Accepted Manuscript

Psychometric evaluation of the Brachial Assessment Tool Part 1: Reproducibility

Bridget Hill Grad, Dip (Physio), Gavin Williams, PhD, John Olver, MBBS, Scott Ferris, MBBS, Andrea Bialocerkowski, PhD

PII: S0003-9993(17)31339-4

DOI: [10.1016/j.apmr.2017.10.015](https://doi.org/10.1016/j.apmr.2017.10.015)

Reference: YAPMR 57066

To appear in: *ARCHIVES OF PHYSICAL MEDICINE AND REHABILITATION*

Received Date: 28 November 2016

Revised Date: 22 August 2017

Accepted Date: 18 October 2017

Please cite this article as: Grad BH, Williams G, Olver J, Ferris S, Bialocerkowski A, Psychometric evaluation of the Brachial Assessment Tool Part 1: Reproducibility, *ARCHIVES OF PHYSICAL MEDICINE AND REHABILITATION* (2017), doi: 10.1016/j.apmr.2017.10.015.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Title Page

Running head: Reproducibility of the BrAT

Title: Psychometric evaluation of the Brachial Assessment Tool Part 1: Reproducibility

Authors

Bridget Hill Grad Dip (Physio)

Menzies Health Institute, Queensland, Australia

Epworth Monash Rehabilitation Medicine Unit Epworth HealthCare, Melbourne, Victoria,
Australia

Gavin Williams, PhD

Epworth Monash Rehabilitation Medicine Unit Epworth HealthCare, Melbourne, Victoria,
Australia

John Olver MBBS

Epworth Monash Rehabilitation Medicine Unit Epworth HealthCare, Melbourne, Victoria,
Australia

Scott Ferris MBBS

The Alfred, Melbourne, Australia

Andrea Bialocerkowski PhD

Menzies Health Institute, Queensland, Australia

Acknowledgements

The authors wish to acknowledge the assistance of Jaslyn Gibson, Occupational Therapist,
Perth, Australia; Mr. David McCombe, Victorian Hand Surgery Associates and St Vincent's

Hospital, Melbourne, Australia and Melanie McCulloch, Re-wired Hand Therapy, Melbourne, Australia for their valuable contribution to this project during the recruitment and data collection phase.

No financial support was provided for this project

There are no conflicts of interest

Corresponding author

Bridget Hill, Grad Dip (Physio)

Griffith Health Institute, Griffith University, Gold Coast, Queensland, Australia

Epworth Monash Rehabilitation Medicine Unit Epworth HealthCare, Melbourne, Victoria, Australia

89 Bridge Rd Richmond Vic 3122 Australia

Business Tel 61+ (03) 9426 8785

Email: bridget.hill@epworth.org.au

Reprints are not available

Psychometric evaluation of the Brachial Assessment Tool. Part 1: Reproducibility

Abstract

Objective: To evaluate reproducibility (reliability and agreement) of the Brachial Assessment Tool (BrAT) a new patient-reported outcome measure for adults with traumatic Brachial Plexus Injury (BPI)

Design: Prospective repeated measure design

Setting: Outpatient clinics

Participants: Adults with confirmed traumatic BPI

Intervention: 43 people (age range 19-82) with BPI completed the 31-item 4-response BrAT twice, 2 weeks apart. Results for the 3 subscales and summed score were compared at time 1 and time 2 to determine reliability including systematic differences using paired t tests; test retest using Intraclass Correlation Coefficient (ICC 1,1) and internal consistency using Cronbach alpha. Agreement parameters included standard error of measurement, minimal detectable change and limits of agreement.

Main outcome measure: The BrAT

Results: Test retest reliability was excellent (ICC [1,1] = 0.90 – 0.97). Internal consistency was high (Cronbach alpha 0.90 - 0.98). Measurement error was relatively low (SEM range

3.1 - 8.8). A change of >4 for subscale 1, >6 for subscale 2, >4 for subscale 3 and >10 for the summed score is indicative of change over and above measurement error. Limits of agreement ranged from ± 4.4 (subscale 3) to 11.61 (summed score).

