mean difference was smaller for

pronation. This disparity between ICCs and mean difference have previously been

demonstrated(29). These authors advise that ICCs "do not always reflect the clinical implications

of measurement errors" and recommend the utilization of other statistical methods such as those

used in this study. The results further highlight the risk of basing assumptions on clinical

acceptability of a measuring instrument solely on ICCs. Szekeres *et al.*(14) reported that the MFG

had excellent inter-rater reliability for pronation (ICC: 0.86) and supination (ICC: 0.95)(14),

similar to results for the GoWatch, however no systematic bias was identified with the MFG

using linear regression analysis. SEM_{95} scores for GoWatch pronation was 7.33° and for

supination was 6.44°, both outside the acceptable *apriori* limits and a larger error than previously

reported for the MFG (SEM_{95} pronation = 2.1° and supination = 1.2°). SEM_{95} will decrease as

sample size increases and Szekeres *et al.*(14) measured ICCs and SEM_{95} using pooled samples of

60 participants over six sessions producing a sample size of 360 possibly explaining smaller

SEM_{95} values. With systematic bias SEM_{95} values may not reduce much further regardless of a

larger sample size. Results were not pooled from both trials because the second trial measured

effect size and therefore pooling to assess agreement and reliability would be a flaw in the

statistical approach. Non-normality of MFG data meant comparisons with GoWatch responsiveness could not be made.

MFG supination measures had non-normal platykurtik kurtosis meaning that the values were uniform with few outliers. Rater 1 pronation measures had non-normal leptokurtic kurtosis meaning that there were frequent outliers. Similar trends were observed in the GoWatch measures and MFG rater 2 pronation measures but kurtosis fell within acceptable limits (±1.0). Systematic bias was also observed between all measures of supination and pronation with a residual error of 36.21°. This was hypothesized as most people have limitations in supination rather than pronation after a distal radius fracture(3) and this phenomenon was observed in this cohort (82.5% of study participants). This may also explain the negative skewness of the pronation measures which were observed in both raters though with rater 2's measures just within normal distribution.

Systematic bias between raters may be an isolated event, due to weaknesses in blinding or the potentially arbitrary reference point for placing the GoWatch. The wide area of the anatomical snuffbox may produce variability in placement and therefore in measurement readings. In larger or oedematous limbs, the anatomical snuffbox may be less defined and further affect accuracy. Participants were on average assessed in this study at around three months from injury or surgery though details on oedema or visibility of the anatomical snuffbox was not collated for this study.

Several suggestions on future iterations may help reduce bias. Adjusting the position of the watch face whereby it is positioned on the dorsum of the wrist where the flatter and wider surface area may offer a more stable base for the watch and limit unwanted movement. Aligning the centre of the watch face with the third metacarpal base by a reference line on the snap bracelet could further improve precision, and a serrated edge on the interior of the watch face may minimise oscillating movements when end ROM is held.

14

Both raters in this study were experienced in hand therapy and were inducted via a pilot study on both methods which were novel to them. Both raters did not find the MFG intuitive, with the Chief Investigator interjecting to provide instruction to flip the goniometer to enable measurements in opposing directions. Formal feedback on each rater's preference of instrument was however beyond the scope of this study.

Limitations included systematic bias of rater-instrument interactions restricting a thorough understanding of the reliability, validity and responsiveness of the GoWatch. Using only two raters limited generalisation of the results to hand therapy practitioners. Some datasets were not normally distributed and as such some important aspects of analysis could not be completed. The study commenced in November 2019 and delays in recruitment were encountered due to COVID-19.

