J. Phys. Ther. Sci. 30: 1271–1277, 2018

The Journal of Physical Therapy Science

Original Article

Test-retest reliability of physiotherapists using the action research arm test in chronic stroke

Polykarpos Angelos Nomikos, MSc[1)*], Nicola Spence, PhD[2)], Mansour Abdullah Alshehri, MSc[3)]

[1)] *Academic Rheumatology, School of Medicine, University of Nottingham: Nottingham, Nottinghamshire, United Kingdom*
[2)] *Sport, Exercise and Physiotherapy Department, School of Health Sciences, University of Salford, United Kingdom*
[3)] *Physiotherpy Department, Faculty of Applied Medical Sciences, Umm Al-Qura University, Saudi Arabia*

Abstract. [Purpose] The aim of this study was to determine whether physiotherapists (PT) scores are consistent over time when using Action Research Arm Test (ARAT) to assess upper limb (UL) function on a videotaped chronic stroke patient. [Participants and Methods] Quantitative correlational study. A convenience-snowball sample of 20 international PT (mean age and experience=32 ± 6.8 and 7.55 ± 7.4 years) used ARAT to score chronic stroke patient's UL function, observing a video at baseline and again ≈ 2 weeks later. Two sets of non-parametric ordinal data were assessed with Spearman's (rho) and the alpha (a) value was set at 0.01. Line of equality, Bland-Altman plots and Wilcoxon signed rank test were also considered. [Results] Spearman's rho was found ≈ 0.78 at a significance level of 0.00. ARAT was scored with a mean difference of 16.6 days and a mean change of 0.6 points was observed. Limits of agreement and coefficient of reproducibility were ± 2.3 and ± 2.6 respectively. The patient's arm impairment was categorised as moderate and floor or ceiling effects were not detected. [Conclusion] The results suggest that ARAT is consistent, valid and should be used by PT in chronic stroke.
Key words: Reliability, Stroke, Physiotherapist

(This article was submitted May 23, 2018, and was accepted Jul. 24, 2018)

INTRODUCTION

Stroke is a leading cause of disability or death, consisting of progressive neurological signs of cerebral dysfunction that last 24 hours, with a prevalent cause being of vascular origin[1)]. While stroke mortality in UK accounts for 71 people per 100,000, the incidence rate is 150/100,000[2, 3)], contributing to high costs to the NHS, with a total estimated cost of £9 billion annually for care[4)]. After stroke, modifications in cortical representation due to spontaneous recovery and neuroplasticity occur over time[5)]. Furthermore, 60% of survivors have impaired manual dexterity six months post-stroke and remain unable to perform arm-hand movements[6–9)]. Dexterity reflect peoples' performance in activities of daily living (ADL) and the aim of stroke rehabilitation is regaining upper limb (UL) function[6–11)].

Determining efficacy of treatments accurately and in a reproducible manner is crucial, thus clinicians remain concerned about the selection criteria of the most repeatable scale and its psychometric properties[12, 13)]. Furthermore, measuring UL's impairment in chronic stroke is essential for rehabilitation[12, 13)], and several outcome measures (OM) have been discussed with regard to the level of disability and relevance to UL progress[14–24)]. OM with good psychometric properties and feasible administration time conducted by appropriately trained staff are recommended for stroke by national clinical guidelines[25)];

*Corresponding author. Polykarpos Angelos Nomikos (E-mail: mbxpan@nottingham.ac.uk)

however, recommendations for tools or specific raters are not provided. The Action Research Arm Test (ARAT)[14–17, 19, 20–24] is one of the most commonly used tools to capture change in hemiplegia. ARAT measures the ability to handle and move qualitatively either large or small objects regarding dexterity and proximal strength for both UL. KNGF (royal Dutch society of physiotherapists) clinical guidelines[26] recommend ARAT to be applied by physiotherapists (PT) for UL examination in chronic stroke; however, EBRSR (evidence-based review of stroke rehabilitation)[27] and ASA (American stroke association)[28] clinical guidelines do not clarify who can use it and report significant floor and ceiling effects. Prior studies have reported the reliability and validity of ARAT to be high[14–17, 19, 21–24].

