Interrater and intrarater reliability of the single arm military press (SAMP) test for upper limb function in patients with non-specific neck pain

ALRENI, ASE <http://orcid.org/0000-0001-8402-2415>, ABOALMATY, HRA, DE HERTOGH, W <http://orcid.org/0000-0001-8856-2507> and MCLEAN, Sionnadh <http://orcid.org/0000-0002-9307-8565>

Citation:

ALRENI, ASE, ABOALMATY, HRA, DE HERTOGH, W and MCLEAN, Sionnadh (2021). Interrater and intrarater reliability of the single arm military press (SAMP) test for upper limb function in patients with non-specific neck pain. Musculoskeletal Science and Practice, 55, p. 102428. [Article]

Interrater and intrarater reliability of the Single Arm Military Press (SAMP) test for upper limb function in patients with non-specific neck pain

Lead & Corresponding Author:
Ahmad Salah Eldin Alreni
PhD & Post-Doctoral Researcher
Faculty of Medicine and Health Science
University of Antwerp
Antwerp, Belgium.
Email: Ahmad.Alreni@uantwerpen.be Email
Email: ahmadalreni@hotmail.co.uk

Co-Authors:
1. Heba Rohy Abdo Aboalmaty
Professor and Senior Lecturer
Department of Sports Training and Kinesiology
Tanta University
Tanta, Egypt.
Email: hebaabdoaboalmaty@outlook.com

2. Willem De Hertogh
Professor and Senior Lecturer
Faculty of Medicine and Health Science
University of Antwerp
Antwerp, Belgium.
Email: willem.dehertogh@uantwerpen.be

3. Sionnadh Mairi McLean
Senior Lecturer and Reader in Physiotherapy
Centre for Health and Social Care Research Sheffield Hallam University
Sheffield, UK
Email: s.mclean@shu.ac.uk

Ethics
Ethical approval for this study was obtained from Sheffield Hallam University Research Committee (SHUREC) and Tanta University, Egypt (reference 2014- 5/ HWB-HSC-18).

PII: S2468-7812(21)00112-0
DOI: https://doi.org/10.1016/j.msksp.2021.102428
Reference: MSKSP 102428
To appear in*: Musculoskeletal Science and Practice*

Received Date: 5 January 2021
Revised Date: 4 May 2021
Accepted Date: 7 July 2021

Abstract

Background

Performance measures that assess the upper limb disability (ULD) in patients with neck pain can provide useful information for making clinical decisions regarding the optimal management of those patients. The Single Arm Military Press (SAMP) test is a performance based ULD measure developed specifically for populations with neck pain. In this test, patients are asked to lift a 1kg weight repetitively overhead for 30 seconds with repetitions counting as the score. Whilst the test has been shown to be acceptable and feasible for use by clinicians and patients, its reliability in a patient group is still unknown.

Objective

To assess the interrater, intrarater reliability and measurement error of the SAMP test in patients with non-specific neck pain (NSNP).

Methods

A total of 210 patients with NSNP and 81 healthy subjects were recruited for this study. The Disabilities of the Arm, Shoulder and Hand (DASH) and the Neck Disability Index (NDI) were assessed at baseline to ensure eligibility of the participants. The SAMP test was assessed at baseline and repeated 4 to 7 days later. A VAS symptom score was used to establish the stability of the participants across time. Interrater, intrarater reliability and measurement error were evaluated using Interclass Correlation Coefficient (ICC2,1) and the standard error of measurement (SEM).

Results

The ICCs for interrater and intrarater reliability for the SAMP test ranged from 0.993 to 0.996 in the patient group. The SEM was ≤ 1 and smaller than the Smallest Detectable Change (SDC) and Bland-Altman plot indicated that the test is accurate.

Conclusion

The almost perfect interrater and intrarater reliability and low levels of measurement error indicate that the 1kg SAMP test has potential for evaluating upper limb functional capacity in female patients with NSNP. Before the test can be fully recommended, further studies are required to evaluate the validity and responsiveness of the SAMP test in population with NSNP and other neck disorders.

Keywords: Neck Pain, Upper Limb Function, Disability, Outcome Measure, Reliability.

Funding acknowledgements: This research was unfunded

1. Background

Non-specific neck pain (NSNP) is defined as pain or discomfort in the neck and/or shoulder girdle with or without pain referred to the arms. Symptoms vary with physical activity and over time and frequently no disease or pathoanatomical cause can be identified. The cause is usually multifactorial and may include poor posture, neck strain, sporting and occupational activities, anxiety and depression [National Institute for Health and Care Excellence (NICE), 2018). It is common and frequently causes pain, motor weakness and impairment in the neck and upper limbs [Walker-Bone et al., 2002]. NSNP can have a substantial effect on quality of life, work absenteeism and loss of productive capacity [Walker-Bone et al., 2004; Huisstede et al., 2006]. Upper limb disability (ULD), which is defined here as the limitation an individual may have when performing physical activity using the upper limbs such as carrying, lifting and overhead activity [World Health Organisation (WHO), 2001], can arise from a spectrum of clinical conditions including NSNP [Frank et al., 2005; McLean et al., 2007]. ULD and NSNP often co-exist and more than 80% of patients with NSNP report difficulties with daily activities that involve functional loading of the upper limbs [Osborn and Jull, 2013; McLean et al., 2011]. The mechanisms which cause these conditions to co-exist are not clear but may relate to the mechanical attachment between the neck and upper limb via skeletal, muscular, neural structures or through psychological mechanisms such as low pain self-efficacy [Lee et al., 2015; Ahmed et al., 2019]. Consequently, a thorough assessment of NSNP will include the use of a suitable upper limb functional capacity measurement instrument to quantify any co-existing ULD. This information may then be used to help develop an upper limb rehabilitation plan as indicated [Osborn and Jull, 2013; McLean et al., 2011].

