BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (http://bmjopen.bmj.com).

If you have any questions on BMJ Open's open peer review process please email editorial.bmjopen@bmj.com

BMJ Open

Inter-tester reliability of clinical shoulder instability and laxity tests in subjects with and without self-reported shoulder problems

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018472
Article Type:	Research
Date Submitted by the Author:	30-Jun-2017
Complete List of Authors:	Eshoj, Henrik; Syddansk Universitet Det Sundhedsvidenskabelige Fakultet, Department of Sports Science and Clinical Biomechanics; Odense Universitetshospital, Quality of Life Research Center, Department of Haematology Ingwersen, Kim Gordon; Hospital Lillebaelt - Vejle Hospital, Vejle, Denmark, Research Unit at Department of Physical- and Occupational Therapy,; Syddansk Universitet Det Sundhedsvidenskabelige Fakultet, Department of Sports Science and Clinical Biomechanics Larsen, Camilla; Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics Kjær, Birgitte; Bispebjerg and Frederiksberg University Hospital, Copenhagen, Denmark , Research Unit at Department of Physical- and Occupational Therapy, ; Syddansk Universitet Det Sundhedsvidenskabelige Fakultet, Department of Sports Science and Clinical Biomechanics Juul-Kristensen, Birgit; Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics
Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Sports and exercise medicine
Keywords:	Reliability, Shoulder < ORTHOPAEDIC & TRAUMA SURGERY, Instability, Laxity, Clinical test

SCHOLARONE™
Manuscripts

Inter-tester reliability of clinical shoulder instability and laxity tests in subjects with and without self-reported shoulder problems

Henrik Eshoj[1], PhD[a,b] Kim Gordon Ingwersen, PhD [a,c], Camilla Marie Larsen, PhD [a,d] Birgitte Hougs Kjær, Msc [a,e], Birgit Juul-Kristensen, PhD [a,f]

Affiliations

[a] Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark

[b] Quality of Life Research Center, Department of Haematology, Odense University Hospital, Odense, Denmark

[c] Research Unit at Department of Physical- and Occupational Therapy, Hospital Lillebaelt - Vejle Hospital, Vejle, Denmark

[d] Health Sciences Research Centre, University College Lillebaelt, Odense, Denmark

[e] Research Unit at Department of Physical- and Occupational Therapy, Bispebjerg and Frederiksberg University Hospital, Copenhagen, Denmark

[f] Institute of Occupational Therapy, Physiotherapy and Radiography, Department of Health Sciences, Bergen University College, Norway

Correspondence:

Birgit Juul-Kristensen

Department of Sports Science and Clinical Biomechanics

University of Southern Denmark, Campusvej 55, 5230 Odense M, DK

Tel.: +45 2942 3335

Email: bjuul-kristensen@health.sdu.dk

Word counts: 2866

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Abstract

Objective: Firstly, to investigate the inter-tester reliability of clinical shoulder instability and laxity tests, and secondly, to describe the mutual dependency of each of the individual tests for identifying self-reported shoulder problems.

Method: A standardized protocol for conducting reliability studies was used to test the inter-tester reliability of the six clinical shoulder instability and laxity tests; apprehension, relocation, surprise, load-and-shift, sulcus sign and Gagey. Cohens kappa (*k*) with 95% confidence intervals (CI) besides Prevalence-Adjusted-Bias-Adjusted-Kappa (PABAK), accounting for insufficient prevalence and bias, were computed to establish the inter-tester reliability and mutual dependency.

Results: Forty individuals (13 with instability and laxity related shoulder problems and 27 shoulder healthy individuals) aged 18-60 were included. Fair (relocation), moderate (load-and-shift, sulcus sign) and substantial (apprehension, surprise, Gagey) inter-tester reliability were observed across tests (*k* 0.39-0.73; 95% CI: 0.00-1.00). PABAK improved reliability across tests, resulting in substantial to almost perfect inter-tester reliability for the apprehension, surprise, load-and-shift and Gagey tests (*k* 0.65-0.90). Mutual dependencies between each test and self-reported shoulder problem showed apprehension, relocation and surprise to be the most often used tests to characterize such musculoskeletal shoulder conditions.

Conclusions: Four tests (apprehension, surprise, load-and-shift and Gagey) out of six, were considered inter-tester reliable for clinical use, while relocation and sulcus sign tests need further

For peer review only

standardization before acceptable evidence. Furthermore, the validity of the tests for instability and laxity needs to be studied.

Article summary section

Strengths and limitations of this study

- The strength of the study is the use of a three-phased standardized study protocol
- Presentation of raw findings increases transparency and interpretation of study findings
- No valid gold standard for including shoulder instability and laxity subjects was used
- A 50/50 prevalence of positive and negative tests for all six tests was not accomplished

Keywords

Reliability, Shoulder instability, laxity, clinical tests

Acknowledgements

We would like to thank Physiotherapists Rasmus Fitzner, Pernille Madsen and Jacob Hansen from Metropolitan University College, Copenhagen, Denmark for recruitment and testing of study participants. Furthermore, a special thanks to Bispebjerg Frederiksberg University Hospital, Copenhagen, Denmark for providing facilities for data collection.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Introduction

Shoulder complaints, affecting shoulder-related quality of life (QoL), are frequent and may be caused by shoulder instability and/or laxity[1] due to traumatic or non-traumatic injuries to the shoulder joint.[2] The traumatic shoulder instability is mainly prompted by a high impact injury during sports participation, resulting in a shoulder dislocation, predominantly in anterior direction.[3] The non-traumatic shoulder instability is usually related to repetitive overhead activities and/or patients with generalised joint hypermobility or glenohumeral hyperlaxity, often refered to as multidirectional shoulder instability.[2,4,5]

Irrespectively of aetiology, shoulder instability and laxity is often accompanied by a variety of symptoms including shoulder discomfort, pain besides glenohumeral subluxations and/or repeated dislocations.[6,7,8] Clinically, shoulder instability and laxity, is diagnosed and verified by a group of shoulder pain and instability provoking/relieving tests, supplemented by shoulder laxity tests.[9,10] The former tests usually include the anterior shoulder instability and laxity tests; apprehension, relocation and surprise, and the laxity tests consisting of the load-and-shift, sulcus sign and Gagey tests.[11,12,13] An ongoing discussion is the use of pain as diagnostic criterion in diagnosing anterior shoulder instability with the shoulder-provoking tests apprehension, relocation and surprise. In one way, it may be a confounding factor, since pain has shown to be less predictive and reliable as a diagnostic criterion.[14] On the contrary though, others have suggested that unrecognized and underlying glenohumeral instability may lead to repetitive microtrauma and painful shoulder conditions,[15,16] justifying pain as diagnostic criterion when testing for anterior shoulder instability.