ARAT[18] is based on the upper extremity function test (UEFT)[16], which comprises four subscales: grasp, grip, pinch, and gross. Lyle[16] eliminated items that were inconsistent, reducing the time required to complete the test by 50% for patients with chronic hemiparesis (cortical damage) as a consequence of multiple brain injuries. ARAT's test-retest reliability (+0.98) was calculated with the Pearson product-moment correlation coefficient (r); a statistical measure useful for interval/ratio data[29], and characterised by overestimation of the results due to its inability to indicate the proportion of agreement obtained by chance[12, 30]. The disease was not specified by Lyle and statistics were misused. Wagenaar et al.[20] enhanced the consistency of the questionnaire, setting a time limit of 1 minute per task, to clarify scores 2 (takes abnormally long) and 3 (normal performance). Another reliability study[14], stated that Lyle's protocol would be inadvisable and impracticable to therapists with less than five-years-experience. Yet, test-retest reliability was not considered and inter-rater reliability (+0.98) was estimated by three therapists with five years of experience. The intra-class correlation coefficient (ICC) was analysed considering the final score of both hands that implies a deterioration of the results due to a positive estimation ipsilateral to the lesion[21]. Platz et al.[17] developed a standardised protocol identifying the test-retest (+0.96) of ARAT using a mixed small sample with different neurological deficits. Lin et al.[24] reported test-retest (+0.99) reliability of ARAT using ICC in chronic stroke, however, they do not provide information on this type of reliability essential for appraising the evidence. Van der Lee et al.[21] videotaped the sample measurements and utilised ICC, weighted kappas and Bland and Altman plots to report clinically significant results. Yozbatiran[23] tested inter-rater (+0.99) and intra-rater (+0.99) reliability of the standardised protocol of ARAT in chronic stroke (n=12). However, this study set the chronic phase as 3 months after the onset, did not explain statistics, and changed the value for MCID (minimal clinically important difference) from −2.4 to +2.8 points. Above all, only three studies[16, 17, 24] established test-retest reliability of ARAT in chronic stroke, but none of them stated to whom the protocol is referred.

The use of a video tape for analysis can enhance the accuracy of the assessment and minimise bias[31–34]. This project will use videotaped evidence of a chronic stroke patient performing ARAT, to determine whether PT's ARAT scores are consistent over time. The following research question was formulated: Is ARAT consistent when the same physiotherapists use it over time to score upper limb function on a videotaped chronic stroke patient?

PARTICIPANTS AND METHODS

A convenience and snowball sampling of 20 PT (mean age=32 ± 6.8 years, mean experience=7.55 ± 7.4 years) scored a videotaped administration of ARAT performed by a chronic stroke patient. The COSMIN checklist[35] provided guidance to the study design, ensuring that all the appropriate information required was reported to ease quality evaluation. The recruitment was via emails and posters. Previous studies[31–33] selected licenced PT with a minimum of one-year clinical experience to observe videotaped gait analysis and rate lower limb's outcome measures in chronic stroke. Therefore, the present study included qualified PT with more than a year of clinical experience regardless of ethnicity that met one of the following criteria: (1) male/female; (2) staff-student members from Salford University; (3) volunteers; (4) specialist/senior or expert; and (5) PT with prior familiarisation with the tool. Prior to testing, information sheets and consent forms were given to the participants.

The measurements were performed in a familiar location and a minimum time interval of one week was defined sufficient regarding the retesting, as being performed by other studies[17, 24, 31, 33], to minimise data recall from the first measurement and not cause maturation. Maturation refers to alternations within the indivisual factors that occur over time which may provide a non-consistent outcome to the later measurement[36]. The researcher(s) explained the scoring of ARAT to PT. The video was not paused during the observation to reflect a real time patient interaction. The participants (n=15) invited more PT (n=5) to increase the sample size using a non-probability technique of snowball sampling[36, 37]. "Intention to Treat" (ITT) was used for incomplete data[38]. The protocol received ethical approval from School Research Ethics, University of Salford, Salford, United Kingdom. The data collection occurred in May–June 2017.

The experimental (*H1*) and a null hypothesis (*H0*) are stated below:

• *H1*: If the same physiotherapists score the same videotaped performance of a chronic stroke patient twice using ARAT and their scores are similar, then results are due to consistency.

• *H0*: If the same physiotherapists score the same videotaped performance of a chronic stroke patient twice using ARAT and their scores are not related, then results are due to chance factors.