Conclusion: These findings support the use of the BrAT as a reproducible patient reported outcome measure for adults with traumatic BPI with evidence of appropriate reliability and agreement for both individual and group comparisons. Further psychometric testing is required to establish the construct validity and responsiveness of the BrAT.

Key words: Brachial Plexus Injury; Outcome Assessment (Health Care), Upper Extremity, Reproducibility of results.

List of abbreviations

BrAT – Brachial Assessment Tool

BPI - Brachial Plexus Injury

DASH - Disabilities of the Arm Shoulder and Hand

CI - Confidence interval

SD - Standard deviation

LoA – Limits of agreement

MDC – Minimal Detectable change

SEM – Standard Error of Measurement

ICC - Intraclass Correlation Coefficient

Traumatic Brachial Plexus Injury (BPI) is a serious condition that generally affects previously healthy younger people.⁽¹⁾ People with BPI present with an extremely wide range of ability to use their arm based on the site and severity of the initial injury. They may undergo many months if not years of expensive and time-consuming surgery and ongoing therapy to reanimate their arm with varying degrees of success.⁽²⁻⁵⁾ Historically outcome assessment following BPI has been primarily impairment based.⁽⁶⁻⁸⁾ Day-to- use of the affected limb has not been routinely assessed despite this being key to the long-term outcome and overall satisfaction for the person with BPI.⁽⁹⁻¹²⁾ Where activity has been assessed the measures have not been psychometrically evaluated for BPI.⁽⁷⁾ The most commonly used patient-reported outcome measure is the Disabilities of the Arm Shoulder and Hand (DASH).^(6, 7) However the DASH has been shown to be multidimensional so scores must be viewed with caution, not to contain items that truly reflect how people with BPI use their affected limb⁽¹³⁾ and are likely to address compensation or adaption rather than actual use of the affected limb.⁽¹⁴⁾

The Brachial Assessment Tool (BrAT) is a new unidimensional 31-item four-response patient-reported outcome measure designed to address some of these issues. Based on the International Classification of Functioning, Disability and Health definition of activity ‘*execution of a task or action by the individual*,’⁽¹⁵⁾ items for inclusion were generated by experts in the field including people with BPI.⁽¹³⁾ Developed using Rasch analysis the BrAT is a unidimensional measure assessing solely ‘activity following adult traumatic BPI’.⁽¹⁶⁾ To assess actual day-to-day use of the arm responses are attributed directly to the affected limb. The BrAT may be used as three separate subscales: i) eight ‘*Dressing and grooming*’ items, ii) 17 ‘*Whole arm and hand*’ items, iii) six ‘*No hand*’ items; or alternatively all 31-items may

be added to produce a summed score. BrAT item responses range from '0 - *cannot do now*', '1 - *very hard to do now*', '2 - *a little hard to do now*', to '3 - *easy to do now*'.

Recovery from BPI occurs over a prolonged period of time and has a significant impact on a person's psychological and emotional state.^(9, 11, 12) Further, people with a BPI often report ongoing severe pain.^(17, 18) These variables may influence how the person with a BPI perceives the day-to-day use of their affected limb and be a source of random error that may affect the reliability of the BrAT.⁽¹⁹⁾ The BrAT was designed using Rasch analysis and has appropriate evidence supporting content validity and unidimensionality, i.e. all the items appear to be measuring the same underlying construct.⁽¹⁶⁾ To further support the use of the BrAT for adults with BPI in the clinical setting and to aid in the interpretation of BrAT scores, evidence of additional psychometric properties is required. All outcome measures must be reproducible, i.e. people who are stable will obtain similar results from repeated assessment.⁽²⁰⁾ Reproducibility is fundamental to all aspects of measurement and proof of reproducibility can ensure confidence in the data from which rational conclusions can be drawn.⁽²¹⁾