Systematic bias may be resolved by further research with more than two raters and provision of more extensive training on each instrument. Future study procedures should also consider strategies to ensure both participants and raters are completely blind to measurements. Further iterations to the GoWatch may also improve precision and reduce systematic bias. The hinge at the base of the GoWatch allows it to fold 90° degrees parallel to the forearm and it can also rotate 360° in its fulcrum. Subsequently the GoWatch may be used to measure other joint ROM (e.g. elbow flexion).The raters in this study found the GoWatch more intuitive to use than the MFG however this is not particularly reflected in the results. Further, exploration of clinician perceptions would be an important area of research. This study highlights the need for caution when interpreting ICC scores of measuring instruments. Using the categories defined by Shrout and Fleiss(28) can be misleading when choosing an instrument for clinical use.

The GoWatch is easy to use and further iterations may help improve its precision so that it becomes an instrument of choice for many clinicians. Thereafter, further research will be necessary to evaluate its measurement properties and to ascertain if systematic bias is endemic across multiple raters.

Declaration of Conflicting Interests

The Authors declare that there is no conflict of interest

REFERENCES

1.   Soubeyrand M, Assabah B, Bégin M, Laemmel E, Dos Santos A, Crézé M. Pronation and supination of the hand: Anatomy and biomechanics. *Hand Surg Rehabil.* 2017;36(1):2-11.

2.   Kleinman WB. Distal radius instability and stiffness: common complications of distal radius fractures. *Hand Clin.* 2010;26(2):245-64.

3.   Naughton N, Algar L. Therapy Management of Distal Radius Fractures. In: Skirven TM, Osterman AL, Fedorczyk JM, Amadio PC, Feldscher SB, Shin EK, editors. *Rehabilitation of the Hand and Upper Extremity* 7th ed. Elsevier Health Sciences: Philidelphia, 2021. p. 833-850

4.   MacIntyre NJ, Dewan N. Epidemiology of distal radius fractures and factors predicting risk and prognosis. *J Hand Ther.* 2016;29(2):136-45.

5.   Yang W, Houtrow A, Cull DS, et al. Quality and Outcome Measures for Medical Rehabilitation. In: Cifu DX, editor. *Braddom's Physical Medicine and Rehabilitation* 6th ed. Elsevier: Philidephia 2021. p. 100-14.e2.

6.   World Health Organisation. International classification of functioning, disability, and health, https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health (2018, accessed 12 October 2023)

16

7.   MacDermid J, Solomon G, Valdes K. *Clinical Assessment Recommendations* 3ʳᵈ ed. Amercian

Society of Hand Therapists: Mount Laurel, NJ, 2015

8.   Armstrong AD, MacDermid JC, Chinchalkar S, Stevens RS, King GJ. Reliability of

range-of-motion measurement in the elbow and forearm. *J Shoulder Elbow Surg.* 1998;7(6):573-80.

9.   McRae R. *Clinical Orthopaedic Examina*tion. 1st ed. Churchill Livingstone: Edinburgh;

1981.

10.   Flowers KR, Stephens-Chisar J, LaStayo P, Galante BL. Intrarater reliability of a new

method and instrumentation for measuring passive supination and pronation: a preliminary

study. *J Hand Ther.* 2001;14(1):30-5.

11.   McGarry G, Gardner E, Muirhead A. Measurement of forearm rotation: an evaluation of

two techniques. *J Hand Surg Br.* 1988;13(3):288-90.

12.   Karagiannopoulos C, Sitler M, Michlovitz S. Reliability of 2 functional goniometric

methods for measuring forearm pronation and supination active range of motion. *J Orthop Sports

Phys Ther.* 2003;33(9):523-31.

13.   Szekeres M, MacDermid JC, Rooney J. A new method for measuring forearm rotation

using a modified finger goniometer. *J Hand Ther.* 2015;28(4):429-31; quiz 32.

14.   Szekeres M, MacDermid JC, Birmingham T, Grewal R. The inter-rater reliability of the

modified finger goniometer for measuring forearm rotation. *J Hand Ther.* 2016;29(3):292-8.

15.   Guo Q, Zheng H, Chen W, Chen Z. Modeling bistable behaviors in morphing structures

through finite element simulations. *Biomed Mater Eng.* 2014;24(1):557-62.