This is a 2-tailed hypothesis testing, since direction is not yet identified[29, 37], and the level of statistical significance Alpha (a) value was set at 0.01. This study focuses on the stability of ARAT over time, so both direction and relationship among variables (time−ARAT scores) are being predicted naturally without manipulation[29, 37]. Due to the study's correlational design, there are no independent or dependent variables[29]. Extraneous variables such as situational variables (noise,

temperature, lighting degree, cleanliness, neatness, size of the room, convenience of the participants, an isolated room), personal variables (fatigue, mental health, bad/good mood, headaches and dizziness, intelligence and previous knowledge in neurology, anxiety, nerves, concertation, distraction, stress and bad visual acuity) and other variables (interpretation, balance in sound, duration of video, size of the screen quality of video and angle of the camera, surroundings, researcher's attitude and personality) may have an impact on ARAT performance scores[32, 37, 39, 40]; however, not all of them were recorded during the study due to unforeseen circumstances. The ARAT subtests formulate scores in an ordinal set of data[16–18].

The correlational study was analysed with the Spearman rank order correlation coefficient (rho)[29] and the sum scores of each questionnaire were used to investigate the reliability as proposed by previous research[15, 27]. Non-parametric correlation tests are appropriate for ordinal data[15, 29] and Spearman's rank correlation coefficient measures the strength and direction of association between two variables (test, retest)[29] using SPSS. The means and standard deviation (SD) were estimated from continuous variables (age, time, experience). However, a high correlation (rho) does not always represent good agreement between two repeated measurements[30]; thus, the line of equality was also utilised to illustrate visual agreement among the two methods (test against retest)[31,37, 41] and the Bland and Altman plot[30, 41, 42] that addresses the proportion of agreement between two clinical measurements (2 sets of data) and searches for possible outliers[30]. It must be noted that the mean difference (δ) is zero, since the same process has been used[30, 41, 42]. The Wilcoxon signed rank test was used for ordinal data to assess the mean rank scores (cases) and the statistical significance of differences between scores[43], as in previous studies[17, 19].

The video was displayed on two different screens. The time limit of 1 minute was set for the scores to be distinguishable[20] and the mean difference between the two measurements was 16.6 days ± 4.9, ensuring independent administration of the tool. Lyle's protocol[16] was assigned to be scored by PT and no specific instructions were given. In actual clinical practice, a brief review and familiarisation with an instrument prior to its use is mandatory[33]; this was achieved by giving adequate time for the PT to familiarise themselves with the protocol. After the first session, PT was advised not to discuss, utilise, or further investigate the questionnaire until their final observation.

RESULTS

From the 20 sets of first data, 19 were collected after the second measurement. None of the PT had prior experience scoring the ARAT. One of the participants dropped out of the study due to personal reasons. A statistical approach known as ITT analysis was found useful on this occasion, since it ignored withdrawals and non-compliance in this study[38]. The observed tasks were completed within the time limit. The international sample had a median of 5 years-experience as PT and a mean age of 32 years ± 6.8. Their demographics and origin are presented in Table 1 and Table 2 respectively. The reliability of the sum scores was calculated, and the value for the Spearman (rho) was found higher than 0.77 (Table 3), indicating good agreement between the measurements over time. The statistical significance p-value was determined as 0.00, which is small and less than the a-value, {p<a, (0.00<0.01)}, so the *H0* was rejected and *H1* is accepted.

Non-parametric correlations are shown in Table 3. Line of equality showed that the range of scores from 14.00 to 22.00. A non-parametric Wilcoxon signed rank test was conducted to evaluate whether a statistically significant difference existed between the mean ARAT scores of individuals that watched the video over time *(H1)*. The value of mean scores on test (*T1*) was 17.4, while on retest (*T2*) decreased to 16.8, thus the mean change is 0.6 (small), suggesting that there are no statistically significant differences between them as shown in Table 3. Table 3 also indicates the ranks and there are seven cases where retest scores were lower than test (a), one case where retest was higher than test (b) and twelve which were the same (c). Table

Table 1. Demographics

Items	Value
Age (years)	32 ± 6.81
Gender (male/female)	12/8
Years of experience (years)	7.55 ± 7.4
Test retest (days)	16.6 ± 4.9
Staff/students/other	3/14/3
Instrument calibration (times)	2
Duration of the video (min)	12
Size of the screen (inches)	14/120
Image quality (pixels)	1,920 × 1,080/1,280 × 720
During the study: bad visual acuity −anxiety	=3
During the study: ARAT knowledge −neuroscience	=4
Adverse events (headaches, dysphoria, dizziness)	None

SD: standard deviation.