Nonetheless, symptoms may become chronic, and lead to reduced work and sports capability,[17,18,19] and with exercise-based management as the most often recommended first-

For peer review only

choice treatment.[20,21] Hence, early diagnosis using reliable and accurate clinical tests to guide focused treatment is essential. Few studies though, have investigated the reliability of clinical shoulder instability and laxity tests showing large variations in reliability and with limited methodological quality, hampering interpretation and comparison with other studies.[14,22,23]

Therefore, the objective of this study was to investigate the inter-tester reliability of commonly used clinical shoulder instability and laxity tests in a group of sports-active individuals with and without self-reported shoulder problems.

For peer review only

Materials and methods

Study design

An inter-tester reliability study was conducted involving two physiotherapists as inter-tester examiners. A third physiotherapist (study coordinator), not involved in the actual inter-tester reliability study, managed all practical aspects during the study period. The Guidelines for Reporting Reliability and Agreement Studies (GRASS) were followed.[24] A standardized protocol for reliability studies, consisting of three phases: preparation and training of clinical tests, overall agreement, and study phase (the actual reliability study) were applied.[25] Two early career physiotherapists with six months clinical experience were involved in the inter-tester reliability study. A test-protocol describing each clinical test was developed and subsequently used by the two testers to practice all tests in order to reach uniformity and mutual agreement in performing and interpreting each test. In the overall agreement phase, the two testers examined 19 individuals (eight shoulder cases and 11 shoulder healthy). The two testers were mutually blinded to the health status of the individuals (shoulder cases vs. shoulder healthy) and also to each other's test results. Before proceeding to the final study phase, the two testers needed an overall agreement of at least 80% based on findings from the six clinical shoulder tests.[25] In the actual inter-tester reliability study phase, the two testers examined a new group of shoulder cases and shoulder healthy individuals with the six clinical shoulder tests. The procedure was the same as in the agreement phase, meaning that testers were blinded to the health status of the individuals and each other's test results.

For peer review only

Study subjects

A sample size of at least 40 individuals was targeted based on recommendations for performing clinical reliability studies.[25] Sixty-five (women and men (aged 18-60 years)) were recruited and screened for eligibility from Metropolitan University College, Copenhagen, and Bispebjerg Frederiksberg University Hospital, Copenhagen resulting in an included number of 13 individuals with instability and/or laxity related shoulder problems (hereinafter referred to as cases) vs. 27 shoulder healthy individuals, respectively.

Cases answering yes to at least one of two questions (‘*Do you have a sense of shoulder instability*?’ and ‘*Have you ever had a shoulder injury*?’) were eligible for a clinical shoulder examination performed by the study coordinator. Cases were then included if they present with at least one positive clinical shoulder test out of the following; apprehension, relocation, surprise, load-and-shift, sulcus sign or Gagey. The shoulder healthy individuals were included if they present with no self-reported shoulder pathology or complaints. In general, any individuals with prior shoulder surgery were excluded. In the actual study phase, individuals completed a short questionnaire with basic demographic details (age, gender, weight, height), in addition to the following: pain level during rest and activity (numeric pain rating scale, NPRS),[26] shoulder injury ever (yes/no), subjective shoulder instability (yes/no) and sports-related activity (hours/week). Further, all individuals filled in the patient-reported Western Ontario Shoulder Instability (WOSI) questionnaire designed to measure shoulder function and quality of life in patients with shoulder instability and laxity symptoms.[27] The study was exempted for notification to the Danish Health Research Study Board due to the non-invasive/-non-treating study design. However, oral and written consent was provided from all individuals and, ethical guidelines were followed according to the Helsinki declaration.[28]

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Clinical tests

The clinical shoulder tests consisted of three shoulder joint-provoking tests for anterior shoulder instability (apprehension, relocation and surprise) besides three shoulder laxity tests (load and shift, sulcus sign and Gagey) (Table 1).[11,13,14,22,23,29]

The apprehension and surprise tests were positive if glenohumeral apprehension and/or pain were evoked during testing whereas relieve of symptoms with the relocation test was regarded as a positive test. The load and shift test was rated on a four-point scale ranging from 0-3 (best to worst; 0= little glenohumeral movement; 3=humeral head moves beyond the glenoid rim and remains dislocated).[12] Also, to enhance mutual agreement between testers when performing the load-and-shift test, only the direction (anterior vs. posterior) with most glenohumeral head translation was rated. Sulcus sign was objectively measured in centimeter (continuous scale) by use of a small ruler according to previously used grading scales as follows; I (<1 cm translation), II (1 to 2.0 cm) or III (>2.0 cm).[29] Finally, Gagey test was rated as positive with passive abduction above 105 degrees.[13]

Statistics

Demographics and descriptive data were tested for normality by visual inspection of histograms and Shapiro-Wilk´s test. Group differences (shoulder cases vs. shoulder healthy) were tested by Fisher´s Exact test for categorical variables, whereas student´s t-test and Mann-Whitney U-test was used for parametric, respectively non-parametric distributed data.