A wide range of ULD measurement instruments have been proposed for patients with neck pain [Alreni et al., 2017]. Some are Patient-Reported Outcome Measures (PROMs), whereas others are Performance-Based Outcome Measures (PBOMs) [Stock et al., 2003; Huisstede et al., 2009; Mehta et al., 2010; Lomond et al., 2011]. The Single Arm Military Press (SAMP) test is a PBOM designed to measure ULD in patients with NSNP [McLean et al., 2010]. It is a simple test that can be efficiently administered by clinicians with varying experience in any setting using readily available and inexpensive equipment (one dumbbell) in less than two minutes [Alreni et al., 2020]. Furthermore, since it is a PBOM, the SAMP test has the theoretical advantages of better reliability, greater sensitivity to change and lower vulnerability to external variance e.g. culture, cognition, language and level of education [Curb et al., 2006; Pinheiro et al., 2016].

Reliability is an essential requirement of all outcome measures; poor reliability with a high level of measurement error would limit the extent to which the findings of an instrument can be trusted. Consequently, this would reduce the usefulness and the clinical utility of that instrument [de vet et al., 2011]. Reliability concerns the extent to which the measurement of stable patient can be reproduced when the same instrument is used at different moments, in different conditions, by different examiners or by the same examiner at different times [Streiner et al., 2015].

Content validity of the 1kg SAMP test was established and preliminary investigation suggested that the test was deemed a feasible and suitable measure of ULD by patients with NSNP and clinicians [Alreni et al., 2020]. However, it is not known whether the 1kg SAMP test is reliable for use by clinicians for patients with NSNP. Consequently, the aim of this study was to investigate the interrater and intrarater reliability and measurement error of the

SAMP test in female patients with NSNP and healthy subjects. Based on previous studies [McLean et al., 2010], we hypothesised that (1) the ICC would be ($\geq$ 0.90) for interrater reliability, (2) the ICC would be ($\geq$ 0.90) for intrarater reliability, (3) the Standard Error of Measurement (SEM) would be ($\leq$ 1) and smaller than the Smallest Detectable Change (SDC) for measurement error.

2. Methods

2.1 Study design

This large-scale reliability study was conducted as a part of a larger research project, which explored the clinical management and measurement of NSNP and its associated disabilities. The study investigated the reliability of the 1kg SAMP test in female patients with NSNP. This study was conducted in accordance with the "**CO**nsensus-based **S**tandards for the selection of health **M**easurement **IN**struments" (COSMIN) checklist recommendations [Mokkink et al., 2010; Terwee et al., 2012, 2018] and reported in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational-cohort studies [von Elm et al., 2007].

2.2 Study sample and recruitment

Participants were recruited from the Rheumatology and Physical Therapy Department at the TUTH and female patients were included if they: (1) had acute, sub-acute or chronic NSNP with or without referred symptoms into the upper limbs; (2) were at least 18 years of age, (3) were able to travel independently to the testing hospital; and (4) scored at least 10 (out of 100) in the Neck and Disability Index (NDI) and 26 (out of 100) in the Disability of the Arm, Shoulder and Hand (DASH) questionnaires. Patients were excluded if they had: (1) any potentially serious conditions (e.g. systemic disease, progressive or worsening neurological disorders, inflammatory conditions or major trauma), (2) a neck condition that requires urgent treatment or (3) previous traumatic injury to the neck (e.g. Whiplash Associated Disorder 'WAD', Cervical Radiculopathy), upper limbs and/or shoulder that resulted in current or prolonged disability. Meanwhile, healthy subjects were recruited from the general population by announcement via social network sites, personal networking and posters/flyers within Tanta University and Tanta city centre. Healthy subjects were included if they: (1) were at least 18 years of age, (2) had no history of head/neck/upper limb trauma; (3) had no current or recent neck or upper limb problems (within the last 3 months); and (4) were females. Healthy subjects were frequency matched with prospective patient participants regarding age, weight and height.

A list of patients was obtained from TUTH and potential participants, patients and healthy subjects, were invited to attend a further assessment and two testing sessions, if initial eligibility was confirmed. In the first testing occasion, participants completed the NDI and DASH questionnaires in order to confirm the presence of NSNP and ULD [Vernon and Mior, 1991; Hudak et al., 1996]. This was followed by a subjective examination, in which standardised clinical questions were used and the neck and upper limb symptoms severity was measured using the Visual Analogue Scale (VAS) 0-10 [Bond and Lader, 1974]. VAS scores were used to investigate whether subjects remained stable between testing sessions. Participants that met the eligibility criterion were asked to sign a consent form and were then allocated to the first SAMP testing.

2.3 Outcome measures

The SAMP test and two PROMs (NDI and DASH) alongside the VAS were used in this study. A brief description of each instrument used are given below.

2.3.1 The SAMP test

The SAMP test is a PBOM which measures the strength and endurance of the upper limb in an overhead activity, with the expectation that an individual's ability to lift a dumbbell and sustain repetitive overhead activity within 30 seconds would discriminate between healthy subjects and patient groups with varying degrees of ULD. The test is conducted with the patient in a standing position with their feet positioned at shoulder width. The patient is asked to hold a 1kg dumbbell and lift it, using their dominant hand, to shoulder level (see Figure 1A). The patient is requested to raise their hand with the dumbbell directly overhead by extending the elbow (see Figure 1B) and repeat this process as fast and as frequently as possible for 30 seconds [McLean et al., 2010]. The SAMP test is a quick and easy assessment requiring the use of readily available and inexpensive equipment (1kg dumbbell) for less than 2 minutes. Scoring the SAMP test involves counting the number of correctly performed repetitions completed within 30 seconds. The feasibility and content validity of the 1kg SAMP test was previously established in female patients with NSNP [Alreni et al., 2020].

