ACRM
AMERICAN CONGRESS OF REHABILITATION MEDICINE

Archives of Physical Medicine and Rehabilitation

journal homepage: www.archives-pmr.org

Archives of Physical Medicine and Rehabilitation 2017;■:■■■-■■■

ORIGINAL RESEARCH

Psychometric Evaluation of the Brachial Assessment Tool Part 1: Reproducibility

Bridget Hill, PhD,[a,b] Gavin Williams, PhD,[b] John Olver, MBBS,[b] Scott Ferris, MBBS,[c] Andrea Bialocerkowski, PhD[a]

From the [a]Menzies Health Institute, Brisbane, QLD; [b]Epworth Monash Rehabilitation Medicine Unit Epworth HealthCare, Melbourne, VIC; and [c]The Alfred, Melbourne, VIC, Australia.

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 (N=43; age range, 19−82y).

Interventions: People 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 model 1,1 ($ICC_{1,1}$), and internal consistency using Cronbach α. Agreement parameters included standard error of measurement, minimal detectable change, and limits of agreement.

Main Outcome Measure: BrAT.

Results: Test-retest reliability was excellent ($ICC_{1,1}$=.90−.97). Internal consistency was high (Cronbach α=.90−.98). Measurement error was relatively low (standard error of measurement 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).

Conclusions: 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.

Archives of Physical Medicine and Rehabilitation 2017;■:■■■-■■■

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 after BPI has been primarily impairment based.[6-8] Day-to-day 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 total scores must be viewed with caution. Further, the DASH may not 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 4-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,

Disclosures: none.

0003-9993/17/$36 - see front matter © 2017 by the American Congress of Rehabilitation Medicine
https://doi.org/10.1016/j.apmr.2017.10.015

"execution of a task or action by the individual,"[15(p.5)] 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 after 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 3 separate subscales: (1) 8 dressing and grooming items, (2) 17 whole arm and hand items, and (3) 6 no hand items; or alternatively, all 31-items may be added to produce a summed score. The BrAT item responses are scored as 0 (cannot do now), 1 (very hard to do now), 2 (a little hard to do now), and 3 (easy to do now).

Recovery from BPI occurs over a prolonged period of time and has a significant effect 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 (ie, 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 (ie, 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 2 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-retest and intrarater reliability, defined as "the degree to which one rater can obtain the same rating on multiple occasions of measuring the same variable."[21(p.870)] Internal consistency indicates how interconnected the items are (ie, all the items appear to be related to each other and measuring something similar).[21,26] Agreement is related to absolute measurement error (ie, how close repeated-measure scores are), expressed in the actual units of the measure. In essence, reliability coefficients enable discrimination of people, whereas agreement addresses how scores differ.

The purpose of this article was to investigate the 2 parameters of reliability (test-retest reliability and internal consistency) and 3 parameters of agreement (standard error of measurement, minimal detectable change [MDC], and Bland-Altman limits of agreement [LoA]). A priori hypotheses were established based on the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) guidelines. We expected that (1) the BrAT will demonstrate high test-retest reliability with an intraclass correlation coefficient (ICC) of >0.8, (2) the BrAT will demonstrate high internal consistency with a Cronbach α of $\geq$0.7 and, (3) 95% of the Bland-Altman LoA scores will fall within 2 SDs above and below the mean difference score.

List of abbreviations:

BPI	**brachial plexus injury**
BrAT	**Brachial Assessment Tool**
CI	**confidence interval**
COSMIN	**COnsensus-based Standards for the selection of health status Measurement INstruments**
DASH	**Disabilities of the Arm, Shoulder and Hand**
ICC	**intraclass correlation coefficient**
LoA	**limits of agreement**
MDC	**minimal detectable change**
MDC_{90}	**minimal detectable change based on a 90% confidence interval**