Apprehension, relocation, surprise and Gagey test were dichotomous variables whereas the load and shift and sulcus sign test were dichotomized to also allow for nominal statistics. Thus, load-and-shift was rated positive when scored 2 or 3, while for sulcus sign a positive rating was equal to

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml
For peer review only

measurements exceeding one centimeter.[29] For transparency, data from each test is presented by 2 x 2 contingency tables besides the use of McNemar´s test for significant between-tester differences. Furthermore, observed and expected agreements are presented along with prevalence and bias[30] indexes. Reliability was evaluated with the use of Cohen´s kappa (k) coefficients including 95% confidence intervals (CI).[25] Also, since kappa is sensitive to imbalances in prevalence and bias (e.g. the number of positive and negative tests not close to 50%) a Prevalence-Adjusted-Bias-Adjusted Kappa (PABAK) was calculated.[30,31] By definition, PABAK reflects the ideal situation, thereby accounting for variation of prevalence and bias between testers (as presented in the "real" world).[32] The relationship between the individual tests and the classification (mutual dependency) by self-reported shoulder problems was tested by Cohen´s kappa (k) coefficients and the characterization of the groups was tested with Fischer's exact tests. The classification system proposed by Landis and Koch was used to interpret reliability as follows: 0.00-0.20 (Slight); 0.21-0.40 (Fair); 0.41-0.60 (Moderate); 0.61-0.80 (Substantial) and 0.81-1.00 (Almost perfect).[33]

Statistical Package for the Social Sciences (SPSS inc., Chicago, IL, USA), version 22, was used for all statistical analyses, with p-value of <0.05 interpreted as significant.

For peer review only

Inter-tester reliability of clinical shoulder instability and laxity tests

Table 1. Performance and evaluation of the clinical shoulder instability and laxity tests

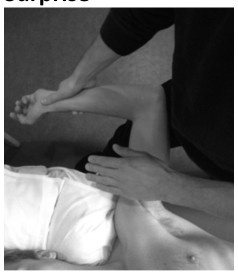
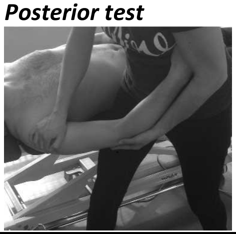
Verbal introduction:

1. I am going to perform six clinical shoulder tests on you
2. I will ask if you experience any symptoms (apprehension and/or pain) during the three first tests.
3. I will guide you through each test

Clinical tests	Description	Placing of hands, etc.	Evaluation
Apprehension 	Individuals placed supine with the shoulder being tested close to the edge of the examination table. Shoulder positioned in 90˚of abduction, elbow flexed to 90˚. Examiner moves the shoulder into maximal external rotation.	One hand around the wrist of the individual with the other hand gently placed in front of the shoulder. Elbow supported at the examiners thigh.	Subjective or objective presence of apprehension and/or pain? Rated as either positive or negative. (Nominal, dichotome data)
Relocation 	From the end position of the apprehension test the humeral head is gently forced posteriorly	Examiners fifth finger placed close to the lateral part of the acromion with the wrist positioned anteriorly of the humeral head.	Diminish of apprehension symptoms and/or pain? Rated as either positive or negative. (Nominal, dichotome data)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Inter-tester reliability of clinical shoulder instability and laxity tests

Surprise 	From end position of the relocation test the posteriorly directed force at the humeral head is quickly removed.	Removal of examiners wrist from the anterior part of the shoulder.	Subjective or objective reproduction of apprehension symptoms and/or pain? Rated as either positive or negative. (Nominal, dichotome data)
Load-and-shift	Individual placed supine with scapula resting at the examination table. Humeral head is loaded gently into the glenoid through axial pressure at the elbow.	Examiners one hand placed at the olecranon with the individual's hand positioned between the examiners torso and elbow.	Humeral head movement evaluated by the use of a four-level laxity scale. 0 = little to almost no movement
Anterior test 	*Anterior load-and-shift:* Shoulder positioned in the scapular plane in 90˚ of abduction with elbow flexed. Humeral head gently shifted in anterior direction.	*Anterior load-and-shift:* Examiners hand placed on top of the shoulder with the fingers on the backside of the glenohumeral head to move it anteriorly.	1 = humeral head moves up onto the glenoid 2 = humeral head moves beyond the glenoid, but relocates spontaneously once pressure is released
Posterior test 	*Posterior load-and-shift:* Shoulder positioned in the scapular plane in 20˚ of abduction with elbow flexed. Humeral head gently shifted in posterior direction.	*Posterior load-and-shift:* Examiners wrist placed at the anterior part of the humeral head to move it posteriorly.	3 = humeral head moves beyond the glenoid and remains dislocated Rated as positive when scored 2 or 3.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Inter-tester reliability of clinical shoulder instability and laxity tests

			(Nominal, dichotome data)
Sulcus sign 	Individual sitting upright. Examiner pulls the distal part of the humerus in a caudal direction. Distance from the top of the humeral head and the acromion is measured with a ruler.	One hand placed above the epicondyles of humerus. Other hand is used to measure the subacromial distance with a ruler	Rated as positive with measurements exceeding 1 centimeter. (Nominal, dichotome data)
Gagey 	Individual sitting upright. The shoulder girdle is gently depressed with the arm passively moved into end range in horizontal abduction. A mirror in front of the individual is used to evalate shoulder abduction angle.	Forearm placed on top of the shoulder girdle with the other hand placed around the elbow joint.	Rated as positive with abduction exceeding 105 degrees. (Nominal, dichotome data)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Results

Characteristics of the participating individuals are presented in Table 2. Demographics showed no difference between the shoulder cases (n=13) and shoulder healthy (n=27) individuals. Further, both groups (92 and 74%; p=0.18) were relatively active with a weekly participation in sports-related activity for more than four hours per week. However, as expected due to the design, shoulder cases had significantly higher pain during activity (4.23 vs. 1.44; p=0.02), higher frequency of shoulder injury ever (62% vs. <1%; p<0.001), higher subjective shoulder instability (69 vs. 11%; p<0.001) and worse total WOSI score (506 vs. 136; p=0.001) (Table 2).

Table 2. Study phase. Participant characteristics.