Reproducibility is comprised of two different but essential components; reliability and agreement.^(20, 22, 23) Reliability addresses how stable a measure is over repeated use and how well people can be differentiated despite measurement error.^(21, 24, 25) Measures of reliability include test re-test and intra-rater reliability, defined as 'the degree to which one rater can obtain the same rating on multiple occasions of measuring the same variable.'⁽²¹⁾ Internal consistency indicates how interconnected the items are, i.e. all the items appear to be related to each other and measuring something similar.^(21, 26) Agreement is related to absolute measurement error, i.e. how close repeated measure scores are, expressed in the actual units

of the measure. In essence, reliability coefficients enable discrimination of people, while agreement addresses how scores differ.

Aims and Hypotheses

The purpose of this paper was to investigate the two parameters of reliability, (test retest reliability and internal consistency), and three parameters of agreement, SEM, MDC and Bland Altman Limits of Agreement (LoA), *A priori* hypotheses were established based on the COSMIN guidelines. We expected that (1) the BrAT will demonstrate high test retest reliability with an Intraclass Correlation Coefficient of >0.8 , (2) the BrAT will demonstrate high internal consistency with a Cronbach alpha of ≥ 0.7 and, (3) 95% of the Bland Altman limits of agreement scores (LoA) will fall within two standard deviations above and below the mean difference score.

Method

This project employed a multicentre, prospective repeated measure design. Ethical approval was gained from three Human Research and Ethics Committees, (Griffith University PES_12_13_HREC, Alfred Health 425/11, Melbourne Health 2011.220) and all participants provided signed informed consent prior to commencement of the project.

Participants

Participants comprised a convenience sample recruited from the 106 people with BPI who participated in the Rasch analysis arm of a previously reported study. Data were collected

concurrently.⁽¹⁶⁾ Participants were recruited to the reproducibility arm if they had a diagnosis of traumatic BPI confirmed by MRI, nerve conduction studies, intraoperative findings or clinical assessment, and were over 18 years of age at the time of recruitment. To ensure participants to this arm of the project remained stable during the assessment period only those greater than 12 weeks post injury, and who had not undergone surgery to reanimate the upper limb within the previous two years were invited to take part. Thus the function of their arm was likely to remain stable for the duration of this project as minimal recovery may be expected. Exclusion criteria included inability to provide informed consent, pre-existing upper limb conditions that affected day-to-day activity, evidence of spinal cord injury confirmed by MRI, or a diagnosis of brachial plexus birth injury.⁽¹⁶⁾

Data collection

Once participants consented to participate, they were posted a copy of the questionnaire used for the Rasch analysis together with a reply paid envelope. Two weeks after its return, a second identical questionnaire was posted to them to complete. A 2-week period was selected to prevent recall bias while participants would not be expected to show any change in the day-to-day use of their arm.^(26, 27) To determine whether participants felt that the use of their affected limb remained stable during the study period a five-point global change score was used as a reference criterion.^(28, 29) Response options were attributed directly to the affected limb and ranged from 1 – ‘*Much less than last time*’, 2 – ‘*A little less than last time*’, 3 ‘*No change to last time*’, 4 – ‘*A little better than last time*’ 5 – ‘*Much better than last time*’.

6.3.3 Data analyses

All statistical analyses to address the ‘a priori’ hypotheses were undertaken using SPSS Statistics version 22.0. On the basis of recent tabled calculations, in order to have 90% probability or ‘assurance’ of obtaining a 95% confidence interval with a precision of 0.15, (i.e. a total width of 0.30), for an intraclass correlation of 0.80 a sample size of 41 participants is required.⁽³⁰⁾ To allow for non-completion 43 people were recruited. The COSMIN checklist informed the analyses undertaken in this study.⁽³¹⁾ Descriptive statistics were used to describe the sample. Data were analysed separately for each of the three subscales and the summed score. Normality of the data was evaluated using visual inspection together with skewness and kurtosis statistics and checked for any missing responses. Data were first analysed for systematic error by comparing the mean change between the two data collection times using paired *t* tests ($p = 0.05$).