Methods

This project used a multicenter, prospective repeated-measure design. Ethical approval was gained from 3 human research and ethics committees (Griffith University PES_12_13_HREC, Alfred Health 425/11, and 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 magnetic resonance imaging, nerve conduction studies, intraoperative findings, or clinical assessment, and were >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 >12 weeks postinjury and who had not undergone surgery to reanimate the upper limb within the previous 2 years were invited to take part. Therefore, the function of their arm was likely to remain stable for the duration of this project because minimal recovery may be expected. Exclusion criteria included inability to provide informed consent, preexisting upper limb conditions that affected day-to-day activity, evidence of spinal cord injury confirmed by magnetic resonance imaging, or a diagnosis of brachial plexus birth injury.[16]

Data collection

Once participants consented to participate, they were mailed 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 mailed 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 5-point global change score was used as a reference criterion.[28,29] Response options were attributed directly to the affected limb and were scored as 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), and 5 (much better than last time).

Data analyses

All statistical analyses to address the a priori hypotheses were undertaken using SPSS Statistics version 22.0.[a] On the basis of recent tabled calculations, to have 90% probability or assurance of obtaining a 95% confidence interval (CI) with a precision of .15 (ie, a total width of .30), for an intraclass correlation of .80, a sample size of 41 participants is required.[30] To allow for

Table 1 Participant demographics (N=43)

Demographic	n (%)
Sex	
Male	38 (89)
Female	5 (11)
Injury level	
C5-6	12 (28)
C5-7	5 (11)
C5-8	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)
Gun shot	0 (0)
Preinjury dominance	
Right	37 (86)
Left	6 (14)
Injured limb	
Right	23 (53)
Left	20 (47)

noncompletion, 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 analyzed separately for each of the 3 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 analyzed for systematic error by comparing the mean change between the 2 data collection times using paired t tests ($P=.05$).

Reliability analyses

Test-retest reliability was assessed using a 1-way repeated model analysis of variance ICC model 1.1[32,33] with 95% CIs. An ICC of >.70 was considered an acceptable standard for good reliability.[20] Internal consistency was examined using Cronbach α. A .80 to .95 score was considered an acceptable measure of internal consistency, with a reliability coefficient >.80 suitable for group comparisons and >.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. The first agreement parameter was the standard error of measurement, a measure of response stability expressed in the same units as the original measure.[26] The formulae to calculate the standard error of measurement was SD ($\sqrt{1-ICC}$).[26] Then 95% CIs were calculated based on the observed score$\pm$1.96$\times$standard error of measurement. The second agreement parameter was the MDC (ie, the smallest amount of change that can be considered above the threshold of error to determine what score may reflect actual change).[21] The MDC based on a 90% confidence interval (MDC_{90}) was calculated as 1.65$\times$standard error of measurement$\times\sqrt{2}$, together with 95% CIs. Because 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. The third agreement parameter, Bland-Altman plots, enables an analysis of the observed error and identifies 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 was plotted on the y axis. In an ideal situation, all differences would equal zero; however, in the real world this is unlikely because some degree of error will always occur.[37] The LoA represents the range within which most differences lie (ie, 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 2 time points, not random error. LoA were calculated as the mean difference $\pm$ SD 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 nonlinear relation on visual examination.[38-40]

Results

Forty-three participants, recruited from 4 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 8 participants who felt they had better use of their arm, none changed by >1 SD. 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 postinjury between the reproducibility cohort and the Rasch only cohort ($t=3.13$, $P=.003$), meaning that the reproducibility group was longer postinjury 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. 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 .90 for subscale 3 to .97 for the summed score and subscale 2 (table 4). These results supported hypothesis 1. Internal consistency was also high, ranging from a Cronbach α of .90 to .98 (see table 4). This result indicated that the 3 subscales and the summed score

Table 2 Participant characteristics (N=43)

Characteristic	Mean $\pm$ SD
Time postinjury (wk)	214$\pm$166.15
Age at time of injury (y)	39$\pm$16.54
Age at recruitment (y)	42$\pm$16.12
Initial summed BrAT (max, 93)	48$\pm$26.21
Initial subscale 1 (max, 24)	16$\pm$5.9
Initial subscale 2 (max, 51)	22$\pm$16.7
Initial subscale 3 (max, 18)	10$\pm$5.7

Abbreviation: max, maximum.