16.   Wagner C. Determination of the rotary flexibility of the elbow joint. E*ur J Appl Physiol

Occup Physiol.* 1977;37(1):47-59.

17.   Darcus HD, Salter N. The amplitude of pronation and supination with the elbow flexed

to a right angle. *J Anat.* 1953;87(2):169-84.

18.	Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al.

Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol. 2*011;64(1):96-106.

19.	Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Medical Research Methodology.* 2020;20(1):293.

20.	Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet.* 1986;1(8476):307-10.

21.	Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. *Korean J Anesthesiol.* 2017;70(2):144-56.

22.	Marx RG, Bombardier C, Wright JG. What do we know about the reliability and validity of physical examination tests used to examine the upper extremity? J *Hand Surg Am.* 1999;24(1):185-93.

23.	Field A. *An Adventure in Statistics: The Reality Enigma.* 2nd ed. Sage Publications Ltd: London, 2022. p 61

24.	Lange RT. Inter-rater Reliability. In: Kreutzer JS, DeLuca J, Caplan B, editors. *Encyclopedia of Clinical Neuropsychology.* Springer: New York, 2011. p. 1348.

25.	Cohen J. *Statistical Power Analysis for the Behavioral Sciences* 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates, 1988.

26.	Terwee CB. Responsiveness to Change. In: Michalos AC, editor. *Encyclopedia of Quality of Life and Well-Being Research.* Springer: Netherlands, 2014. p. 5547-5550.

27.	Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med.* 2016;15(2):155-63.

28.	Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin.* 1979;86(2):420.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