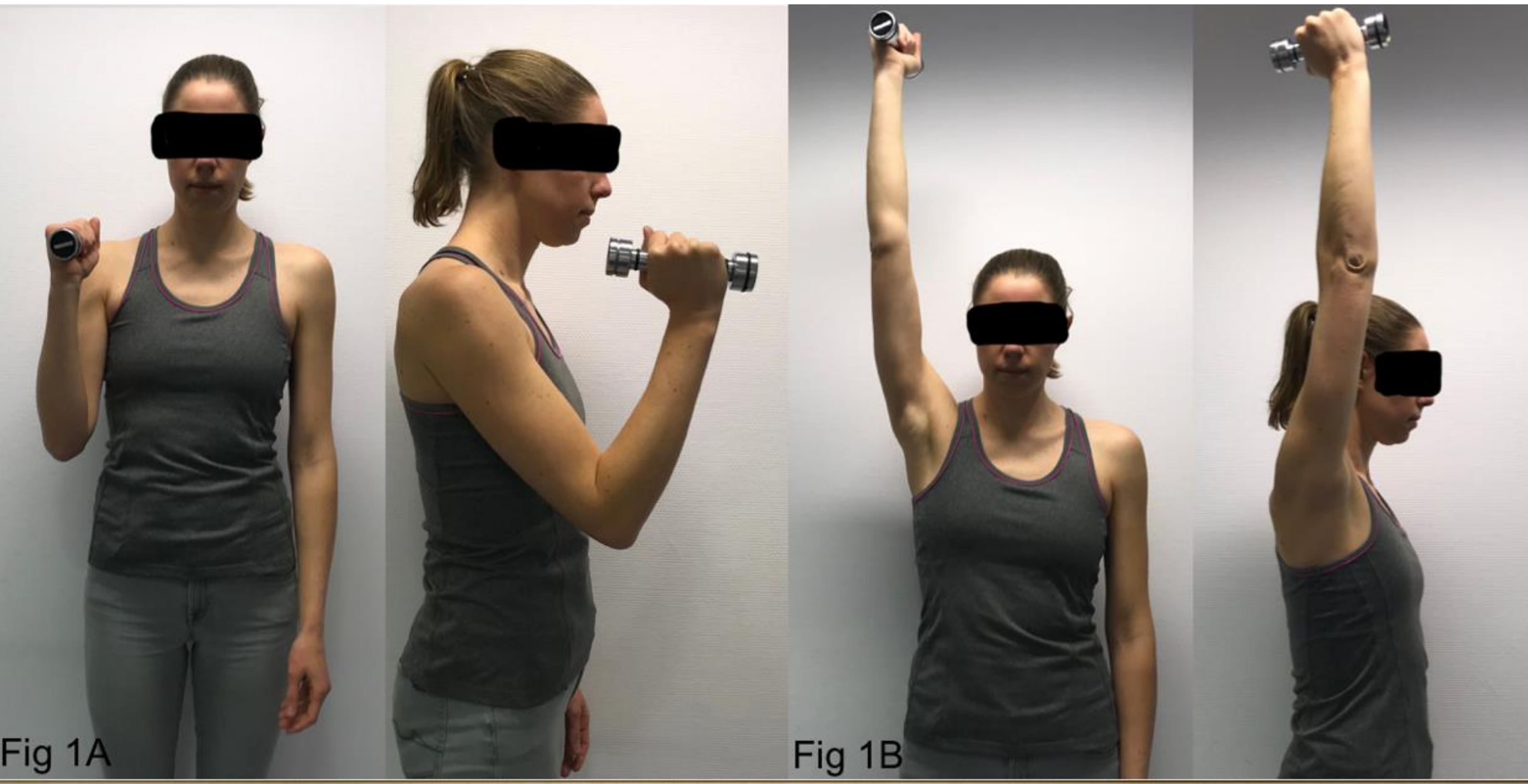

Figure 1 SAMP Test Protocol

2.3.2 The Neck Disability Index (NDI) Questionnaire

The NDI is a standardised PROM developed and extensively validated to measure a patient's disability due to neck pain [Vernon and Mior, 1991]. It has 10 items; 7 items related to activities of daily living, 2 items related to pain, and 1 item related to concentration. Each item is scored from 0-5 and a total score is expressed as percentage score, with higher scores indicating greater disability. The NDI has been found to be reliable, valid and responsive in numerous patient populations, including patients with acute and chronic NSNP, as well as those with neck pain due to whiplash-associated disorders and cervical radiculopathy [MacDermid et al., 2009; Bobos et al., 2018]. The NDI was translated and culturally-adapted in Arabic and its reliability and validity were determined in Arabic-speaking patients with neck pain [Shaheen et al., 2013].

2.3.3 The Disability of the Arm, Shoulder and Hand (DASH) questionnaire

The DASH is a standardised PROM developed primarily to evaluate the upper limb symptoms as a single functional unit [Hudak et al., 1996]. The DASH uses 30-items related to difficulty when performing activities which use the upper limb (arm, shoulder and hand). The dimension physical function comprised 21-items, pain 5-items and emotional/social function comprised 4-items. Each item is scored on a 1-5 scale. A total score is calculated by summing item scores and transforming them into a score from 0-100 where 0 equals no disability and 100 equals the most severe disability. Since its development, the measurement properties of the DASH questionnaire have been extensively evaluated for a variety of upper limb conditions, translated and cross-culturally adapted into over 40 different languages, including Arabic-Speaking patient populations [Beaton et al., 2001; Bot et al., 2004; Roy et al., 2009; Alotaibi et al., 2010]. The DASH was also validated to measure ULD in patients with NSNP [Huisstede et al., 2009; Mehta et al., 2010].

2.3.4 Visual Analogue Scale (VAS)

The VAS with 0-10 response categories, where 0 indicates no symptoms and 10 indicates the worst possible symptoms, was used to measure the NSNP and ULD severity for patient participants pre-testing in the two sessions. The VAS (0-10) scale has been extensively validated and found to be reliable, valid and responsive in measuring symptom severity in patients with various musculoskeletal conditions including neck and upper limb symptoms [Bond and Lader, 1974; McCormack et al., 1988; Wewers and Lowe, 1990; Jaeschke et al., 1990].

2.4 Testing procedure and data collection

A total of four examiners, physicians, with at least 3-years of experience of working with musculoskeletal patients were involved in the data collection. Prior to the first testing, all examiners attended a 30-minute practical training and information session delivered by the lead author. This covered the purpose of the study, a brief outline of the SAMP test description and practical application, a standardised demonstration of the warm-up and the SAMP scoring system. Two pairs of examiners were then formed for testing the participants, each pair consisted of an A (rater) and B (co-assessor).