Reliability analyses: Test retest reliability was assessed using a one-way repeated model analysis of variance Intraclass Correlation Coefficient (ICC 1.1)^(32, 33) with 95% confidence intervals. An ICC of greater than .0.70 was considered an acceptable standard for good reliability.⁽²⁰⁾ Internal consistency was examined using Cronbach alpha. A 0.80 – 0.95 score was considered an acceptable measure of internal consistency, with a reliability coefficient above 0.80 suitable for group comparisons and above 0.90 for individual comparisons.^(20, 34)

Agreement analyses: Agreement parameters assist in the interpretation of change scores over time.^(31, 35) Three agreement parameters were examined; i) the Standard Error of Measurement (SEM), a measure of response stability expressed in the same units as the original measure.⁽²⁶⁾ The formulae to calculate the SEM was $SD (\sqrt{1 - ICC})$.⁽²⁶⁾ Ninety five percent confidence intervals were calculated based on observed score $\pm 1.96 \times SEM$. ii) Minimal Detectable Change (MDC), i.e. the smallest amount of change that can be

considered above the threshold of error to determine what score may reflect actual change.

⁽²¹⁾ The MDC_{90} was calculated as $1.65 \times SEM \times \sqrt{2}$, together with 95% CIs. As the MDC is calculated from reliability statistics, it was included in this study as a further measure to quantify error, as recommended in the COSMIN guidelines. iii) Bland Altman plots enable an analysis of the observed error, identification of any systematic differences and outliers by plotting the spread of the scores around zero. ^(21, 36) The mean score for each participant was plotted on the x-axis, and the difference between scores on the y-axis. In an ideal situation all differences would equal zero, however, in the real world this is unlikely as some degree of error will always occur. ⁽³⁷⁾ The limit of agreement (LoA) represents the range within which most differences lie, i.e. the magnitude of the error. Greater variability indicates larger error and data points that occur outside the LoA are likely to represent real difference between the two time points, not random error. LoA were calculated as the mean difference $\pm$ the standard deviation of the mean difference multiplied by 1.96. ⁽³⁶⁾ Heteroscedasticity was considered to be absent if the difference between time 1 and time 2 followed a non-linear relationship on visual examination. ⁽³⁸⁻⁴⁰⁾

Results

Forty-three participants, recruited from four outpatient clinics throughout Australia, completed the reproducibility study. No participants rated themselves as having much better use of their arm at time point 2 and none as having less use based on the global change score, and there was no missing data. Of the eight who felt they had better use of their arm, none changed by greater than 1SD. All data were retained for analyses. Table 1 outlines the demographic characteristics and demonstrates a wide spread of injury level consistent with the BPI population. Table 2 outlines the participant characteristics. There was a significant

difference in time post injury between the reproducibility cohort and the Rasch only cohort ($t = 3.13, p = 0.003$) meaning that the reproducibility group was longer post injury and more likely to be stable in their ability to use their arm for day-to-day tasks. Visual inspection and skewness statistics confirmed a normal distribution (skewness range - 0.52 sub scale 1 time 1 to - 0.15 sub scale 3 time 1). The results of the paired t tests showed no statistically significant differences between the scores for each of the subscales and summed scores indicating no systematic bias in the data (Table 3).

Reliability

Test retest reliability was high, with ICCs ranging from 0.90 for subscale 3, to 0.97 for the summed score and subscale 2 (Table 4). These results supported hypothesis 1. Internal consistency was also high, ranging from a Cronbach alpha of 0.90 to 0.98 (Table 4). This result indicated that the three subscales and the summed score consisted of homogeneous sets of items that appear to be measuring a single construct. This supported hypothesis 2.