29.   Lee KM, Lee J, Chung CY, Ahn S, Sung KH, Kim TW, et al. Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. *Clin Orthop Surg.* 2012;4(2):149-55.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 1: Measuring pronation (top left) and supination (top right) using the goniometer watch (GoWatch) and measuring pronation (bottom left) and supination (bottom right) using the modified finger goniometer (MFG)
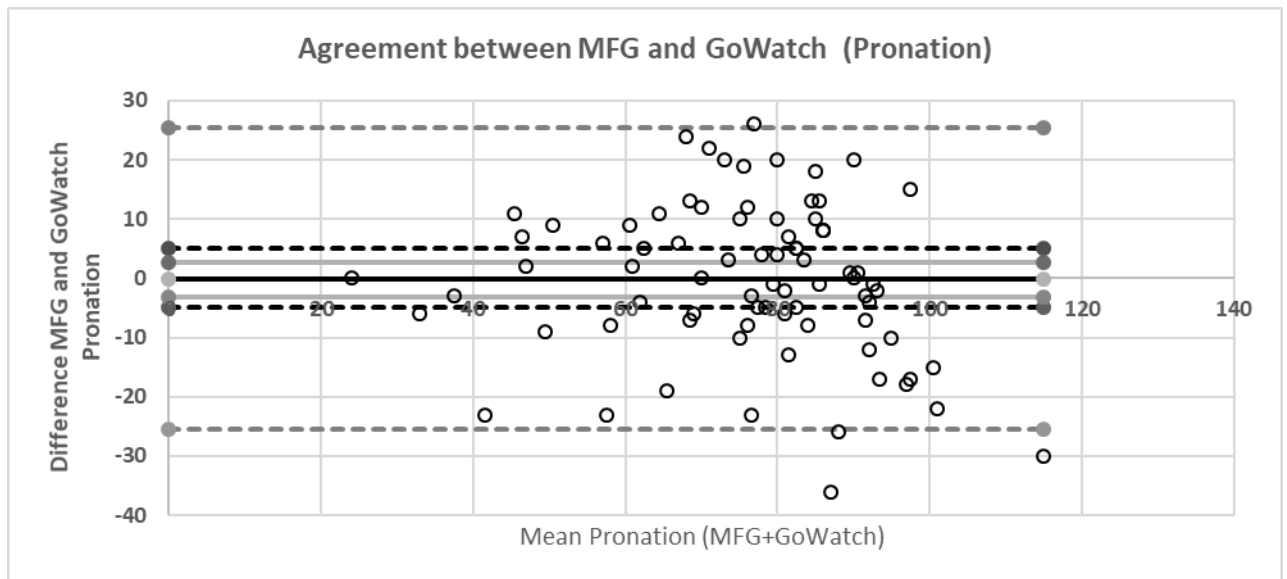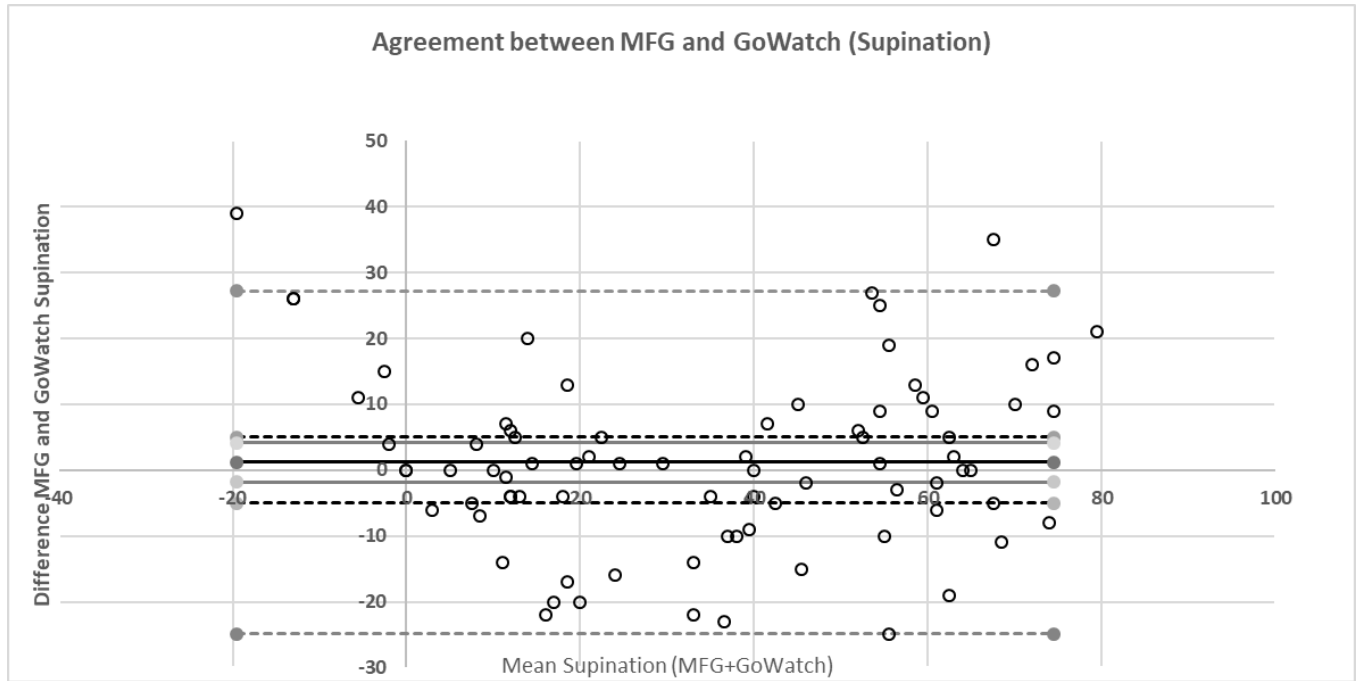
Figure 2: Agreement between MFG and GoWatch in measuring pronation and supination. Black central

solid line = mean difference; two solid grey lines = 95% CI; black dotted lines = *apriori* Limits of

Agreement (LOA); Outside grey dotted lines = LOA based on mean of the two values, minus and plus

1.96 standard deviations

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Agreement Rater 1 and Rater 2 (Gowatch Supination)

Agreement Rater 1 and Rater 2 (GoWatch Pronation)