Table 3 Paired differences between T1 and T2

BrAT Scales	Mean Difference T1/T2 ± SD	95% CI	*t*	Significance (2-tailed)
Summed items	−.86±6.00	−2.7 to 1.0	−.95	.349
Subscale 1	.70±2.97	−2.2 to 1.6	1.54	.131
Subscale 2	−.93±4.13	−2.2 to .3	−1.48	.147
Subscale 3	−.62±2.34	.3 to −1.4	−1.62	.112

Abbreviations: T1, time point 1; T2, time point 2.

consisted of homogeneous sets of items that appear to be measuring a single construct. This supported hypothesis 2.

Agreement parameters

The standard error of measurement scores ranged from 1.6 to 4.5 (see table 4). The MDC_{90} ranged from 3.7 for subscale 3 to 10.3 for the summed score (see table 4). Bland-Altman plots are presented in figure 1. Data were evenly distributed above and below the mean for all subscales and the summed score, indicating no systematic differences for any data set and no evidence of heteroscedasticity. No plot demonstrated >3 data points >2 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 after adult traumatic BPI. To our knowledge, this 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 intrarater 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 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 highly sufficient for both individual- and group-level comparisons for all 3 subscales and the summed score.[21,26] The Cronbach α values were also high pointing to the internal consistency or interrelatedness of the items. One issue with the Cronbach α 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 after adult BPI. Further, they support the use as both a total score or as a series of 3 separate subscales.[16]

The standard error of measurement and MDC_{90} scores provide evidence of absolute reliability and aid in the interpretation of individual scores in the clinical setting. For example, a change of >4 for subscale 3 (no hand items) or >10 for the summed score (31 items) may indicate real change has occurred that is greater than random error. Although the amount of change is relatively large (approximately 10% of the score for the total score and subscale 2), it compares favorably 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]; however, 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 being 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, whereas 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 standard error of measurement and MDC scores.[42] For this study, the ICC was used as the reliability coefficient to determine the standard error of measurement and therefore the MDC scores. The ICC is considered by some to be a more accurate way to express measurement error because it takes into account any systematic difference between the data collection points, yet may have resulted in a higher standard error of measurement and therefore MDC scores.[43,44] Additional testing is required to confirm the standard error of measurement and MDC scores in larger cohorts. Further, although standard error of measurement 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 standard error of measurement, 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 2 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 because the mean increased with values evenly distributed for all 3 subscales and the summed score across all scores (see fig 1).[40] Therefore, the assumption of homoscedasticity was not violated.

Study limitations

Although 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

Table 4 Reliability and agreement of the BrAT

Raw Scores	ICC Model 1,1	ICC 95% CI	Cronbach α	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

Abbreviation: SEM, standard error of the measurement.