Agreement parameters

The SEM scores ranged from 1.6 to 4.5 (Table 4). The MDC_{90} ranged from 3.7 for subscale 3 to 10.3 for the summed score (Table 4). Bland Altman plots are presented in Figure 1. Data were evenly distributed above and below the mean for all sub scales and the summed score, indicating no systematic differences for any data set and no evidence of heteroscedasticity. No plot demonstrated more than 3 data points greater than two SDs away from the mean difference for any of the subsets or the summed score. This supported hypothesis 3.

Discussion

The BrAT is a new patient-reported outcome measure developed to assess solely activity following adult traumatic BPI. To the best of our knowledge is the first outcome measure specifically developed and psychometrically evaluated for this population. The results of this study support the psychometric properties of test-retest intra-rater reliability, internal consistency and agreement parameters indicating the BrAT is a reproducible outcome measure for this group. All results were within the boundaries of the *a priori* hypotheses and provide preliminary evidence to support to the use of the BrAT in the clinical setting as either a series of subscales or as a single summed score.

Test retest values were high sufficient for both individual and group level comparisons for all three subscales and the summed score.^(21, 26) The Cronbach Alpha values were also high pointing to the internal consistency or inter-relatedness of the items. One issue with the Cronbach Alpha is that it is not a measure of unidimensionality, only a measure of interrelatedness of the items. These results do not imply that the item sets are unidimensional, only that the items appear to be measuring one concept. However, they do support the use of the BrAT as a unidimensional measure of ‘activity of the upper limb’ following adult BPI. Further they support the use as both a total score or as a series of three separate subscales.⁽¹⁶⁾

The SEM and MDC₉₀ scores provide evidence of absolute reliability and aid in the interpretation of individual scores in the clinical setting. For example, a change of greater than 4 for subscale 3 (*No hand items*) or greater than 10 for the summed score (31 items) may indicate real change has occurred that is greater than random error. While the amount of change is relatively large (approximately 10% of the score for the total score and subscale 2)

it compares favourably with other patient-reported outcome measures for upper limb conditions. The DASH, for example, is the most widely used patient-reported outcome measure for BPI^(6, 7) although it has not been psychometrically evaluated for this population so direct comparisons are not possible. However, the MDC score for the DASH is variously reported as between 10 and 17 for a variety of upper extremity diagnostic conditions.⁽⁴¹⁾

BPI is a heterogeneous condition, which results in high within group variability. Some people present with almost no use of their arm while others may have almost full use. However, high within group variability is also known to result in a lower ICC, which leads to a higher SEM and MDC scores.⁽⁴²⁾ For this study the ICC was used as the reliability coefficient to determine the SEM and therefore the MDC scores. The ICC is considered by some to be a more accurate way to express measurement error as it takes into account any systematic difference between the data collection points, yet may have resulted in a higher SEM and therefore MDC scores.^(43, 44) Additional testing is required to confirm the SEM and MDC scores in larger cohorts. Further, while SEM and MDC are measures of observed change that occurred as a result of error or true change in a stable population, these results do not indicate if the observed change is clinically important or meaningful to adults with a BPI.⁽²⁷⁾

Agreement statistics such as SEM, MDC and LoA express error in the actual BrAT measurement units. The use of these statistics relies on the assumption of heteroscedasticity where the observed difference between scores at the two time points does not change with increasing mean values.⁽³⁸⁾ Absolute statistics cannot be used where the observed variance is dependent of the variable mean or heteroscedastic. Visual inspection of the Bland Altman plots did not reveal any evidence of increasing error as mean increased with values evenly

distributed for all 3 subscales and the summed score across all scores (Figure 1).⁽⁴⁰⁾

Therefore the assumption of homoscedasticity was not violated.