Table 2. Ethnicity and potential threats to internal validity

County	Number of participants	Location of test	PT	Major factors that may have an impact on ARAT performance scores
United Kingdom	3	Familiar	2 Staff	Attentiveness-instrumentation
Saudi Arabia	7	Familiar	Students	Fatigue −motivation −drop out
Nigeria	1	Familiar	Student	Visual acuity −fatigue
India	1	Familiar	Student	✓
Brazil	4	2 not familiar	1 Staff	Motivation
Greece	4	1 not familiar	1 Student	Viewing distance

PT: physiotherapist; ARAT: action research arm test.

Table 3. Statistical Analysis (n=20)

1. Non-parametric correlations (Spearman's rho)	Test	Retest	p value		
	1.000	0.775	0.000*		
2. Descriptive statistics (mean ± SD)	Test	Retest			
	17.4 ± 1.9	16.8 ± 1.85			
3. Wilcoxon Signed Rank (Ranks)		Number	Mean Rank	Sum of Ranks	
	Negative ranks	7[a]	4.71	33	
	Positive ranks	1[b]	3	3	
	Ties	12[c]	-	-	
	Total	20	-	-	
	Z value (retest−test)	−2.15[¥]			
	p value	0.031			
4. Coefficients		Unstandardized		Standardized	t value	p value
		B	Std. error	Beta		
	Constant	0.09	2.7	-	0.033	0.97
	Mean	0.03	0.16	0.045	0.19	0.85

*p<0.05; [a]retest<test; [b]retest>test; [c]retest=test; [¥]based on positive ranks.

3 shows a z-value of −2.154 associated with a p-value of 0.031, that is more than the a-value of 0.01 (p>a), so the *H0* that there are no statistically significant differences between the scores of the sample is accepted. Coefficient of reproducibility (CR) is expressed as two times the difference of SD, CR=2 × SD and the 95% limits of agreement (LOA) are (δ=0) −1.96 SD and (δ=0) + 1.96 SD[30, 41, 42]. CR=2SD (d2-d1)[30]=$(-1)^2$ +1+$(2)^2$ +$(1)^2$ +$(1)^2$ +$(4)^2$ +$(3)^2$ +$(1)^2$ ÷ N=34÷20=1.7 →CR=2 ×(√1.7) →CR=±2.6. The LOA are upper line (green): 1.96SD (+2.31) and lower line (red): −1.96SD (−2.31). The plot indicates two outliers from the LOA. A linear regression procedure was administrated to detect if there is any proportional bias, since there is a trend of points being above the mean difference. As can be seen in Table 3, the t-value (t=0.190) is linked with a statistical significance of p=0.851, greater than the a-value (0.01), so again *H0* (null hypothesis) is accepted. Therefore, there is no proportional bias and there is a certain level of agreement between measurements.

DISCUSSION

The study investigated test-retest reliability of PT using ARAT in chronic stroke. Yozbatiran's standardised instructions[23] did not correspond to the videotape assessment, so Lyle's decision rules[16] in conjunction with the time limit of 1 minute[20] were implemented. The correlation coefficient (rho) of ≈ 0.78 linked with a statistical significance of 0.00, shows that ARAT is a reliable, consistent, and reproducible OM when used by PT in chronic stroke. The line of equality showed a positive correlation considering that the results lie along a line. However, relative reliability may give misleading results and a positive correlation is not necessarily related to good agreement[30, 37, 41, 42]. A previous study[16] overestimated test-retest reliability, using Pearson correlation that applies to parametric data, without converting the non-parametric ARAT data. The statistically significant results contradict Hsieh's findings[14], which suggested that therapists with less than five years of experience are not consistent using ARAT. Our results are relatively consistent with previous findings[24]; the mean change of 0.6 points associated with a median of 16 days in this study is consistent with the mean change of 0.3 points and a median of 15 days

found by van der Lee[44]. CR is not mentioned in their study[44], the mean difference (δ) is not considered 0 in plots, as usual for repeated measurements[30], and LOA (−5.7, +6.2) were greater than the MCID. The Wilcoxon signed rank test was used to analyse descriptive statistics and mean rank scores. Lin[24] targeted the examiners for ARAT without defining terms such as 'specially trained PT', thus Lin's[24] sample is not representative and their results cannot be generalised due to sampling error. To the best of our knowledge, none of the previous studies[16–18, 21, 23, 24, 44] considered Bland and Altman plots and CR to estimate ARAT's test-retest reliability.