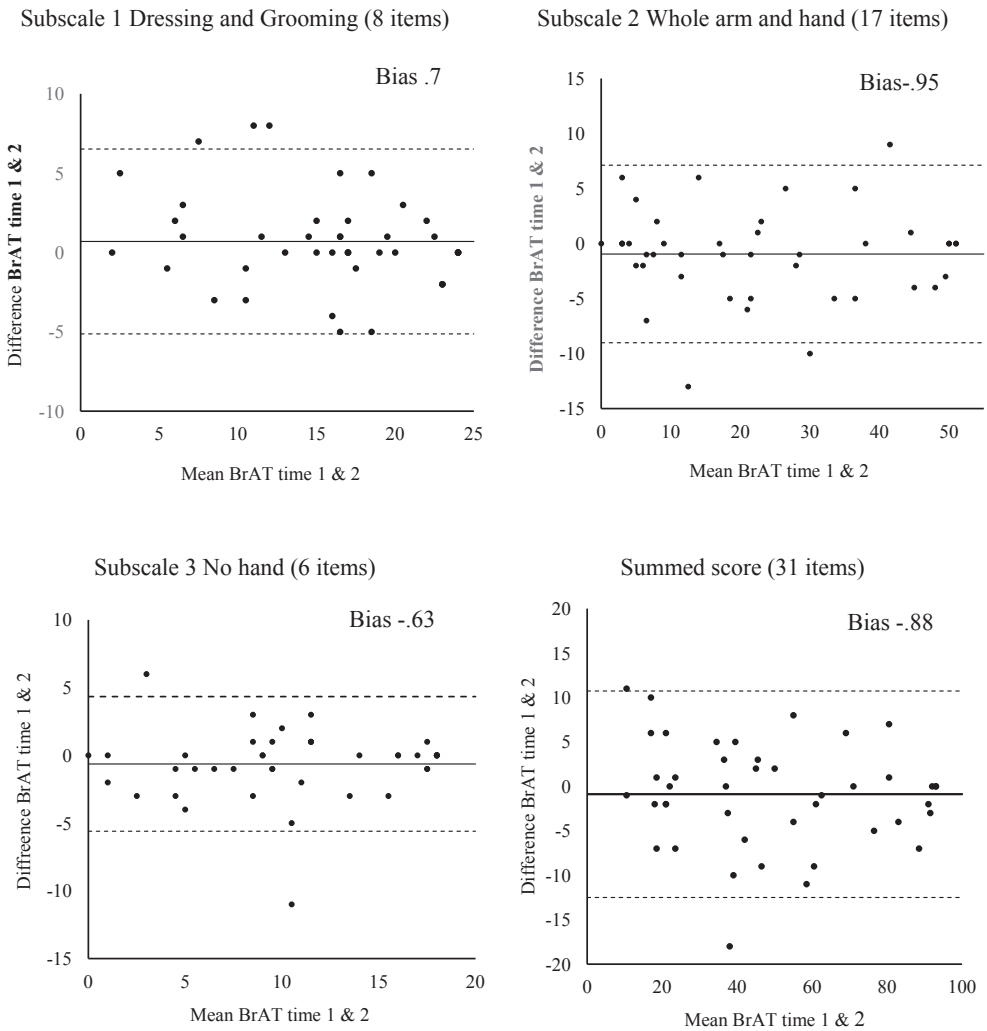

Fig 1 Bland-Altman plots. The solid line represents the mean difference score; dashed lines, 95% upper and lower LoA (2 SDs above and below the mean difference).

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 3 subscales and the summed score. Reliability on its own, although fundamental to the ability of a measure to evaluate outcome over time, cannot be used to justify an outcome measure's use because 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.

Supplier

a. SPSS Statistics version 22.0; IBM.

Keywords

Brachial plexus; Outcome assessment (health care); Rehabilitation; Reproducibility of results

Corresponding author

Bridget Hill, PhD, Epworth Monash Rehabilitation Medicine Unit Epworth HealthCare, 89 Bridge Rd, Melbourn, VIC 3122, Australia. *E-mail address:* bridget.hill@epworth.org.au.

Acknowledgments

We thank Jaslyn Gibson, BSc(OT), Occupational Therapist, Perth, WA, Australia; David McCombe, M.B.B.S., Victorian Hand Surgery Associates and St Vincent's Hospital, Melbourne, VIC, Australia; and Melanie McCulloch, BSc(OT), Re-wired Hand Therapy, Melbourne, VIC, Australia, for their valuable contribution to this project during the recruitment and data collection phase.