The SAMP testing was conducted on two different occasions with 4-7 days interval [de Vet et al., 2011], by a pair of examiners independently but simultaneously for each patient. Meanwhile healthy subjects were tested by one examiner only. The testing started with a brief warm-up followed by description and demonstration of the SAMP procedure by the rater. The participant performed 2-3 reps of the test to ensure correct performance of the technique, then after a short rest was asked to perform the SAMP test for 30 seconds. A data collection sheet was completed by a rater and co-assessor for each patient, recording the SAMP score. On the second occasion, the neck and upper limb symptoms severity were remeasured using VAS to ensure that participants remained stable between testing sessions. Participants were SAMP tested again by the same examiners, though they had swapped their rater and co-assessor. The second occasion was in the same venue at a similar time of the day. Participants were discharged following the conclusion of the second test.

2.5 Sample size and data analysis

A sample size of more than 100 is recommended for reliability testing based on the COSMIN checklist recommendations in order to obtain a Confidence Interval (CI) > 0.90 around Interclass Correlation Coefficient (ICC) of 0.90-0.95 [de Vet et al., 2011]. A larger than required sample was recruited for further subsequent analysis conducted in additional validity studies.

Data were transferred into Excel and then to SPSS (IBM SPSS Statistical Software, version 24.0) for further analysis. Descriptive statistics (mean, standard deviation, standard error of mean, and 95% confidence interval) were computed for the SAMP test for patient participants and healthy subjects. The Shapiro-Wilk test was used to check for normal distribution of the primary data and for suitability for analysis using parametric statistics [Cramer, 1998; Doane and Seward, 2011; Razali and Wah, 2011].

The reliability (interrater and intrarater) and measurement error of the SAMP test were evaluated using the COSMIN checklist recommendations [Mokkink et al., 2010; Terwee et al., 2012, 2018]. *Interrater reliability* concerns the extent to which scores of the same subject are unchanged when using the same instrument on the same occasion by different examiners. It was calculated for the two independent but simultaneous examiners (rater and co-assessor) across two testing occasions. *Intrarater reliability* concerns the extent to which scores for the same subject are unchanged for repeated measurement by the same examiner on different occasions with appropriate time interval. It was calculated for examiner A and B across two testing occasions. $ICC_{2,1}$ with 95% Confidence Interval (CI) was used to calculate the interrater and intrarater reliability of the SAMP test [de Vet et al., 2011]. The $ICC_{2,1}$ scoring interval for interpretation was categorised as follows: $0.000 > ICC_{2,1}$ as poor; $0.00 < ICC_{2,1} \leq 0.20$ as slight; $0.21 < ICC_{2,1} \leq 0.40$ as fair; $0.41 < ICC_{2,1} \leq 0.59$ as moderate; $0.60 < ICC_{2,1} \leq 0.79$ as substantial (high); and $0.80 < ICC_{2,1} \leq 1.00$ as almost perfect (very high) [Landis and Koch, 1977].

Measurement error is the systematic and random error of a subject's score, which cannot be attributed to a true change in the construct being measured. It is the absolute measurement error over repeated measurements and expressed by the Standard Error of Measurement (SEM). It is the standard deviation of the errors of measurement that are associated with the instrument's scores and is equal to the square root of the error variance ($\sqrt{}\ \sigma^2$ error) [de Vet et al., 2006]. SEM was derived using a two-way analysis of variance (ANOVA) $ICC_{2,1}$ [McGraw and Wongs, 1996]. The Smallest Detectable Change (SDC) represents the minimal change that a patient must show on an instrument to ensure that the observed change is true and not related to measurement errors [Bland and Altman, 1996]. It was calculated using the formula: ($SDC = 1.96 \times \sqrt{2} \times SEM$). Low levels of SEM which should be smaller than the SDC indicate high levels of score accuracy [Vincent and Weir, 2012]. The Bland-Altman 95% limit of agreement (LoA), which calculates the measurement error was also reported. Bland-Altman plot was used to examine the agreement between the scores obtained with SAMP test administered on two occasions. The 95% limits of agreement between the scores was calculated using the formula: (95% CI Upper Limits = (SD ×1.96) + Mean; 95% CI Lower Limits = Mean − (SD × 1.96). The plot is in number of repetitions, which is the unit of measurement of the SAMP test [Bland and Altman, 2007].

3. Results

The flow of patient participants and controls through each stage is presented in Figure 2. A list of 300 patients was obtained from the Rheumatology and Physical Therapy Department TUTH. Following a phone screening, 250 patients were eligible and willing to voluntarily participate in the study. Thirty patients were ineligible, 20 patients declined and 40 patients did not turn-up for their first assessment and testing session. Following the subjective examination in the first session, 210 patient participants and 81 healthy subjects were found eligible for SAMP testing, consented in writing and participated in testing at timepoint 1. Participants from timepoint 1 were retained and participated in testing at timepoint 2 (no drop-out).