PT with multiple nationalities were involved and either a laptop or a projector was utilised to play the videotape in each case accordingly. Situational variables were controlled; measurements being performed late morning provided an adequate degree of lighting, while the rooms for PT were clean, quiet and ensured a comfortable temperature to avoid distraction and injuries[45]. The individuals were observed naturally and examined verbally for their health (headaches, injury, convenience) and well-being (stress, thoughts, feelings). Natural observation refers to the process when the researcher joins the group to obtain a deeper understanding of PT[46]. Fatigue and distance from the display are two of the personal variables not recorded that may have influenced the outcome of the study; 8 PT that scored the protocol twice were Muslims and measurements were conducted during Ramadan. Fasting is directly linked with fatigue and headaches[47], however, none of the PT complained about headaches. Three PT (1 Greek, 2 Brazilian) that scored the protocol were not familiar with the environment. The mean difference was 16.6 ± 4.9 (days) when the measurements performed, with a range of days (7–24). More than one week was highlighted by previous studies[33] to reduce the potential for memory effects. Although, standardisation regarding maturation of a sample has not been yet identified and it basically depends on the nature of research[48]. Most PT that participated in the study were students and staff members, so knowledge acquisition may comprise a factor of maturation, particularly in longer term changes[48]. However, the mean difference of 16.6 days among the two measurements was within the time period suggested by previous studies[33, 49]. Motivation, reward, and interest are three extraneous variables that may have influenced the scoring procedure and were not assessed during the study.

Considering potential threats to internal validity, instrumentation and history may influence the findings of a study[48]. On this occasion, instrumentation occurs when the device used alters over time, so instrumental bias may have affected the results. In total, 39 observations were performed (1 drop out): 16 observations in which the video was played through a 14-inch laptop and 18 observations via a 120-inch projector, matched for both test-retest. The device provided was changed from projector to laptop for 4 of the observations regarding retesting. While Neuman[50] suggests that larger screens with a poor-quality image may provide a greater sense of presence, enjoyment, and intense responses, Lombard[51] gives more value to smaller displays with high resolution. Consequently, stability and balance within electronic devices were verified in this study and instrumentation was not a limiting factor.

Chanubol et al.[19] defined a classification scale to track the severity of arm impairment based on ARAT, with a score less than 10 determined severe. Using Chanubols' classification system in conjunction with the range of scores (14−22) and the mean scores of 16.8 and 17.4 respectively, the present study categorises patient's arm impairment as moderate. This was confirmed by Nijland et al.[52], who compared the Wolf Motor Function Test (WMFT) with ARAT and recommended the protocol for patients with moderate hemiparesis, while van der Lee et al.[22] stated that ARAT may not be sensitive enough to identify change in chronic stroke patients approaching severe or normal arm function. International guidelines[27] and previous studies[22, 24, 52] demonstrated that ARAT has notable floor and ceiling effects when capturing changes in severe or mild stroke patients. Floor and ceiling effects of ARAT are calculated as 5% of the scale, and to be identified, 20% of the total scores need to be appointed above or below the LOA. Bland and Altman plot showed two outliers, so there are no significant floor and ceiling effects in this study and the patient is likely to have moderate impairment[52], in agreement with Chanubol et al.[19], but if the patient had severe or near normal UL impairment, this study would have given robust evidence regarding floor and ceiling effects. This study did not confirm the patient's non-affected side.