Agreement Rater 1 and Rater 2 (MFG Supination)
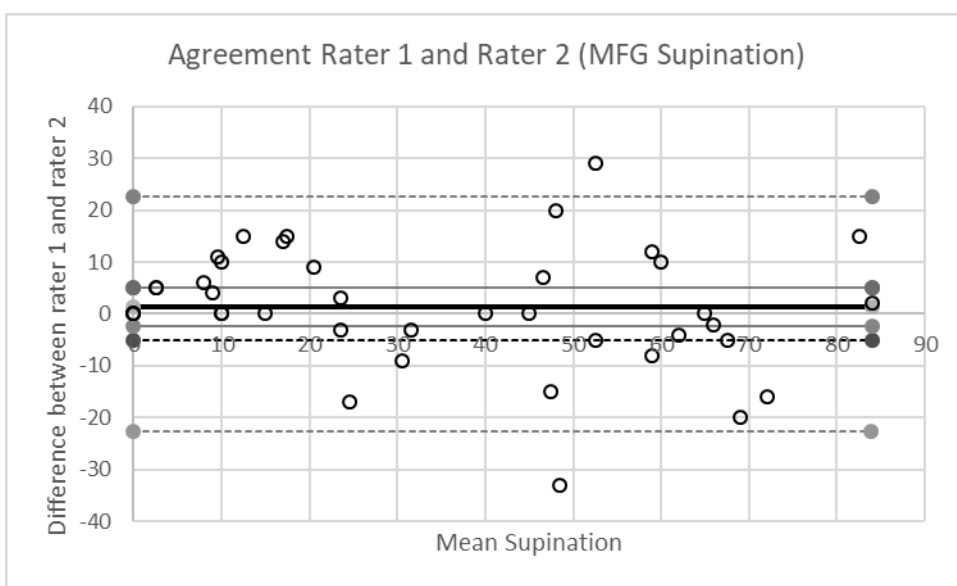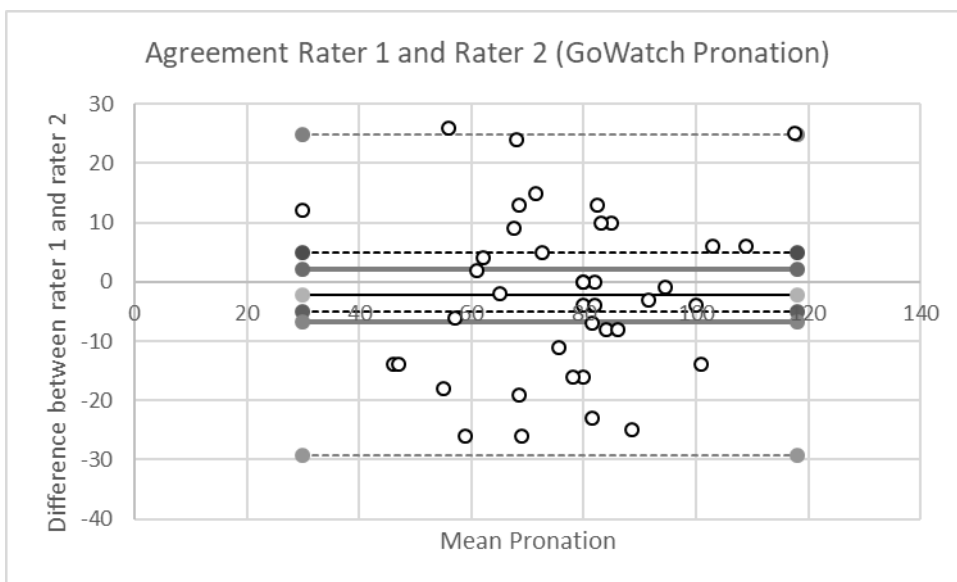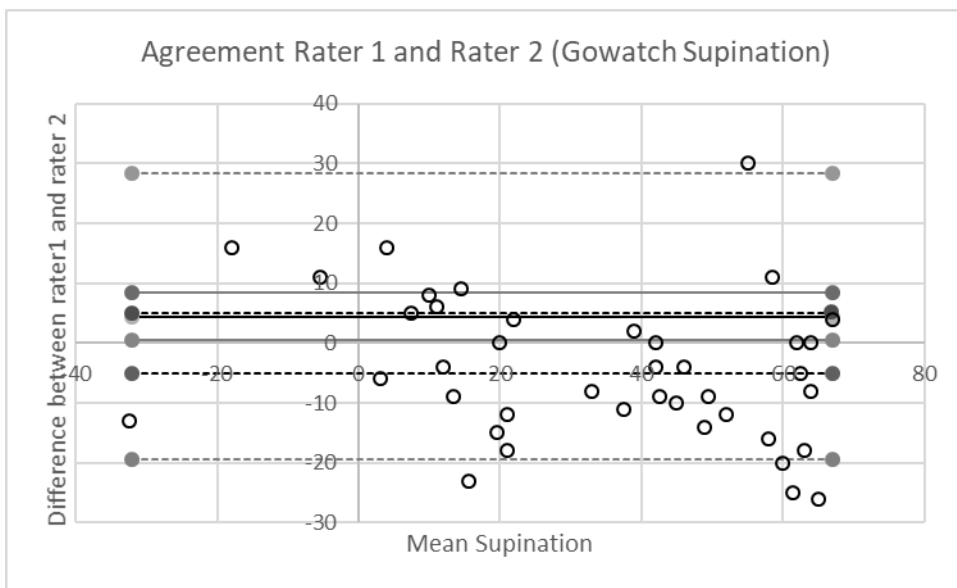
Figure 3: Agreement between Rater 1 and 2 in measuring supination and pronation using the GoWatch and supination using the MFG. Black central solid line = mean difference; two solid grey lines = 95% CI; black dotted lines = *apriori* Limits of Agreement (LOA); Outside grey dotted lines = LOA based on mean of the two values, minus and plus 1.96 standard deviations