	Shoulder cases (n=13)	**Shoulder healthy (n=27)**	***P-value***
Sex (women/men)	8/5	21/6	0.28
Age (years) mean (SD)	28 (9)	29 (7)	0.72
Weight (kg)* mean (SD)	71.0 (12.8)	74.9 (23.4)	0.59
Height (cm) mean (SD)	174.0 (8.6)	173.4 (7.9)	0.82
Pain, rest (NRS 0-10) mean (SD)	1.08 (1.44)	0.41 (1.15)	0.12
Pain, activity (NRS 0-10) mean (SD)	4.23 (2.92)	1.44 (2.12)	<0.05
Shoulder injury ever, n (%)	8 (62)	1 (<1)	<0.001
Subjective shoulder instability, n (%)	9 (69)	3 (11)	<0.001
Sports-related activity (>4 hours/week), n (%)	12 (92)	20 (74)	0.18
WOSI domains, mean (SD)			
- Physical symptoms (0-1000)	225 (165)	60 (78)	<0.05
- Sports, recreation, work (0-400)	103 (93)	24 (47)	<0.05
- Lifestyle (0-400)	58 (57)	13 (21)	<0.05
- Emotions (0-300)	121 (94)	39 (49)	<0.05
WOSI total score (0-2100) mean (SD)	506 (362)	136 (174)	<0.001

***SD* Standard deviation; *kg* kilo; * significance level p < 0.05; *cm* centimeter *NRS* Numeric Rating Scale**

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Prevalence of positive tests was especially low for the load-and-shift test (table 3), and significant between-tester differences were found for relocation and sulcus sign tests (p=0.021) (not shown in tables).

Table 3. Contingency tables with findings from tester A and B.

Apprehension		**A** Yes	**A** No
B	Yes	14	4
	No	3	19

Relocation		**A** Yes	**A** No
B	Yes	6	2
	No	8	24

Surprise		**A** Yes	**A** No
B	Yes	14	4
	No	3	19

Load-and-shift		**A** Yes	**A** No
B	Yes	1	0
	No	2	37

Sulcus		**A** Yes	**A** No
B	Yes	7	1
	No	9	23

Gagey		**A** Yes	**A** No
B	Yes	8	3
	No	1	28

Reliability varied between *k*: 0.39-0.73 (95% CI: 0.00-1.00), indicating fair (relocation; *k* 0.39), moderate (load-and-shift, sulcus sign; *k* 0.43 and 0.48) and substantial (apprehension, surprise, Gagey; *k* 0.65-0.73) reliability (Table 4). The prevalence index of all six tests ranged from 0.05-0.44, (lowest for load-and-shift, relocation and sulcus; 0.05, 0.28 and 0.30), whereas the bias index ranged from and 0.03-0.20 (highest for relocation and sulcus). PABAK improved reliability for relocation, load-and-shift, sulcus sign and Gagey test, now corresponding to moderate (relocation and sulcus sign; *k* 0.50), substantial (Gagey; *k*: 0.80) and almost perfect (load and shift; *k*: 0.90) reliability (Table 4).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 4. Study phase. Reliability of six clinical shoulder instability and laxity tests.

	Observed agreement	Expected agreement	Prevalence index	Bias index	Kappa (95% CI)	PABAK
Apprehension	0.83	0.51	0.44	0.03	0.65 (0.38; 0.85)	0.65
Relocation*	0.75	0.59	0.28	0.15	0.39 (0.07; 0.68)	0.50
Surprise	0.83	0.51	0.44	0.03	0.65 (0.38; 0.85)	0.65
Load and Shift	0.95	0.90	0.05	0.05	0.48 (0.00; 1.00)	0.90
Sulcus sign*	0.75	0.56	0.30	0.20	0.43 (0.17; 0.72)	0.50
Gagey	0.90	0.62	0.40	0.05	0.73 (0.46; 0.94)	0.80

223 ******Significant inter-tester differences*, **CI** Confidence intervals, **PABAK** *prevalence-and-bias-adjusted*
224 *kappa*

226 The *k* values for mutual dependency indicate that apprehension, relocation and
227 surprise tests for both examiners were the most frequently used tests for
228 characterizing self-reported shoulder problems (Table 5). This was further confirmed
229 by the significant group difference in the presence of positive tests.

231 **Table 5.** Kappa statistics for mutual dependency of the individual tests and self-
232 reported shoulder problems.

	Observed agreement	Expected agreement	Prevalence index	Kappa	p-value (cases/controls)
Apprehension					
Examiner A	0.75	0.53	0.38	0.47	0.003
Examiner B	0.68	0.52	0.40	0.33	0.04
Relocation*					
Examiner A	0.83	0.55	0.35	0.61	<0.001
Examiner B	0.73	0.63	0.25	0.27	0.08
Surprise					
Examiner A	0.75	0.53	0.38	0.47	0.003
Examiner B	0.68	0.52	0.40	0.33	0.04
Load and Shift					
Examiner A	0.75	0.65	0.20	0.29	0.03
Examiner B	0.70	0.67	0.18	0.10	0.33
Sulcus sign*					
Examiner A	0.63	0.61	0.28	0.05	0.52
Examiner B	0.63	0.54	0.38	0.19	0.19
Gagey					
Examiner A	0.70	0.60	0.28	0.26	0.10
Examiner B	0.70	0.58	0.30	0.29	0.08

For peer review only

Discussion

The inter-tester reliability across the selected six clinical shoulder instability and laxity tests ranged from fair to substantial. Use of PABAK calculations improved inter-tester reliability to substantial and almost perfect across most tests, except for the relocation and sulcus sign tests. The tests most often used to characterize self-reported shoulder problems (mutual dependency) were apprehension, relocation and surprise tests.

The inter-tester reliability for the apprehension, relocation and surprise was higher than, or equivocal, to previously reported results of these tests using the same diagnostic procedures (apprehension and/or pain).[23] Specifically for the apprehension and surprise test, the present *k* values were somewhat higher than previously reported (0.65 vs. 0.44-0.45). The reason for this may be that the current study included both shoulder cases and shoulder healthy individuals as opposed to only including symptomatic subjects.[23] This may have increased subject variation, known to affect reliability positively. Also, PABAK calculations did not affect the overall reliability of the apprehension and surprise tests, probably due to an optimal prevalence index of positive and negative tests (close to 0.50). For the relocation test, the existing inter-tester reliability was almost similar to previously reported (*k* 0.39 vs. 0.44),[23] however, fairly lower. Apparently, the primary reason for the current poor reliability in relocation was presence of systematic bias between testers, as indicated by the actual raw data (contingency tables) and the statistical significant inter-examiner difference. Likewise, systematic bias between testers was also fund for the sulcus sign test in the present study. Hypothetically, this may be explained by inter-tester variability in the

For peer review only

force produced to relocate, respectively inferiorly translate the glenohumeral head during the relocation and sulcus sign tests in the current study. This is, however, only speculative and further studies are needed to standardize these tests.