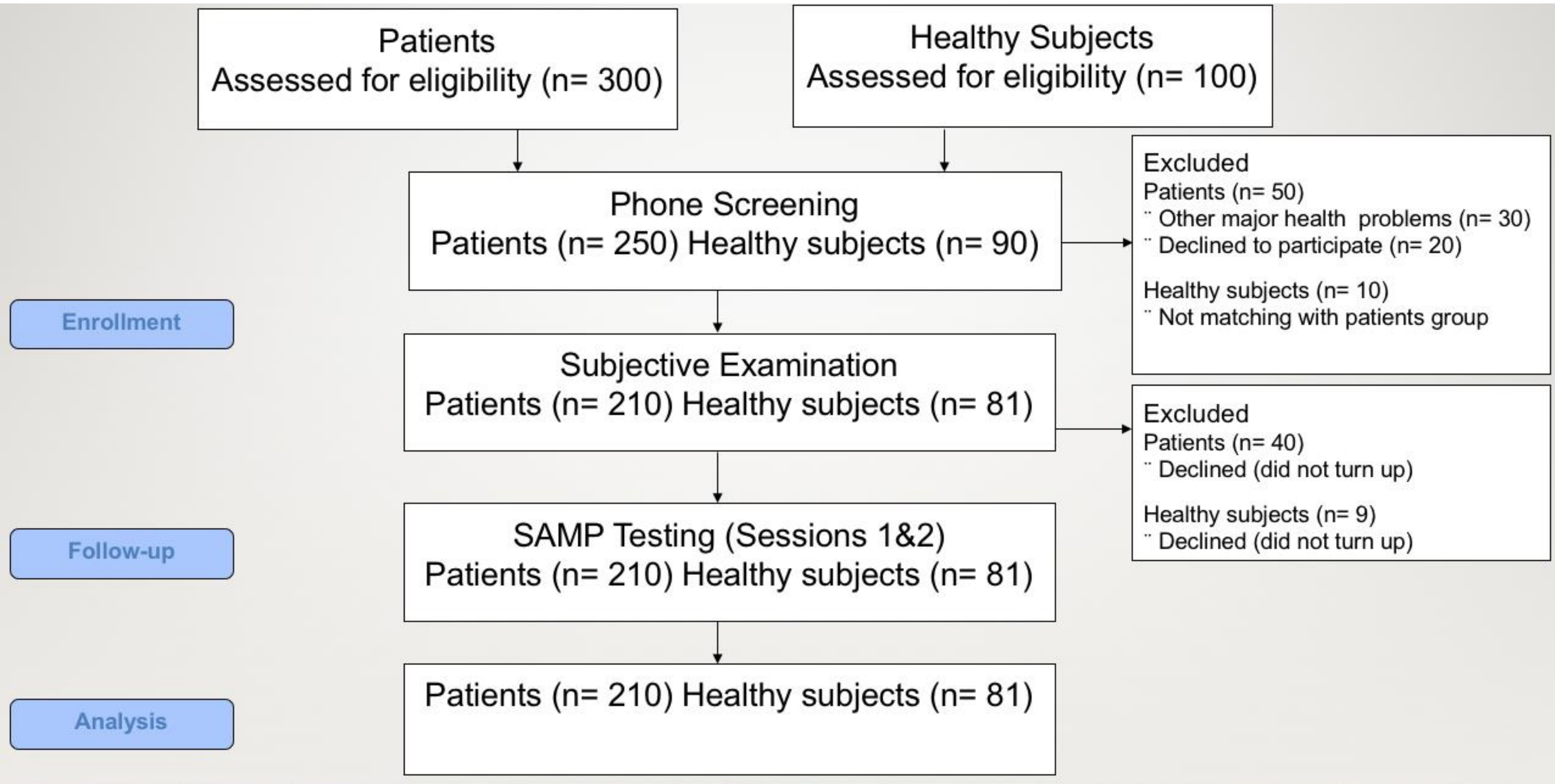
Patients
Assessed for eligibility (n= 300)
Healthy Subjects
Assessed for eligibility (n= 100)
Phone Screening
Patients (n= 250) Healthy subjects (n= 90)
Excluded
Patients (n= 50)
¨ Other major health problems (n= 30)
¨ Declined to participate (n= 20)
Healthy subjects (n= 10)
¨ Not matching with patients group
Enrollment
Subjective Examination
Patients (n= 210) Healthy subjects (n= 81)
Excluded
Patients (n= 40)
¨ Declined (did not turn up)
Healthy subjects (n= 9)
¨ Declined (did not turn up)
Follow-up
SAMP Testing (Sessions 1&2)
Patients (n= 210) Healthy subjects (n= 81)
Analysis
Patients (n= 210) Healthy subjects (n= 81)

Figure 2 Flow-Chart of Participants in the Reliability Study

3.1 Demographic characteristics and baseline data

The demographic characteristics and baseline data of participants in this study are summarised in Table 1. The age (mean and standard deviation) of the 210 patient participants was 40.4 ± 4.9 years and the 81 healthy subjects was 36.54 ± 4.9 years. There were no significant differences between patients and healthy subjects on age, weight or height, and this indicates that these groups were well matched. However, as expected, there were clear and substantial differences (*P*<0.05) between these groups regarding the severity of NSNP and ULD in all measures. In the second testing session, patients group reported slight, but non-significant, improvements on their neck and upper limb symptoms scores (VAS), indicating that these groups were stable between the testing sessions.

Table 1
Participants characteristics at baseline

Variables	Healthy subjects (n=81)	Patients (n=210)
Age in years: (mean - SD)	36.54 - 4.9	40.41 - 4.9
Weight (kg): Frequencies (%)		
75-80	4 (5)	10 (4.8)
81-85	4 (5)	11 (5)
86-90	23 (28.3)	60 (28.6)
91+	50 (61.7)	129 (61.4)
Height (cm): Frequencies (%)		
155-160	25 (30.8)	65 (31)
161-165	51 (63)	130 (62)
166+	5 (6.2)	15 (7)
Neck VAS scores (0-10): (mean – SD)		
Timepoint 1	0	4.40 – 1.48
Timepoint 2	0	4.04 – 1.45
Upper Limb VAS scores (0-10): (mean – SD)		
Timepoint 1	0	2.45 – 1.41
Timepoint 2	0	2.28 – 1.29
NDI Scores (0-100): (mean – SD)	4.63 – 0.798	43.38 – 14.47
DASH Scores (0-100): (mean – SD)	4.04 – 1.17	31.66 – 16.42

SD: Standard Deviation, NDI: Neck Disability Index, DASH: Disability of the Arm, Shoulder and Hand.

3.2 Descriptive statistics of the SAMP test

Participants in the study were SAMP tested on two occasions approximately 1 week apart. Descriptive statistics of the SAMP scores for patients and healthy subjects are summarised in Table 2. As expected the scores on the SAMP test were significantly lower (P<0.0001) for the patient group versus the healthy subjects on both testing occasions (17.90 ± 6.167 versus 35.23 ± 3.348 on timepoint 1; 17.99 ± 6.140 versus 35.07 ±2.692 on timepoint 2).