Note: MFG supination upper *apriori* LOA black dotted line not visible as overlaps with upper 95% CI (solid grey line)
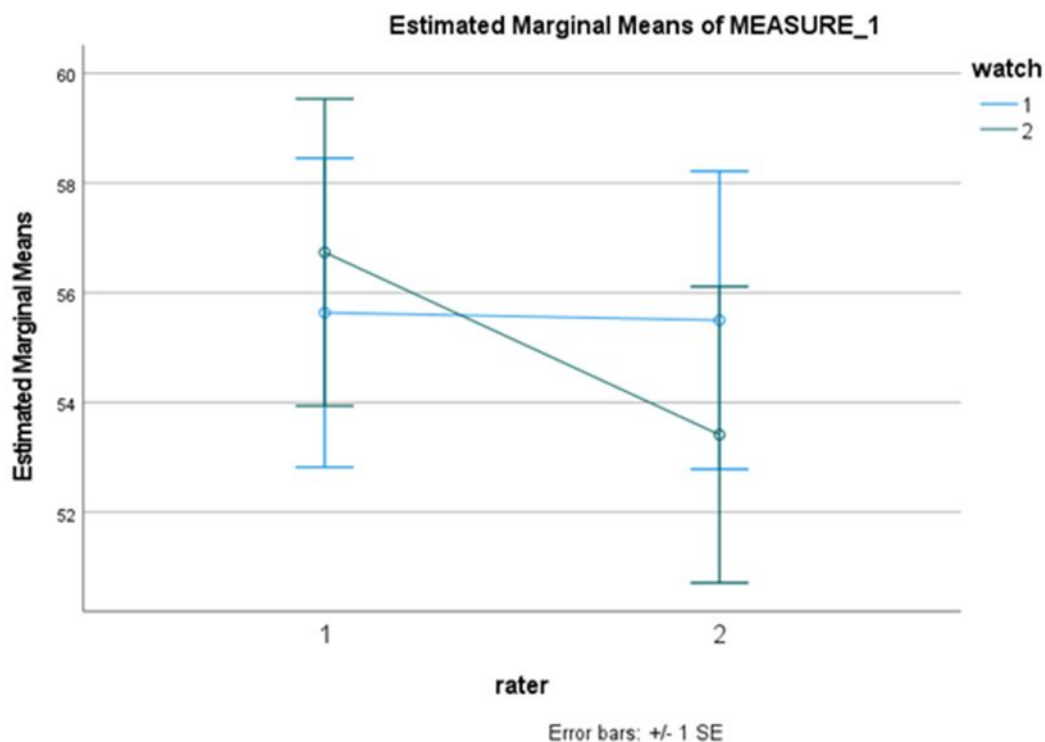
Figure 4: Estimated marginal mean for pooled total ROM (°) as measured by rater 1 and rater 2 in trial