Reliability for the present sulcus sign test was slightly lower than previously reported (*k* 0.39 vs. >0.50).[22,23] The discrepancy in reliability observed may be due to the use of different test positions with participants in the current study sitting upright[29] as opposed to a previous lying test position.[22] Although, using PABAK it applies for both the relocation and sulcus sign test that overall reliability was not affected much, due to the presence of systematic bias.

For the load-and-shift test, reliability was relatively low (including wide CI). This may be due to the current poor prevalence index. However, the present dichotomous rating of the load-and-shift test (meaning that only individuals that could either subluxate or dislocate the shoulder during testing was deemed positive) may have influenced the prevalence of positive tests largely. Therefore, using PABAK, reliability of the load-and-shift test improved considerably (from moderate to almost perfect). Nevertheless, different statistics (kappa vs. Intra-class-Correlation Coefficients), different scoring systems (dichotomous rating (positive yes/no) vs. four point grading scale (0-3)[23] and inclusion of shoulder asymptomatic athletes only[22] make comparison across studies difficult.

Finally, reliability of the Gagey test was substantial and PABAK did not affect reliability much due to a nearly optimal prevalence and low bias between testers. Unfortunately, there is no other study to compare with.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Although the current study was designed to investigate reliability, and not diagnostic accuracy, the mutual dependency between the individual tests and self-reported shoulder problems was analysed. It revealed that the tests most often used to characterize those with and without self-reported shoulder problems (mutual dependency) proved to be the apprehension, relocation and surprise tests. This may indicate a relationship between these tests, which may come as no surprise, since these tests are a continuum of the apprehension test and, thus, closely related.[9] Nevertheless, for clinicians it is of interest to specify the clinical characteristics of patients with self-reported shoulder problems. Thus, the current prevalence of positive tests may mirror these characteristics of the included patients and should be taken into consideration in the management of such musculoskeletal conditions. It could be suggested to develop and test the clinimetric properties of a more comprehensive test battery for evaluating self-reported shoulder problems. No prior studies were found addressing mutual dependency of the current tests for shoulder instability and laxity, which hampers comparison.

The present study has several limitations. First, despite the use of a standardized test protocol, including training and an overall agreement phase, two of the tests (relocation and sulcus sign tests) failed to reach recommendable reliability values. Further standardization in both performance and interpretation are therefore needed. Secondly, no valid gold standard for classifying shoulder instability/laxity was used. To compensate for this, self-reported confirmation of shoulder-related problems was applied, but this was not reflected in the current WOSI scores, which were relatively

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

low. Lack of a more objective gold standard may have decreased diagnostic accuracy, however, not reliability, which was the primary objective of the present study. Thirdly, although the recommended prevalence of 0.50 in reliability studies[25] in all six tests was not accomplished, use of PABAK calculations was able to represent a valid alternative for the original kappa values. Nevertheless, future studies should use inclusion criteria of more established shoulder instability and laxity conditions, and, if possible, verified by objective criteria as surrogate for a gold standard of shoulder instability and laxity. This may optimize prevalence as well as diagnostic accuracy in studies where this is a further aim.

The strengths of the study are the use of standardized procedures (including blinding to patient status and the use of a three-phased protocol for conducting reliability studies). Also, presentation of raw data, using contingency tables, along with kappa and PABAK values, increases data transparency and improves interpretation of the reliability study.

Conclusions

This study showed acceptable inter-tester reliability for four of six clinical shoulder instability and laxity tests in relatively sports active individuals with and without self-reported shoulder disability. However, relocation and sulcus sign tests need further standardisation before being recommended for use in clinical practice. Based on the frequency and mutual dependency of the current tests, especially apprehension and surprise tests seem important in the characterisation of self-reported shoulder

For peer review only

problems. Future research on the validity of tests for shoulder instability and laxity is needed.

For peer review only

328 **Author statement:** HE, KGI, CML and BJK conceived and designed the study. HE and
329 BHK recruited study participants and collected data. HE performed the statistical
330 analysis. HE, KGI, CML, BJK interpreted the results. HE drafted the manuscript with
331 KGI, CML, BHK and BJK contributing to the manuscript. All authors have read and
332 approved the final manuscript. HE is the guarantor.

333

334 **Funding:** This work was supported by Region of Southern Denmark's Research fund
335 and The Danish Rheumatism Association.

336

337 **Competing interest:** None declared

338

339 **Data sharing statement:** No additional data are available

340

For peer review only

References

1. Luime, J. J., B. W. Koes, I. J. Hendriksen, A. Burdorf, A. P. Verhagen, H. S. Miedema, and J. A. Verhaar. 2004. 'Prevalence and incidence of shoulder pain in the general population; a systematic review', *Scandinavian Journal of Rheumatology*, 33: 73-81.

2. Farrar, N. G., J. J. Malal, J. Fischer, and M. Waseem. 2013. 'An overview of shoulder instability and its management', *Open Orthopaedics Journal*, 7: 338-46.

3. Goss, T. P. 1988. 'Anterior glenohumeral instability', *Orthopedics*, 11: 87-95.

4. Cameron, K. L., M. L. Duffey, T. M. DeBerardino, P. D. Stoneman, C. J. Jones, and B. D. Owens. 2010. 'Association of generalized joint hypermobility with a history of glenohumeral joint instability', *J Athl Train*, 45: 253-8.

5. Chahal, J., J. Leiter, M. D. McKee, and D. B. Whelan. 2010. 'Generalized ligamentous laxity as a predisposing factor for primary traumatic anterior shoulder dislocation', *J Shoulder Elbow Surg*, 19: 1238-42.

6. Dodson, C. C., and F. A. Cordasco. 2008. 'Anterior glenohumeral joint dislocations', *Orthop Clin North Am*, 39: 507-18, vii.