Interestingly, the resulting value of CR=± 2.6 points approaches numerically the LOA (± 2.3) in this study, confirming that these two values are statistically related to each other, as proposed by other studies[53, 54]. It must be noted that CR does not represent the MCID[55], and clinicians remain subjective while estimating CR, since many studies proposed the equation differently[30, 41, 42, 53]. CR is a psychometric property of ARAT which reflects the instrument's ability to capture a difference of ± 2.6 points, while MCID is the selected minimal value, previously set by clinicians to determine whether the outcome has been clinically significant due to intervention[53, 55]. The value of CR=± 2.6 in this study is less than the MCID=5.7 points of ARAT (10% of the scale)[21] (CR<MCID), verifying that ARAT is a valid outcome measure that accurately identifies real change. Consequently, responsiveness and reproducibility are directly connected to each other[53, 54]. Since a trend being above the mean difference was detected, a linear regression was completed to search for proportional bias. Bland and Altman plot identifies two outliers likely due to the variability of the measurement. One of the outliers had bad visual acuity (confounding variable). No proportional bias was found and the LOA (± 2.3) was less than the MCID of 5.7 points, so ARAT is capable of detecting clinically relevant changes.

However, the findings in this study must be interpreted with caution. Using a videotape for analysis automatically reduces measurement variability, eliminating the need for the patient to perform the ARAT multiple times that may cause fatigue[33]. Non-probability sampling is vulnerable to selection bias, the PT were not blinded to the video, but the allowed time interval was considered adequate for both data memorisation and sample maturation. Despite the robust statistical methods used, the small a-value (0.01) increased the confidence for significant results. Three researchers with previous experience scored

the ARAT, which may lead to biased results, and captured only the affected side of the patient. Another potential limitation may be instrumentation. The results would have yielded a more precise estimation of reliability if the two outliers have been excluded. Assuming this sample with a mean age of 32 ± 6.81 years and a mean experience of 7.55 ± 7.4 years to be representative of the whole population, then the results can be generalised to a wider population. Despite these limitations, the findings of this study suggest that PT should use the ARAT for the examination of UL function in moderate chronic stroke.

The ARAT demonstrates good test-retest reliability using statistical analysis with Spearman's rank order correlation coefficient, Bland and Altman plots and linear regression. These results illustrate good test-retest and are consistent. The included sample accurately reflects a larger population with a mean age of 32 ± 7 years, mean experience of 7.55 ± 7.4 years and six different nationalities. Therefore, the results are representative internationally, and PT should use and score the protocol. However, the reported results should be interpreted with caution due to the small sample size and the selected type of sampling.

Conflict of interest

The author declares that he has no competing interests.

ACKNOWLEDGEMENT

We take this opportunity to thank all participants who took part in this study.