Table 2
Descriptive statistics of the SAMP test

Variables	Healthy subjects (n=81)				Patients (n=210)				*p*
	Mean	SD 95% CI	Min	Max	Mean	SD 95% CI	Min	Max	
SAMP Score: Timepoint 1									0.0001
Examiner A (Rater) [a]	35.23	3.348	28	39	17.90	6.167	3	30	
Examiner B (Co-Assessor) [b]					17.71	6.023	3	29	
SAMP Score: Timepoint 2									0.0001
Examiner A (Co-Assessor) [b]	35.07	2.692	29	40	17.99	6.140	3	30	
Examiner B (Rater) [a]					18.00	6.116	3	30	

SD: Standard Deviation, CI: Confidence Interval, SAMP: The Single Arm Military Press, Min: Minimum, Max: Maximum, P: P Value, [a] Rater: Examiner who conducted the test and completed data collection sheet, [b] Co-Assessor: Examiner who only completed data collection sheet independently but simultaneously with the rater.

3.3 Interrater and intrarater reliability

The ICC 2,1, SEM, SDC statistics and their 95% Confidence Interval (lower bound and upper bound) for interrater, intrarater reliability and agreement are summarised in Table 3. The ICCs exceeded 0.90 indicating almost perfect interrater and intrarater reliability for the SAMP test [Landis and Koch, 1977]. The SEM was ≤ 1and smaller than the SDC indicating high levels of score accuracy and agreement for the SAMP test [McGraw and Wongs, 1996; Bland and Altman, 1996; de Vet et al., 2006; Vincent and Weir, 2012].

Table 3
Reliability Coefficient and SEM and SDC and their 95% CI for the SAMP test

Variables	Patients (n=210)			
	$ICC_{2,1}$	95% CI (LB–UB)	SEM	SDC
Interrater reliability				
Timepoint 1	0.995	0.993 - 0.996	0.42	1.2
Timepoint 2	0.997	0.996 - 0.998	0.35	1.0
Intrarater reliability				
Examiner A	0.997	0.996 - 0.998	0.35	1.0
Examiner B	0.994	0.998 - 0.996	0.44	1.2

SEM: Standard Error of Measurement, SDC: Smallest Detectable Change, ICC: Interclass Correlation Coefficient, CI: Confidence Interval, LB: Lower Bound, UB: Upper Bound

Figure 3 shows the Bland-Altman plot for the level of agreement between the SAMP test scores for Examiner A across two separate occasions. The mean difference between the 2 occasions was (Mean±Standard Deviation (SD) -0.0952 ±0.48958 Repetition). The 95% confidence interval upper limits and lower limits (0.8643 and -1.0547) respectively, indicating that no systematic differences occurred between testing occasions [Bland and Altman, 2007].

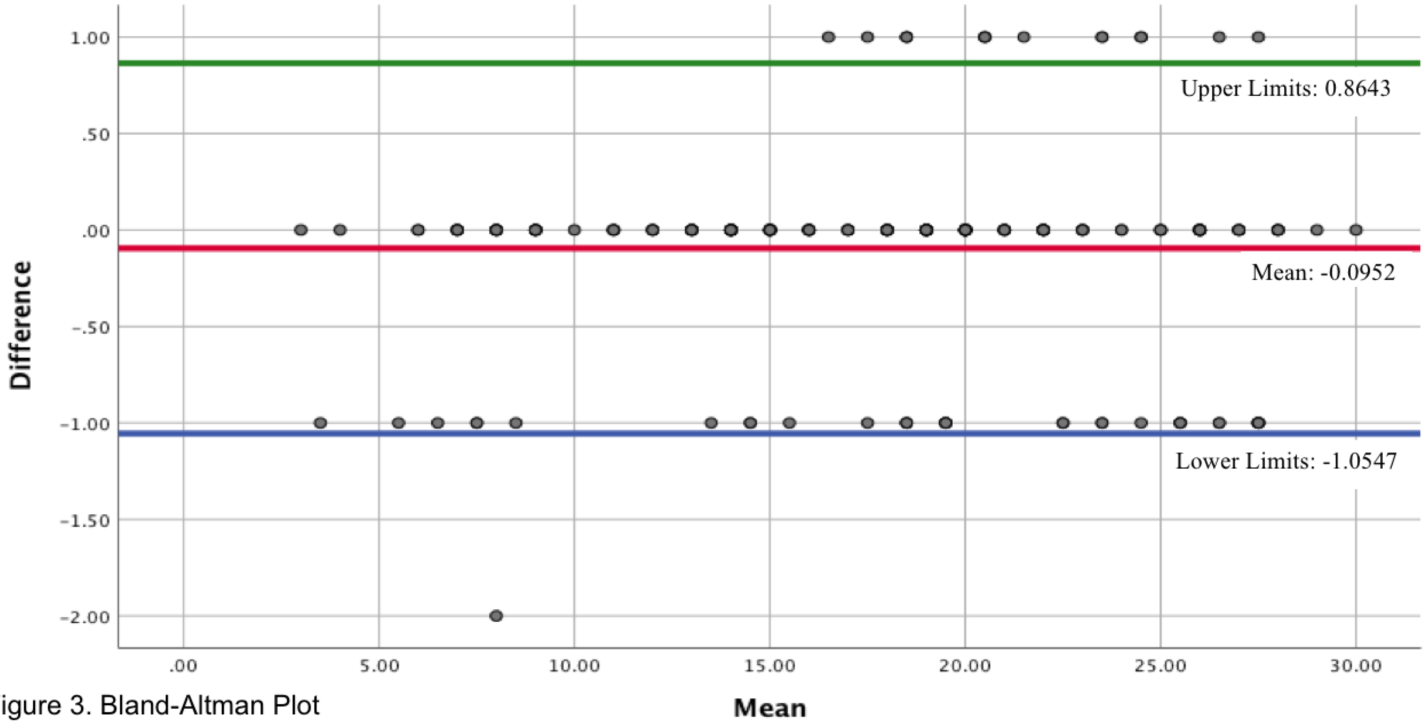

Figure 3. Bland-Altman Plot

The Mean between the 2 measurements is the Red line "d line" (Mean ± SD: -0.0952 ± 0.48958); 95% CI Upper Limits is the Green line ([SD × 1.96] + Mean [0.48958 × 1.96] + -0.0952 = 0.8643); 95% CI Lower Limits is the Blue line (Mean – [SD × 1.96] -0.0952 – [0.48958 × 1.96] = -1.0547). d = Upper Limits – Lower Limits 0.8643 – -1.0547 = 1.9191 Repetition