Limitations

While it is impossible to state that participants' level of ability did not change during the assessment period, the use of a global rating of change score ensured analyses were performed using data from people who perceived that their level of activity remained stable during the assessment period. Further the 2-week time frame between assessments would limit recall bias. The sample size was smaller than that recommended by the COSMIN group, however the sample used was based on sample size calculations specific to reliability studies.⁽³⁰⁾

Conclusions

The BrAT demonstrated reproducibility with high test retest reliability, internal consistency and agreement parameters for each of the three subscales and the summed score. Reliability on its own, while fundamental to the ability of a measure to evaluate outcome over time, cannot be used to justify an outcome measure's use, as a measure may be reliable but not necessarily valid.^(21, 26) Further testing is required to establish the construct validity and responsiveness of the BrAT.

References

- 299 1. Ahmed-Labib M, Golan JD, Jacques L. Functional outcome of brachial plexus
300 reconstruction after trauma. *Neurosurg.* 2007;61:1016-22.
- 301 2. Merrell GA, Barrie KA, Katz DL, Wolfe SW. Results of nerve transfer techniques for
302 restoration of shoulder and elbow function in the context of a meta-analysis of the English
303 literature. *J Hand Surg.* 2001;26:303-14.
- 304 3. Yang LJS, Chang KWC, Chung KC. A systematic review of nerve transfer and nerve
305 repair for the treatment of adult upper brachial plexus injury. *Neurosurg.* 2012;71:417-29.
- 306 4. Ali ZS, Heuer GG, Faught RW, Kaneriya SH, Sheikh UA, Syed IS, et al. Upper
307 brachial plexus injury in adults: comparative effectiveness of different repair techniques. *J*
308 *neurosurg.* 2015;122:195-201.
- 309 5. Garg R, Merrell GA, Hillstrom HJ, Wolfe SW. Comparison of nerve transfers and
310 nerve grafting for traumatic upper plexus palsy: A systematic review and analysis. *J Bone*
311 *Joint Surg.* 2011;93:819-29.
- 312 6. Dy CJ, Garg R, Lee SK, Tow P, Mancuso CA, Wolfe SW. A systematic review of
313 outcomes reporting for brachial plexus reconstruction. *J Hand Surg.* 2015;40:308-13.
- 314 7. Hill B, Williams G, Bialocerkowski A. Clinimetric Evaluation of Questionnaires
315 Used to Assess Activity After Traumatic Brachial Plexus Injury in Adults: A Systematic
316 Review. *Arch Phys Med Rehabil.* 2011;92:2082-9.
- 317 8. Bengtson KA, Spinner RJ, Bishop AT, Kaufman KR, Coleman-Wood K, Kircher MF,
318 et al. Measuring outcomes in adult brachial plexus reconstruction. *Hand Clin.* 2008;24:401-
319 15.
- 320 9. Wellington B. Quality of life issues for patients following traumatic brachial plexus
321 injury - Part 2 research project. *J Orthop Nurs.* 2010;14:5-11.
- 322 10. Franzblau L, Shauver MJ, Chung KC. Patient satisfaction and self-reported outcomes
323 after complete brachial plexus avulsion injury. *J Hand Surg.* 2014;39:948-55.e4.