REFERENCES

1) World Health Organization: WHO STEPS Stroke Manual: The WHO STEPwise approach to stroke surveillance, 2005.
2) Hankey GJ: Stroke treatment and prevention: an evidence-based approach. Cambridge: Cambridge University, 2005.
3) Stroke Association: State of the nation-stroke statistics. https://www.stroke.org.uk/resources/state-nation-stroke-statistics. (Accessed Dec. 31, 2016)
4) Saka O, McGuire A, Wolfe C: Cost of stroke in the United Kingdom. Age Ageing, 2009, 38: 27–32. [Medline] [CrossRef]
5) Shumway-Cook A, Woollacott M: Motor control, 4th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2012.
6) Claflin ES, Krishnan C, Khot SP: Emerging treatments for motor rehabilitation after stroke. Neurohospitalist, 2015, 5: 77–88. [Medline] [CrossRef]
7) Nakayama H, Jørgensen HS, Raaschou HO, et al.: Recovery of upper extremity function in stroke patients: the Copenhagen Stroke Study. Arch Phys Med Rehabil, 1994, 75: 394–398. [Medline] [CrossRef]
8) Kwakkel G, Kollen BJ, Wagenaar RC: Long term effects of intensity of upper and lower limb training after stroke: a randomised trial. J Neurol Neurosurg Psychiatry, 2002, 72: 473–479. [Medline]
9) Broeks JG, Lankhorst GJ, Rumping K, et al.: The long-term outcome of arm function after stroke: results of a follow-up study. Disabil Rehabil, 1999, 21: 357–364. [Medline] [CrossRef]
10) Houwink A, Nijland RH, Geurts AC, et al.: Functional recovery of the paretic upper limb after stroke: who regains hand capacity? Arch Phys Med Rehabil, 2013, 94: 839–844. [Medline] [CrossRef]
11) Kwakkel G, Kollen BJ, van der Grond J, et al.: Probability of regaining dexterity in the flaccid upper limb: impact of severity of paresis and time since onset in acute stroke. Stroke, 2003, 34: 2181–2186. [Medline] [CrossRef]
12) Sheikh K: Disability scales: assessment of reliability. Arch Phys Med Rehabil, 1986, 67: 245–249. [Medline]
13) van der Lee JH, Wagenaar RC, Lankhorst GJ, et al.: Forced use of the upper extremity in chronic stroke patients: results from a single-blind randomized clinical trial. Stroke, 1999, 30: 2369–2375. [Medline] [CrossRef]
14) Hsieh CL, Hsueh IP, Chiang FM, et al.: Inter-rater reliability and validity of the action research arm test in stroke patients. Age Ageing, 1998, 27: 107–113. [Medline] [CrossRef]
15) Koh CL, Hsueh IP, Wang WC, et al.: Validation of the action research arm test using item response theory in patients after stroke. J Rehabil Med, 2006, 38: 375–380. [Medline] [CrossRef]
16) Lyle RC: A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res, 1981, 4: 483–492. [Medline] [CrossRef]
17) Platz T, Pinkowski C, van Wijck F, et al.: Reliability and validity of arm function assessment with standardized guidelines for the Fugl-Meyer Test, Action Research Arm Test and Box and Block Test: a multicentre study. Clin Rehabil, 2005, 19: 404–411. [Medline] [CrossRef]
18) Carroll D: A quantitative test of upper extremity function. J Chronic Dis, 1965, 18: 479–491. [Medline] [CrossRef]
19) Chanubol R, Wongphaet P, Ot NC, et al.: Correlation between the action research arm test and the box and block test of upper extremity function in stroke patients. J Med Assoc Thai, 2012, 95: 590–597. [Medline]
20) Wagenaar RC, Meijer OG, van Wieringen PC, et al.: The functional recovery of stroke: a comparison between neuro-developmental treatment and the Brunnstrom method. Scand J Rehabil Med, 1990, 22: 1–8. [Medline]
21) Van der Lee JH, De Groot V, Beckerman H, et al.: The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil, 2001, 82: 14–19. [Medline] [CrossRef]
22) van der Lee JH, Roorda LD, Beckerman H, et al.: Improving the Action Research Arm test: a unidimensional hierarchical scale. Clin Rehabil, 2002, 16: 646–653. [Medline] [CrossRef]
23) Yozbatiran N, Der-Yeghiaian L, Cramer SC: A standardized approach to performing the action research arm test. Neurorehabil Neural Repair, 2008, 22: 78–90. [Medline] [CrossRef]