4. **Discussion**

4.1 Summary and discussion of the main findings

This study aimed to determine the reliability of the SAMP test in female patients with NSNP. The SAMP test demonstrated almost perfect levels of reliability. Interpretation of the 95% Confidence Intervals around the Interclass correlation coefficient ($ICC_{2,1}$) values suggest that the 'true' estimate of interrater and intrarater reliability of the SAMP test range between ICCs of 0.993 and 0.996, indicating a very high degree of stability of the SAMP scores over time and agreement between examiners. This exceeds the ICC ≥0.90 set out in hypothesis 1 and 2. Given that an ICC of at least 0.70 is considered to be satisfactory for an instrument to detect differences in severity between groups in research practice and an ICC value of 0.90-0.95 is required to enable this instrument to detect differences in severity between individual patients in clinical practice [de Vet et al., 2011], this indicates that the SAMP test can be consistently well used by different examiners or the same examiner in different occasions to measure ULD in female patients with NSNP in clinical and research practice. The reliability results of this study have confirmed previous results reported for the 3kg SAMP test used in healthy subjects and non-patient populations who reported neck pain [McLean et al., 2010]. The reason for this high level of reliability may be due to the simplicity and standardisation of the test. The SAMP test requires simple instructions and minimal training for observers who are required to only count the valid repetitions within 30 second in order to complete the administration and scoring of the test. The SAMP test also demonstrated very low levels of measurement error. The SEM was very low and smaller than SDC as expected. The SEM ranged from 0.35 to 0.42 for interrater reliability and 0.35 to 0.44 for intrarater reliability, indicating a very high level of precision in the patients' scores. This is ≤ 1 and smaller than the SDC set out in hypothesis 3. The hypotheses regarding reliability and agreement have been confirmed.

4.2 Strengths of the study

This study was conducted, analysed and interpreted in accordance with the COSMIN recommendations for developing health-related OM [Mokkink et al., 2010; Terwee et al., 2012, 2018]. Independent but simultaneous examiners were used when assessing the interrater reliability in order to reduce or possibly prevent the risk of fatigue or soreness to patients, which could lead to drop-out and also to avoid the Hawthorne effect [de Vet et al., 2011]. The large sample size achieved (n=290), which was significantly higher than the recommended sample size by the COSMIN checklist (n=100), increased the statistical power of the test of mean differences, prevented potential masking of systematic error and enabled appropriate quantification of the SAMP test reliability and agreement [de Vet et al., 2006]. The use of broad inclusion and exclusion criteria and standardised assessments which ensured that the included participants were representative of typical healthy subjects from the general population and patients with a variety of NSNP severity levels. All patients and healthy subjects who attended the first assessment and testing occasion were retained for the second testing occasion (no drop-out). To ensure robust methodology, the DASH and NDI questionnaires, which are relevant, standardised and extensively validated neck and upper limb PROMs, were used to quantify the degree of NSNP and ULD at the baseline.

4.3 Limitations of the study

This study was conducted on female patients with NSNP, which will prevent the generalisability of the findings to male population as well as those patients with other types of neck disorders such as Whiplash Associated Disorder 'WAD' or Cervical Radiculopathy. This study also involved female participants in the age group (30-50-year) only, which may limit generalisability of the findings to other younger and older patients age groups. These limitations point to the requirement for further feasibility and reliability studies in different neck pain and age populations.

5. Conclusion

This cross-sectional study established that a 1kg SAMP test has almost perfect reliability and agreement levels in female patients with NSNP. The findings suggest that the 1kg SAMP test has potential for use in clinical and research practice for the purpose of evaluating upper limb functional capacity in female patients with NSNP. However, before this test can be fully recommended, further research is required to investigate the convergent and discriminant validity of the SAMP test. In addition, a longitudinal study to explore the responsiveness of the SAMP test is also required before it can be recommended as a treatment outcome measure.

References

Ahmed SA, Shantharam G, Eltorai AEM, Hartnett DA, Goodman A, Daniels AH. The effect of psychosocial measures of resilience and self-efficacy in patients with neck and lower back pain. Spine. 2019;19(2):232-237.

Alreni ASE, Harrop D, Lowe A, Potia T, Kilner K, McLean SM. Measures of upper limb function for people with neck pain. A systematic review of measurement and practical properties. Musculoskelet Sci Pract. 2017;29:155-163.

Alreni ASE, Aboalmaty HRAA, De Hertogh, W, Meirte J, Harrop D, McLean S. Measuring upper limb disability for patients with neck pain: evaluation of the feasibility of the Single Arm Military Press (SAMP) test. 2020;50. https://doi.org/10.1016/j.msksp.2020.102254

Alotaibi N. Cross-cultural adaptation process and pilot testing of the Arabic version of the Disability of the Arm, Shoulder and Hand (DASH-Arabic). Hand Therapy. 2010;15(4): 80-86.

Beaton DE, Katz JN, Fossel AH, Wright JG, Tarasuk V, Bombardier. 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(2):128-146.

Bland JM. and Altman, DG. Measurement Error. BMJ. 1996; 312 (7047):1654. doi: 10.1136/bmj.312.7047.1654.

Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat. 2007;17:571-582.

Bobos P, MacDermid JC, Walton DM, Gross A, Santaguida PL. Patient-Reported Outcome Measures Used for Neck Disorders: An Overview of Systematic Reviews. J Orthop Sports Phys Ther 2018;48(10):775-88.

Bond A, Lader M. The use of analogue scales in rating subjective feelings. Br J med Psychol. 1974;47:211-218.

Bot SDM., Terwee CB., van der Windt DAWM., Bouter LM., Dekker J., de Vet. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann Rheum Dis. 2004; 63(4): 335-341.

Cramer D. Fundamental statistics for social research. London: Routledge; 1998.