7. Jaggi, A., and S. Lambert. 2010. 'Rehabilitation for shoulder instability', *Br J Sports Med*, 44: 333-40.

8. Wilk, K. E., L. C. Macrina, and M. M. Reinold. 2006. 'Non-operative rehabilitation for traumatic and atraumatic glenohumeral instability', *N Am J Sports Phys Ther*, 1: 16-31.

9. Hegedus, E. J., A. P. Goode, C. E. Cook, L. Michener, C. A. Myer, D. M. Myer, and A. A. Wright. 2012. 'Which physical examination tests provide clinicians with the most value when examining the shoulder? Update of a systematic review with meta-analysis of individual tests', *Br J Sports Med*, 46: 964-78.

10. Ann Ganestam, Mikkel L. Attrup, Per Hølmich and Kristoffer W Barfoed. 2015. 'Evaluation of Clinical Practice of Shoulder Examination amon Ten Experienced Shoulder Surgeons', *Journal of Orthopaedic Research and Physiotherapy*, 1.

11. Luime, J. J., A. P. Verhagen, H. S. Miedema, J. I. Kuiper, A. Burdorf, J. A. Verhaar, and B. W. Koes. 2004. 'Does this patient have an instability of the shoulder or a labrum lesion?', *JAMA*, 292: 1989-99.

12. Tzannes, A., and G. A. Murrell. 2002. 'Clinical examination of the unstable shoulder', *Sports Medicine*, 32: 447-57.

13. Gagey, O. J., and N. Gagey. 2001. 'The hyperabduction test', *J Bone Joint Surg Br*, 83: 69-74.

14. Lo, I. K., B. Nonweiler, M. Woolfrey, R. Litchfield, and A. Kirkley. 2004. 'An evaluation of the apprehension, relocation, and surprise tests for anterior shoulder instability', *Am J Sports Med*, 32: 301-7.

15. Silliman, J. F., and R. J. Hawkins. 1993. 'Classification and physical diagnosis of instability of the shoulder', *Clin Orthop Relat Res*: 7-19.

16. Cavallo, R. J., and K. P. Speer. 1998. 'Shoulder instability and impingement in throwing athletes', *Med Sci Sports Exerc*, 30: S18-25.

17. Matsen, F. A., 3rd, and J. D. Zuckerman. 1983. 'Anterior glenohumeral instability', *Clin Sports Med*, 2: 319-38.

18. Robinson, C. M., J. Howes, H. Murdoch, E. Will, and C. Graham. 2006. 'Functional outcome and risk of recurrent instability after primary traumatic anterior shoulder dislocation in young patients', *J Bone Joint Surg Am*, 88: 2326-36.

19. van Kampen, D. A., T. van den Berg, H. J. van der Woude, R. M. Castelein, C. B. Terwee, and W. J. Willems. 2013. 'Diagnostic value of patient characteristics, history, and six clinical tests for traumatic anterior shoulder instability', *J Shoulder Elbow Surg*, 22: 1310-9.

20. Warby, S. A., T. Pizzari, J. J. Ford, A. J. Hahne, and L. Watson. 2014. 'The effect of exercise-based management for multidirectional instability of the glenohumeral joint: a systematic review', *J Shoulder Elbow Surg*, 23: 128-42.

21. Gibson, K., A. Growse, L. Korda, E. Wray, and J. C. MacDermid. 2004. 'The effectiveness of rehabilitation for nonoperative management of shoulder instability: a systematic review', *J Hand Ther*, 17: 229-42.

22. Levy, A. S., S. Lintner, K. Kenter, and K. P. Speer. 1999. 'Intra- and interobserver reproducibility of the shoulder laxity examination', *Am J Sports Med*, 27: 460-3.

23. Tzannes, A., A. Paxinos, M. Callanan, and G. A. Murrell. 2004. 'An assessment of the interexaminer reliability of tests for shoulder instability', *J Shoulder Elbow Surg*, 13: 18-23.

24. Kottner, J., L. Audige, S. Brorson, A. Donner, B. J. Gajewski, A. Hrobjartsson, C. Roberts, M. Shoukri, and D. L. Streiner. 2011. 'Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed', *Int J Nurs Stud*, 48: 661-71.

25. Patijn J, Remvig L. . 2007. "Reproducibility And Validity 2007-Protocol Formats For Diagnostics Procedures In Manual/Musculoskeletal Medicine." In.

26. Downie, W. W., P. A. Leatham, V. M. Rhind, V. Wright, J. A. Branco, and J. A. Anderson. 1978. 'Studies with pain rating scales', *Annals of the Rheumatic Diseases*, 37: 378-81.

For peer review only

27. Kirkley, A., S. Griffin, H. McLintock, and L. Ng. 1998. 'The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI)', *Am J Sports Med*, 26: 764-72.

28. 'World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects'. 2000. *JAMA*, 284: 3043-5.

29. Bahk, M., E. Keyurapan, A. Tasaki, E. L. Sauers, and E. G. McFarland. 2007. 'Laxity testing of the shoulder: a review', *Am J Sports Med*, 35: 131-44.

30. Sim, J., and C. C. Wright. 2005. 'The kappa statistic in reliability studies: use, interpretation, and sample size requirements', *Phys Ther*, 85: 257-68.

31. Byrt, T., J. Bishop, and J. B. Carlin. 1993. 'Bias, prevalence and kappa', *J Clin Epidemiol*, 46: 423-9.

32. Hoehler, F. K. 2000. 'Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity', *J Clin Epidemiol*, 53: 499-503.

33. Landis, J. R., and G. G. Koch. 1977. 'The measurement of observer agreement for categorical data', *Biometrics*, 33: 159-74.