- 324 11. Franzblau L, Chung KC. Psychosocial outcomes and coping after complete avulsion
325 traumatic brachial plexus injury. *Disabil Rehabil.* 2015;37:135-43.
- 326 12. Mancuso CA, Lee SK, Dy CJ, Landers ZA, Model Z, Wolfe SW. Expectations and
327 limitations due to brachial plexus injury: a qualitative study. *Hand.* 2015; 10:741-749
- 328 13. Hill B, Williams G, Olver J, Bialocerkowski A. Do existing patient-report activity
329 outcome measures accurately reflect day-to-day arm use following adult traumatic brachial
330 plexus injury? *J Rehabil Med.* 2015;47:438-44.
- 331 14. Mancuso CA, Lee SK, Dy CJ, Landers ZA, Model Z, Wolfe F. Compensation by the
332 injured arm after Brachial Plexus Injury. *Hand.* 2016;4:410-6.
- 333 15. WHO. International Classification of Functioning, Disability and Health.
334 Geneva.2001.
- 335 16. Hill B, Pallant J, Williams G, Olver J, Ferris S, Bialocerkowski A. Evaluation of
336 Internal Construct Validity and Unidimensionality of the Brachial Assessment Tool, A
337 Patient-Reported Outcome Measure for Brachial Plexus Injury. *Arch Phys Med Rehabil.*
338 2016;97:2146-56.
- 339 17. Zhou Y, Liu P, Rui J, Zhao X, Lao J. The clinical characteristics of neuropathic pain
340 in patients with total brachial plexus avulsion: A 30-case study. *Injury.* 2016;47:1719-1724.
- 341 18. Novak CB, Katz J. Neuropathic pain in patients with upper-extremity nerve injury.
342 *Physio Can.* 2010;62:190-201.
- 343 19. Bialocerkowski AE, Bragge P. Measurement error and reliability testing: Application
344 to rehabilitation. *Int J Ther Rehabili.* 2008;15:422-7.
- 345 20. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et
346 al. Quality criteria were proposed for measurement properties of health status questionnaires.
347 *J Clin Epidemiol.* 2007;60:34-42.

21. Portney L, G. Watkins. M,P Foundations of Clinical Research Applications to
Practice. Third Edition ed: Pearson Prentice Hall; 2009.
22. Hernaez R. Reliability and agreement studies: A guide for clinical investigators. Gut.
2015;64:1018-27.
23. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, et al.
Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin
Epidemiol. 2011;64:96-106.
24. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: What are the necessary
measurement properties? J Clinical Epidemiol. 1992;45:1341-5.
25. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: A
practical guide. Cambridge: Cambridge University Press.; 2011..
26. Streiner DN, GR. Cairney, J. Health Measurement Scales a practical guide to their
developent and use. Fourth Edition ed: Oxford University Press: New York; 2015.
27. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The
COSMIN checklist for evaluating the methodological quality of studies on measurement
properties: A clarification of its content. BMC Med Res Meth. 2010;10.
28. Cleland JA, Childs JD, Whitman JM. Psychometric Properties of the Neck Disability
Index and Numeric Pain Rating Scale in Patients With Mechanical Neck Pain. Arch Phys
Med Rehabil. 2008;89:69-74.
29. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The
COSMIN checklist for assessing the methodological quality of studies on measurement
properties of health status measurement instruments: An international Delphi study. Qual Life
Res. 2010;19:539-49.
30. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with
precision and assurance. Stat Med. 2012;31:3972-81.

- 373 31. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The
374 COSMIN study reached international consensus on taxonomy, terminology, and definitions
375 of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*.
376 2010;63:737-45.
- 377 32. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psych*
378 *Bull*. 1979;86:420-8.
- 379 33. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: An illustration
380 of appropriate statistical analyses. *Clin Rehabil*. 1998;12:187-99.
- 381 34. Nunnally J, Bernstein I. *Psychometric theory*. 3rd Edition ed. New York.: McGraw-
382 Hill; 1994.
- 383 35. Kottner J, Streiner DL. The difference between reliability and agreement. *J Clin*
384 *Epidemiol*. 2011;64:701-2.
- 385 36. Bland JM, Altman DG. Statistical methods for assessing agreement between two
386 methods of clinical measurement. *Lancet*. 1986;1(8476):307-10.
- 387 37. Giavarina D. Understanding Bland Altman analysis. *Biochemia Medica*.
388 2015;25:141-51.
- 389 38. Atkinson G, Nevill AM. Statistical methods for assessing measurement error
390 (reliability) in variables relevant to sports medicine. *Sport Med*. 1998;26:217-38.
- 391 39. De Vet HCW, Bouter LM, Dick Bezemer P, Beurskens AJHM. Reproducibility and
392 responsiveness of evaluative outcome measures: Theoretical considerations illustrated by an
393 empirical example. *Int J Tech Assess Health Care*. 2001;17:479-87.
- 394 40. Brehm MA, Scholtes VA, Dallmeijer AJ, Twisk JW, Harlaar J. The importance of
395 addressing heteroscedasticity in the reliability analysis of ratio-scaled variables: An example
396 based on walking energy-cost measurements. *Dev Med Child Neurol*. 2012;54:267-73.

41. Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier C. Measuring the whole or the parts? Validity, reliability, and responsiveness of the disabilities of the arm, shoulder and hand outcome measure in different regions of the upper extremity. *J Hand Thera.* 2001;14:128-46.
42. Rosengren J, Brodin N. Validity and reliability of the Swedish version of the Patient Specific Functional Scale in patients treated surgically for carpometacarpal joint osteoarthritis. *J Hand Ther.* 2013;26:53-61.
43. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol.* 2006;59:1033-9.
44. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol.* 1999;52:861-73.

Figure legend

Figure 1 Bland Altmen plots. The solid line represents the mean difference score. The dashed lines represent the 95% upper and lower limits of agreement (2 standard deviations above and below the mean difference)

Suppliers list: SPSS Statistics version 22.0

Table 1. Participant demographics

Demographic		n = 43 (%)
Gender	Male	38 (89)
	Female	5 (11)
Injury level	C5/6	12 (28)
	C5 - C7	5 (11)
	C5 - C8	15 (35)
	C8/T1	4 (9)
	Complete avulsion	7 (16)
Mechanism of injury	Motor Car	8 (18)
	Motor Bike	23 (53)
	Bicycle	2 (5)
	Pedestrian	0 (0)
	Work Injury	4 (9)
	Fall from height	3 (7)
	Sporting injury	3 (7)
Pre injury dominance	Right	37 (86)
	Left	6 (14)
Injured limb	Right	23 (53)
	Left	20 (47)

Table 2. Participant characteristics

	Mean (SD)
n	43
Time post injury (weeks)	214 (166.15)
Age at time of injury (years)	39 (16.54)
Age at recruitment (years)	42 (16.12)
Initial Summed BrAT (max 93)	48 (26.21)
Initial Subscale 1(max 24)	16 (5.9)
Initial Subscale 2 (max 51)	22 (16.7)
Initial subscale 3 (max 18)	10 (5.7)

Key: SD, standard deviation.

Table 3. Paired differences between time point 1 and time point 2

	Mean diff T1 / T2	SD	95% CI	<i>t</i> value	Sig. (2-tailed)
Summed items	-.86	6.00	-2.70 - .972	-.95	.349
Sub scale 1	.70	2.97	-.22 – 1.61	1.54	.131
Sub scale 2	-.93	4.13	-2.20 - .34	-1.48	.147
Sub scale 3	-.62	2.34	.39 - -1.41	-1.62	.112

Key: T1, time point 1, T2 time point 2. SD, standard deviation; CI, confidence interval.

Table 4. Reliability and agreement of the Brachial Assessment Tool

Raw scores	ICC (1,1)	ICC 95% CI	Cronbach alpha	SEM	SEM 95% CI	MDC ₉₀	MDC ₉₀ 95% CI	LoA (+/-)
Summed items	.97	.95 - .98	.98	4.5	± 8.8	10.3	±16.9	11.6
Sub scale 1	.91	.86 - .95	.92	1.8	± 3.5	4.1	± 6.7	5.8
Sub scale 2	.97	.95 - .98	.97	2.8	± 5.5	6.5	± 10.7	8.0
Sub scale 3	.90	.84 - .94	.90	1.6	± 3.1	3.7	± 6.1	4.9

Key: ICC intra-class correlation coefficient. CI confidence interval. LoA, limits of agreement. SEM, standard error of the measurement. MDC₉₀ minimal detectable change

Figure 1. Bland Altman plots. The solid line represents the mean difference score. The dashed lines represent the 95% upper and lower limits of agreement (2 standard deviations above and below the mean difference).