Curb JD, Ceria-Ulep CD, Rodriguez BL, Grove J, Guralnik J, Willcox, BJ, et al. Performance-based measures of physical function for high-function populations. J Am Geriatr Soc. 2006;54(5):737-742.

de Vet HCW. Terwee B. Knol DL. Bouter LM. When to use agreement versus reliability measures. Journal of Clinical Epidemiology. 2006; 59:1033-1039.

de Vet HCW, Terwee CB, Mokkink LB, KNOL DL. Measurement in medicine: A Practical Guide. Cambridge: Cambridge University Press; 2011. 338 p.

Doane DP, Seward LE. Measuring skewness. Journal of Statistics Education. 2011;19(2):1-18.

Frank AO, De Souza LH, Frank CA. Neck pain and disability: A cross-sectional survey of the demographic and clinical characteristics of neck pain seen in a rheumatology clinic. Int J Clin Pract. 2005; 59(2):173-182.

Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [Corrected]. The Upper Extremity Collaborative Group (UECG). Am J Ind Med. 1996; 29(6):602-608.

Huisstede BMA, Bierma-Zeinstra SMA, Koes BW, Verhaar JAN. Incidence and prevalence of upper extremity musculoskeletal disorders: A systematic appraisal of the literature. BMC Musculoskeletal Disord. 2006; 7: 7.

Huisstede BM, Feleus A, Bierma-Zeinstra SM, Verhaar JA, Koes BW. Is the disability of arm, shoulder, and hand questionnaire (DASH) also valid and responsive in patients with neck complaints. Spine. 2009;34(4):130-138.

Jaeschke R, Singer J, Guyatt GH. A comparison of seven-point and visual analogue scales. Data from a randomized trial. Control Clin Trials. 1990;11(1):43-51.

Landis JR. and Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174.

Lee H, Hubscher M, Moseley L, Kampre S, Traeger A, Mansell G, McAuley J. How does pain lead to disability? A systematic review and meta-analysis of mediation studies in people with back and neck pain. Pain. 2015;156(6):988-997.

Lomond KV, Cote JN. Shoulder functional assessments in persons with chronic neck/shoulder pain and healthy subjects: reliability and effects of movement repetition. Work. 2011;38(2):169-180.

MacDermid JC, Walton DM, Avery S, Blanchard A, Etruw E, McAlpine C, Goldsmith CH. Measurement properties of the neck disability index a systematic review Journal of Orthopedic and Sports Physical Therapy. 2009 May;39(5):400-17.

McLean SM, May S, Moffett JK, Sharp DM, Gardiner E. Prognostic factors for progressive non-specific neck pain. Physical Therapy Reviews. 2007;12(3):207-220.

McLean SM, Taylor J, Balassoubramien T, Kulkarni M, Patekar P, Darne R, et al. Measuring upper limb disability in non-specific neck pain: a clinical performance measure. International Journal of Physiotherapy and Rehabilitation. 2010;1:44-52.

McLean SM, Moffett JK, Sharp DM, Gardiner E. An investigation to determine the association between neck pain and upper limb disability for patients with non-specific neck pain: a secondary analysis. Man Ther. 2011;16(5):434-439.

Mehta S, MacDermid JC, Carlesso LC, McPhee C. Concurrent validation of the DASH and the QuickDASH in comparison to neck-specific scales in patients with neck pain. Spine. 2010;35(24):2150-2156.

McCormack HM, Horne DJ, Sheather S. Clinical applications of visual analogue scales: a critical review. Psychol Med. 1988;18(4):1007-1019.

McGraw K, Wong S. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1:30–46.

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(4):539-549.

National Institute for Health and Care Excellence (NICE) guidelines, 2018. Neck pain–non-specific. https://cks.nice.org.uk/topics/neck-pain-non-specific/

Osborn W, Jull G. Patients with non-specific neck disorders commonly report upper limb disability. Man Ther. 2013;18(6):492-497.

Pinheiro LC, Callahan LF, Cleveland RJ, Edwards LJ, Reeve BB. The Performance and Association Between Patient-reported and Performance-based Measures of Physical Functioning in Research on Individuals with Arthritis. J Rheumatol. 2016;43(1):131-137.

Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorv-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics. 2011; 2(1):21-33.

Roy J., MacDermid JC., Woodhouse L. Measuring shoulder function: A systematic review of four questionnaires. Arthritis & Rheumatism. 2009; 61(5): 623-632.

Shaheen AA, Omar MT, Vernon H. Cross-cultural adaptation, reliability, and validity of the Arabic version of neck disability index in patients with neck pain. Spin. 2013;38(10):609-615.

Stock S, Loisel P, Durand M, et al. L'indice d'impact de la douleur au cou et aux membres supérieurs sur la vie quotidienne (NULI: neck and upper limb index). Études et Recherches. 2003;R-355. Available at: http://www.irsst.qc.ca/media/documents/pubirsst/R-355.pdf

Streiner DL, Norman GR, Cairney J. Health measurement scales: A practical guide to their development and use. Oxford: Oxford University Press; 2015.

Terwee CB, Mokkink LB, Knol DL, Ostelo R, Bouter LX, de Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Qual Life Res. 2012;21(4):651-657.

Terwee CB, Prinsen CAC, Chiaeotti A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159-1170.

Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther. 1991;14(7):409-415.

Vincent WJ. and Weir JP. Statistics in Kinesiology. 4th Edition; United State: Human Kinetics; 2012.

von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. Ann Intern Med. 2007;147(8):573-577.

Walker-Bone K, Byng P, Linaker C, Reading I, Coggon D, Palmer K, Cooper C. Reliability of the Southampton examination schedule for the diagnosis of upper limb disorders in the general population. Ann Rheum Dis. 2002;61:1103-1106.

Walker-Bone K, Palmer KT, Reading I, Coggon D, Cooper C. Prevalence and impact of musculoskeletal disorders in the upper limb in the general population. Arthritis Rheum. 2004;51(4):642-651.

Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. Res Nurs Health. 1990;13(4):227-236.

World Health Organization (WHO). International Classification of Functioning, Disability and Health: ICF. 2001; http://www.who.int/.