For peer review only

BMJ Open

Inter-tester reliability of clinical shoulder instability and laxity tests in subjects with and without self-reported shoulder problems

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018472.R1
Article Type:	Research
Date Submitted by the Author:	06-Oct-2017

Complete List of Authors:	Eshoj, Henrik; Syddansk Universitet Det Sundhedsvidenskabelige Fakultet, Department of Sports Science and Clinical Biomechanics; Odense Universitetshospital, Quality of Life Research Center, Department of Haematology Ingwersen, Kim Gordon; Hospital Lillebaelt - Vejle Hospital, Vejle, Denmark, Research Unit at Department of Physical- and Occupational Therapy,; Syddansk Universitet Det Sundhedsvidenskabelige Fakultet, Department of Sports Science and Clinical Biomechanics Larsen, Camilla; Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics Kjær, Birgitte; Bispebjerg and Frederiksberg University Hospital, Copenhagen, Denmark , Research Unit at Department of Physical- and Occupational Therapy, ; Syddansk Universitet Det Sundhedsvidenskabelige Fakultet, Department of Sports Science and Clinical Biomechanics Juul-Kristensen, Birgit; Research Unit for Musculoskeletal Function and Physiotherapy, Institute of Sports Science and Clinical Biomechanics

Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Sports and exercise medicine
Keywords:	Reliability, Shoulder < ORTHOPAEDIC & TRAUMA SURGERY, Instability, Laxity, Clinical test

SCHOLARONE™
Manuscripts

Inter-tester reliability of clinical shoulder instability and laxity tests in subjects with and without self-reported shoulder problems

Henrik Eshoj, PhD [a] Kim Gordon Ingwersen, PhD [b,c], Camilla Marie Larsen, PhD [b,d] Birgitte Hougs Kjær, Msc [b,e], Birgit Juul-Kristensen, PhD [b,f]

Affiliations
[a] Quality of Life Research Center, Department of Haematology, Odense University Hospital, Odense, Denmark
[b] Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark
[c] Research Unit at Department of Physical- and Occupational Therapy, Hospital Lillebaelt - Vejle Hospital, Vejle, Denmark
[d] Health Sciences Research Centre, University College Lillebaelt, Odense, Denmark
[e] Research Unit at Department of Physical- and Occupational Therapy, Bispebjerg and Frederiksberg University Hospital, Copenhagen, Denmark
[f] Institute of Occupational Therapy, Physiotherapy and Radiography, Department of Health Sciences, Bergen University College, Norway

Correspondence:

Birgit Juul-Kristensen

Department of Sports Science and Clinical Biomechanics

University of Southern Denmark, Campusvej 55, 5230 Odense M, DK

Tel.: +45 2942 3335

Email: bjuul-kristensen@health.sdu.dk

Word counts: 3001

For peer review only

Abstract

Objective: Firstly, to investigate the inter-tester reliability of clinical shoulder instability and laxity tests, and secondly, to describe the mutual dependency of each test evaluated by each tester for identifying self-reported shoulder instability and laxity.

Method: A standardized protocol for conducting reliability studies was used to test the inter-tester reliability of the six clinical shoulder instability and laxity tests; apprehension, relocation, surprise, load-and-shift, sulcus sign and Gagey. Cohens kappa (*k*) with 95% confidence intervals (CI) besides Prevalence-Adjusted-Bias-Adjusted-Kappa (PABAK), accounting for insufficient prevalence and bias, were computed to establish the inter-tester reliability and mutual dependency.

Results: Forty individuals (13 with self-reported shoulder instability and laxity related shoulder problems and 27 normal shoulders individuals) aged 18-60 were included. Fair (relocation), moderate (load-and-shift, sulcus sign) and substantial (apprehension, surprise, Gagey) inter-tester reliability were observed across tests (*k* 0.39-0.73; 95% CI: 0.00-1.00). PABAK improved reliability across tests, resulting in substantial to almost perfect inter-tester reliability for the apprehension, surprise, load-and-shift and Gagey tests (*k* 0.65-0.90). Mutual dependencies between each test and self-reported shoulder problem showed apprehension, relocation and surprise to be the most often used tests to characterize self-reported shoulder instability conditions.

Conclusions: Four tests (apprehension, surprise, load-and-shift and Gagey) out of six, were considered inter-tester reliable for clinical use, while relocation and sulcus sign tests need further

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

standardization before acceptable evidence. Furthermore, the validity of the tests for instability and laxity needs to be studied.

Article summary section

Strengths and limitations of this study

- The strength of the study is the use of a three-phased standardized study protocol
- Presentation of raw findings increases transparency and interpretation of study findings
- No valid gold standard for including shoulder instability and laxity subjects was used
- A 50/50 prevalence of positive and negative tests for all six tests was not accomplished

Keywords

Reliability, Shoulder instability, laxity, clinical tests

Acknowledgements

We would like to thank Physiotherapists Rasmus Fitzner, Pernille Madsen and Jacob Hansen from Metropolitan University College, Copenhagen, Denmark for recruitment and testing of study participants. Furthermore, a special thanks to Bispebjerg Frederiksberg University Hospital, Copenhagen, Denmark for providing facilities for data collection.

For peer review only

Introduction

Shoulder complaints, affecting shoulder-related quality of life (QoL), are frequent and may be caused by shoulder instability and/or laxity [1] due to traumatic or non-traumatic injuries to the shoulder joint.[2] The traumatic shoulder instability is mainly prompted by a high impact injury during sports participation, resulting in a shoulder dislocation, predominantly in anterior direction.[3] The non-traumatic shoulder instability is usually related to repetitive overhead activities and/or patients with generalised joint hypermobility or glenohumeral hyperlaxity, often refered to as multidirectional shoulder instability.[2,4,5]

Irrespectively of aetiology, shoulder instability and laxity is often accompanied by a variety of symptoms including shoulder discomfort, pain besides glenohumeral subluxations and/or repeated dislocations.[6,7,8] Clinically, shoulder instability and laxity, are diagnosed and verified by a group of shoulder pain and instability provoking/relief tests, supplemented by shoulder laxity tests.[9,10] The former tests usually include the anterior shoulder instability and laxity tests; apprehension, relocation and surprise, and the laxity tests consisting of the load-and-shift, sulcus sign and Gagey tests.[11,12,13] An ongoing discussion is the use of pain as diagnostic criterion in diagnosing anterior shoulder instability with the clinical tests apprehension, relocation and surprise. [14,15,16] In one way, it may be a confounding factor, since pain has shown to be less predictive and reliable as a diagnostic criterion.[14] On the contrary though, others have suggested that unrecognized and underlying glenohumeral instability may lead to repetitive microtrauma and painful shoulder conditions,[15,16] justifying pain as diagnostic criterion when testing for anterior shoulder instability.