References

1. Ahmed-Labib M, Golan JD, Jacques L. Functional outcome of brachial plexus reconstruction after trauma. Neurosurgery 2007;61:1016-22.
2. Merrell GA, Barrie KA, Katz DL, Wolfe SW. Results of nerve transfer techniques for restoration of shoulder and elbow function in the context of a meta-analysis of the English literature. J Hand Surg 2001;26:303-14.
3. Yang LJ, Chang KW, Chung KC. A systematic review of nerve transfer and nerve repair for the treatment of adult upper brachial plexus injury. Neurosurgery 2012;71:417-29.
4. Ali ZS, Heuer GG, Faught RW, et al. Upper brachial plexus injury in adults: comparative effectiveness of different repair techniques. J Neurosurg 2015;122:195-201.
5. Garg R, Merrell GA, Hillstrom HJ, Wolfe SW. Comparison of nerve transfers and nerve grafting for traumatic upper plexus palsy: a systematic review and analysis. J Bone Joint Surg 2011;93:819-29.
6. Dy CJ, Garg R, Lee SK, Tow P, Mancuso CA, Wolfe SW. A systematic review of outcomes reporting for brachial plexus reconstruction. J Hand Surg 2015;40:308-13.
7. Hill B, Williams G, Bialocerkowski A. Clinimetric evaluation of questionnaires used to assess activity after traumatic brachial plexus injury in adults: a systematic review. Arch Phys Med Rehabil 2011;92:2082-9.
8. Bengtson KA, Spinner RJ, Bishop AT, et al. Measuring outcomes in adult brachial plexus reconstruction. Hand Clin 2008;24:401-15.
9. Wellington B. Quality of life issues for patients following traumatic brachial plexus injury - part 2 research project. J Orthop Nurs 2010;14:5-11.
10. Franzblau L, Shauver MJ, Chung KC. Patient satisfaction and self-reported outcomes after complete brachial plexus avulsion injury. J Hand Surg 2014;39:948-955.e4.
11. Franzblau L, Chung KC. Psychosocial outcomes and coping after complete avulsion traumatic brachial plexus injury. Disabil Rehabil 2015;37:135-43.
12. Mancuso CA, Lee SK, Dy CJ, Landers ZA, Model Z, Wolfe SW. Expectations and limitations due to brachial plexus injury: a qualitative study. Hand (N Y) 2015;10:741-9.
13. Hill B, Williams G, Olver J, Bialocerkowski A. Do existing patient-report activity outcome measures accurately reflect day-to-day arm use following adult traumatic brachial plexus injury? J Rehabil Med 2015;47:438-44.
14. Mancuso CA, Lee SK, Dy CJ, Landers ZA, Model Z, Wolfe F. Compensation by the injured arm after brachial plexus injury. Hand (N Y) 2016;4:410-6.
15. World Health Organization. International Classification of Functioning, Disability and Health. Geneva: World Health Organization; 2001.
16. Hill B, Pallant J, Williams G, Olver J, Ferris S, Bialocerkowski A. Evaluation of internal construct validity and unidimensionality of the brachial assessment tool, a patient-reported outcome measure for brachial plexus injury. Arch Phys Med Rehabil 2016;97:2146-56.
17. Zhou Y, Liu P, Rui J, Zhao X, Lao J. The clinical characteristics of neuropathic pain in patients with total brachial plexus avulsion: a 30-case study. Injury 2016;47:1719-24.
18. Novak CB, Katz J. Neuropathic pain in patients with upper-extremity nerve injury. Physio Can 2010;62:190-201.
19. Bialocerkowski AE, Bragge P. Measurement error and reliability testing: application to rehabilitation. Int J Ther Rehabil 2008;15:422-7.
20. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
21. Portney LG, Watkins MP. Foundations of clinical research applications to practice. 3rd ed. Upper Saddle River, NJ: 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, 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 HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine: a practical guide. Cambridge: Cambridge University Press; 2011.
26. Streiner DL, Norman GR, Cairney J. Health measurement scales a practical guide to their development and use. 4th ed. New York: Oxford University Press; 2015.
27. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010;10:22.
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, 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.
31. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.
32. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psych Bull 1979;86:420-8.
33. Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. Clin Rehabil 1998;12:187-99.
34. Nunnally J, Bernstein I. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.
35. Kottner J, Streiner DL. The difference between reliability and agreement. J Clin Epidemiol 2011;64:701-2.
36. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
37. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb) 2015;25:141-51.
38. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sport Med 1998;26:217-38.
39. De Vet HC, Bouter LM, Dick Bezemer P, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures: theoretical considerations illustrated by an empirical example. Int J Tech Assess Health Care 2001;17:479-87.
40. Brehm MA, Scholtes VA, Dallmeijer AJ, Twisk JW, Harlaar J. The importance of addressing heteroscedasticity in the reliability analysis of ratio-scaled variables: an example 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 Ther 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 HC, 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.