1using both the MFG and GoWatch. Watch 1 on x axis refers to MFG and Watch 2 refers to GoWatch.

Overlapping vertical lines illustrate agreement between devices measured by each rater. Horizontal lines

illustrate agreement between raters using each device. Error bars represent standard error. Total ROM

refers to total arc of forearm rotation (pronation + supination)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1: Demographic characteristics of participants (n=40)

Characteristics			
	Mean	**SD**	**Range**
Age	53.6	13.58	20-78
Time from Injury/ Surgery* (Days)	98.79	59.78	24-344
	n	**(%)**	
Female	30	75	
Affected Side			
Left	22	55	
Right	18	45	
Injury Type			
Humeral fracture	3	7.5	
Distal radius fracture	31	77.5	
Distal radius and ulna fracture	2	5	
Radial Head fracture	2	5	
Elbow dislocation	1	2.5	
Triad injury	1	2.5	

n = number of participants, SD = standard deviation

*Days counted from surgery rather than injury in cases where surgery was performed

Table 2. Descriptive statistics of GoWatch and MFG measures. Trial 1 measures carried out prior to therapy. Trial 2 measures

carried out after 30-60 minute interval of hand therapy

GoWatch					MFG		
		n	Mean (SD)	Range	n	Mean (SD)	Range
	Trial 1						
Rater 1	Pronation	40	77.35(18.40)	24-108	40	76.83 (17.44)	24-100
	Supination	40	36.13 (27.81)	-26-78	40	34.45 (27.44)	0-83
Rater 2	Pronation	40	75.13 (19.61)	36-130	40	75.25 (16.83)	30-105
	Supination	40	31.70 (24.67)	-39-70	40	35.75 (25.30)	0-90
	Trial 2						
Rater 1	Pronation	40	82.25 (16.93)	50-119	40	78.08 (15.69)	37-100
	Supination	40	40.25 (27.00)	-21-86	39	37.74 (25.66)	0-84
Rater 2	Pronation	40	76.15 (16.45)	34-110	40	79.88 (17.45)	35-120
	Supination	40	38.15 (23.94)	-26-82	40	39.85 (26.67)	0-100

Note: Trial 1 measures carried out prior to therapy. Trial 2 measures carried out after 30-60 minute interval of hand therapy.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

**COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument**

Table 1. Elements that make up a comprehensive reached question.

	Element of the research question
1	the **name** of the outcome measurement instrument
2	the **version** of the outcome measurement instrument or way of **operationalization** of the measurement protocol
3	the **construct** measured by the measurement instrument
4	a specification whether one is interested in a **reliability parameter** (i.e. a relative parameter such as an ICC, Generalizability coefficient φ, or Kappa κ) or a **parameter of measurement error** (i.e. an absolute parameter expressed in the unit of measurement e.g. SEM, LoA or SDC; or expressed as agreement or misclassification, e.g. the percentage specific agreement).
5	a specification of the **components of the measurement instrument** that will be **repeated** (especially when only part of the measurement instrument is repeated, e.g. only assignment of the score based on the same images)
6	a specification of the **source(s) of variation** that will be **varied** (e.g. time or occasion, the (level of expertise of) professionals, the machines, or other components of the measurement)
7	a specification of the **patient population** studied

ICC = Intraclass correlation coefficient; SEM = standard error of measurement; LoA = Limits of Agreement; SDC = smallest detectable change.