24) Lin JH, Hsu MJ, Sheu CF, et al.: Psychometric comparisons of 4 measures for assessing upper-extremity function in people with stroke. Phys Ther, 2009, 89: 840–850. [Medline] [CrossRef]
25) Party IS: National clinical guideline for stroke. 2012.
26) Van Peppen RP, Kwakkel G, van der Wel BH, et al.: KNGF clinical practice guideline for physical therapy in patients with stroke. Review of the evidence. Nederlands Tijdschrift voor Fysiotherapie, 2004, 114.
27) Salter K, Campbell N, Richardson M, et al.: Outcome Measures in stroke rehabilitation, 16th Ed. Ontario: Canadian Stroke Network; 2013.
28) Winstein CJ, Stein J, Arena R, et al. American Heart Association Stroke Council, Council on Cardiovascular and Stroke Nursing, Council on Clinical Cardiology, Council on Quality of Care and Outcomes Research: Guidelines for adult stroke rehabilitation and recovery: a guideline for healthcare professionals From the American Heart Association/American Stroke Association. Stroke, 2016, 47: e98–e169. [Medline] [CrossRef]
29) Hicks CM: Practical research methods for physiotherapists, 3rd ed. Edinburgh: Churchill Livingstone, 1988.
30) Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1986, 1: 307–310. [Medline] [CrossRef]
31) Brunnekreef JJ, van Uden CJ, van Moorsel S, et al.: Reliability of videotaped observational gait analysis in patients with orthopedic impairments. BMC Musculoskelet Disord, 2005, 6: 17. [Medline] [CrossRef]
32) Eastlack ME, Arvidson J, Snyder-Mackler L, et al.: Interrater reliability of videotaped observational gait-analysis assessments. Phys Ther, 1991, 71: 465–472. [Medline] [CrossRef]
33) Wellmon R, Degano A, Rubertone JA, et al.: Interrater and intrarater reliability and minimal detectable change of the Wisconsin Gait Scale when used to examine videotaped gait in individuals post-stroke. Arch Physiother, 2015, 5: 11. [Medline] [CrossRef]
34) Shrum W, Duque R, Brown T: Digital video as research practice: methodology for the millennium. J Res Pract, 2005, 1: 4.
35) 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: 539–549. [Medline] [CrossRef]
36) Salthouse TA, Tucker-Drob EM: Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology, 2008, 22: 800–811. [Medline] [CrossRef]
37) Price CP: Psychology research methods: core skills and concepts. Creative Commons, California State University, Michael Boezi, 2012.
38) Gupta SK: Intention-to-treat concept: a review. Perspect Clin Res, 2011, 2: 109–112. [Medline] [CrossRef]
39) McLeod SA: Independent, dependent and extraneous variables. https://www.simplypsychology.org/variables.html. 2008. (Accessed Jul. 13, 2017)
40) Coutts F: Gait analysis in the therapeutic environment. Man Ther, 1999, 4: 2–10. [Medline] [CrossRef]
41) Bland JM, Altman DG: Measuring agreement in method comparison studies. Stat Methods Med Res, 1999, 8: 135–160. [Medline] [CrossRef]
42) Bland JM, Altman DG: Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet, 1995, 346: 1085–1087. [Medline] [CrossRef]
43) McCrum-Gardner E: Which is the correct statistical test to use? Br J Oral Maxillofac Surg, 2008, 46: 38–41. [Medline] [CrossRef]
44) van der Lee JH, Beckerman H, Lankhorst GJ, et al.: The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. J Rehabil Med, 2001, 33: 110–113. [Medline] [CrossRef]
45) Wagner A, Gossauer E, Moosmann C, et al.: Thermal comfort and workplace occupant satisfaction—Results of field studies in German low energy office buildings. Energy Build, 2007, 39: 758–769. [CrossRef]
46) McLeod SA: Observation methods. www.simplypsychology.org/observation.html. (Accessed Jul. 13, 2017)
47) Washington Post.: Fasting during the month of Ramadan can cause headaches, fatigue and dehydration. https://www.washingtonpost.com/national/health-science/fasting-during-the-month-of-ramadan-can-cause-headaches-fatigue-anddehydration/2011/08/15/gIQAhuV1WJ_story.html?utm_term=.5e29b2258739. (Accessed Jul. 13, 2017)
48) Lærd Dissertation: Extraneous and confounding variables. http://dissertation.laerd.com/extraneous-and-confounding-variables.php. (Accessed Aug. 11, 2017)
49) Paiva CE, Barroso EM, Carneseca EC, et al.: A critical analysis of test-retest reliability in instrument validation studies of cancer patients under palliative care: a systematic review. BMC Med Res Methodol, 2014, 14: 8. [Medline] [CrossRef]
50) Neuman WR: Beyond HDTV: exploring subjective responces to very high definition television. Media Laboratory, Massachusetts Institute of Technology, 1990.
51) Lombard M, Ditton TB, Grabe ME, et al.: The role of screen size in viewer responses to television fare. Commun Rep, 1997, 10: 95–106. [CrossRef]
52) Nijland R, van Wegen E, Verbunt J, et al.: A comparison of two validated tests for upper limb function after stroke: The Wolf Motor Function Test and the Action Research Arm Test. J Rehabil Med, 2010, 42: 694–696. [Medline] [CrossRef]
53) Vaz S, Falkmer T, Passmore AE, et al.: The case for using the repeatability coefficient when calculating test-retest reliability. PLoS One, 2013, 8: e73990. [Medline] [CrossRef]
54) Beckerman H, Roebroeck ME, Lankhorst GJ, et al.: Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res, 2001, 10: 571–578. [Medline] [CrossRef]
55) Hébert R, Spiegelhalter DJ, Brayne C: Setting the minimal metrically detectable change on disability rating scales. Arch Phys Med Rehabil, 1997, 78: 1305–1308. [Medline] [CrossRef]