Responses to above from "Main Document":
1. Refer to main manuscript: GoWatch
2. Page 5 Line 109 and Page 6 Line 146
3. Page 4 Line 95
4. Page 7 - 8
5. Page 7 "Procedure"
6. Page 7 "Procedure"
7. Page 4-6

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Standards for studies on reliability

	Design requirements	very good	adequate	doubtful	inadequate	NA
1	Were patients stable in the time between the repeated measurements on the construct to be measured?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	Na
2	Was the time interval between the repeated measurements appropriate?	Yes		Doubtful, OR time interval not stated	No	Na
3	Were the measurement conditions similar for the repeated measurements – except for the condition being evaluated as a source of variation?	Yes (evidence provided)	Reasons to assume standard was met, OR change was unavoidable	Unclear	No (evidence provided)	Na
4	Did the professional(s) administer the measurement without knowledge of scores or values of other repeated measurement(s) in the same patients?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
5	Did the professional(s) assign scores or determine values without knowledge of the scores or values of other repeated measurement(s) in the same patients?	Yes (evidence provided)	Reasons to assume standard was met	Unclear	No (evidence provided)	
6	Were there any other important flaws in the design or statistical methods of the study?	No		Minor methodological flaws	Yes	

1. Yes – the manuscript describes randomisation of order of assessment between two raters and also set up procedure for measurement (very good)
2. To assess responsiveness of the device the time of up to one hour was appropriate for initial analysis however assessment of responsiveness over several treatment sessions would have been more insightful. However the responsiveness of the device was not the primary question of this study as initial focus was in interrater reliability and agreement with the gold standard. (very good)
3. Yes – above (very good)
4. Yes – Page 7 Line 165 (doubtful)
5. Yes – Page 7 Line 165 (doubtful)
6. Some datasets relating to the modified finger goniometer (reference outcome measure) were non-normal therefore some analysis relating to ICCs and responsiveness could not be carried out however this was not the instrument of concern in this study (adequate)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

GRRAS checklist for reporting of studies of reliability and agreement

Version based on Table I in: Kottner J, Audigé L, Brorson S, Donner A, Gajeweski BJ, Hróbjartsson A, Robersts C, Shoukri M, Streiner DL. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106

Section	Item #	Checklist item	Reported on page #
Title/Abstract	1	Identify in title or abstract that interrater/intrarater reliability or agreement was investigated.	Page 1
Introduction	2	Name and describe the diagnostic or measurement device of interest explicitly.	Page 6
	3	Specify the subject population of interest.	Page 4
	4	Specify the rater population of interest (if applicable).	Page 6
	5	Describe what is already known about reliability and agreement and provide a rationale for the study (if applicable).	Page 5
Methods	6	Explain how the sample size was chosen. State the determined number of raters, subjects/objects, and replicate observations.	Page 5 and 6
	7	Describe the sampling method.	
	8	Describe the measurement/rating process (e.g. time interval between repeated measurements, availability of clinical information, blinding).	Page 7 and 8
	9	State whether measurements/ratings were conducted independently.	Page 7
	10	Describe the statistical analysis.	Page 7-10
Results	11	State the actual number of raters and subjects/objects which were included and the number of replicate observations which were conducted.	Page 6, Page 8
	12	Describe the sample characteristics of raters and subjects (e.g. training, experience).	Page 5 and 6
	13	Report estimates of reliability and agreement including measures of statistical uncertainty.	Page 7-10
Discussion	14	Discuss the practical relevance of results.	Page 10-13
Auxiliary material	15	Provide detailed results if possible (e.g. online).	See Figures and Tables