Nonetheless, symptoms may become chronic, and lead to reduced work and sports capability,[17,18,19] and with exercise-based management as the most often recommended first-

For peer review only

choice treatment.[20,21] Hence, early diagnosis using reliable and accurate clinical tests to guide focused treatment is essential. Few studies though, have investigated the reliability of clinical shoulder instability and laxity tests showing large variations in reliability and with limited methodological quality, hampering interpretation and comparison with other studies.[14,22,23]

Therefore, the objective of this study was to investigate the inter-tester reliability of commonly used clinical shoulder instability and laxity tests and secondly to describe the mutual dependency for each test evaluated by each tester, in a group of sports-active individuals with and without self-reported shoulder problems.

For peer review only

Materials and methods

Study design

An inter-tester reliability study was conducted involving two physiotherapists as inter-tester examiners. A third physiotherapist (study coordinator), not involved in the actual inter-tester reliability study (test phase), managed all practical aspects during the study period. The Guidelines for Reporting Reliability and Agreement Studies (GRASS), a consensus document on how to report reliability and agreement studies, were followed.[24] A standardized protocol for reliability studies, consisting of three phases: preparation and training of clinical tests, overall agreement, and test phase (the actual reliability study) were applied.[25] Two early career physiotherapists with six months clinical experience were involved in the inter-tester reliability study. A test-protocol describing each clinical test was developed and subsequently used by the two testers to practice all tests in order to reach uniformity and mutual agreement in performing and interpreting each test. In the overall agreement phase, the two testers examined 19 individuals (eight affected shoulders and 11 normal shoulders). The two testers were mutually blinded to the health status of the individuals (affected shoulders vs. normal shoulders) and also to each other's test results. Before proceeding to the final study phase, the two testers needed an overall agreement of at least 80% based on findings from the six clinical shoulder tests.[25] In the actual inter-tester reliability test phase, the two testers examined a new group of individuals with affected, respectively normal shoulders with the six clinical shoulder tests. The procedure was the same as in the agreement phase, meaning that testers were blinded to the health status of the individuals and each other's test results.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Study subjects

A sample size of at least 40 individuals was targeted based on recommendations for performing clinical reliability studies.[25] Sixty-five individuals (women and men (aged 18-60 years)) were recruited and screened for eligibility from Metropolitan University College, Copenhagen, and Bispebjerg Frederiksberg University Hospital, Copenhagen resulting in an included number of 13 individuals with instability and/or laxity related shoulder problems (hereinafter referred to as shoulder affected) vs. 27 normal shoulder individuals, respectively.

Shoulder affected individuals answering yes to at least one of two questions ('*Do you have a sense of shoulder instability*?' and '*Have you ever had a shoulder injury*?') were eligible for a clinical shoulder examination performed by the study coordinator. The shoulder affected individuals were then included if they present with at least one positive clinical shoulder test out of the following; apprehension, relocation, surprise, load-and-shift, sulcus sign or Gagey. Individuals with normal shoulders were recruited through public advertisements followed by a telephone interview and included if they present with no self-reported shoulder pathology or complaints. In general, any individuals with prior shoulder surgery were excluded. In the actual test phase, individuals completed a short questionnaire with basic demographic details (age, gender, weight, height), in addition to the following: pain level during rest and activity (numeric pain rating scale, NPRS),[26] shoulder injury ever (yes/no), subjective shoulder instability (yes/no) and sports-related activity (hours/week). Further, all individuals completed the patient-reported Western Ontario Shoulder Instability (WOSI) questionnaire designed to measure shoulder function and quality of life in patients with shoulder instability and laxity symptoms.[27] The time period between each test phase was approximately 2 weeks, and new subjects were included for each phase. Only the study phase is reported in the current manuscript. The study was exempted for notification to the Danish

Health Research Study Board due to the non-invasive and non-treating study design. However, oral and written consent was provided from all individuals and, ethical guidelines were followed according to the Helsinki declaration.[28]

Clinical tests

The clinical shoulder tests consisted of three shoulder joint-provoking tests for anterior shoulder instability (apprehension, relocation and surprise) besides three shoulder laxity tests (load and shift, sulcus sign and Gagey) (Table 1). [11,13,14,22,23,29]

The apprehension test (Table 1, Figure 1) was positive if glenohumeral apprehension and/or pain were evoked during testing whereas relief of symptoms with the relocation test (Table 1, Figure 2) was regarded as a positive test. As for the apprehension, the surprise test (Table 1, Figure 3) was positive if glenohumeral apprehension and/or pain were evoked during testing. The load and shift test (Table 1, Figure 4 & 5) was rated on a four-point scale ranging from 0-3 (best to worst; 0= little glenohumeral movement; 3=humeral head moves beyond the glenoid rim and remains dislocated).[12] Also, to enhance mutual agreement between testers when performing the load-and-shift test, only the direction (anterior vs. posterior) with most glenohumeral head translation was scored. Sulcus sign (Table 1, Figure 6) was objectively measured in centimeter (continuous scale) by use of a small ruler according to previously used grading scales as follows; I (<1 cm translation), II (1 to 2.0 cm translation) or III (>2.0 cm translation).[29] Finally, Gagey test (Table 1, Figure 7) was rated as positive with passive abduction above 105 degrees.[13]

Statistics

Demographics and descriptive data were tested for normality by visual inspection of histograms

For peer review only

176 and Shapiro-Wilk´s test. Group differences (affected shoulders vs. normal shoulders) were tested
177 by Fisher´s Exact test for categorical variables, whereas student´s t-test and Mann-Whitney U-test
178 was used for parametric, respectively non-parametric distributed data.
179 Apprehension, relocation, surprise and Gagey tests were dichotomous variables whereas the load
180 and shift and sulcus sign tests were dichotomized to also allow for nominal statistics. Thus, load-
181 and-shift was rated positive when scored 2 or 3, while for sulcus sign a positive rating was equal to
182 measurements exceeding one centimeter.[29] For transparency, data from each test is presented by
183 2 x 2 contingency tables besides the use of McNemar´s test for significant between-tester
184 differences. Furthermore, observed and expected agreements are presented along with
185 prevalence and bias[30] indexes. Reliability was evaluated with the use of Cohen´s kappa (*k*)
186 coefficients including 95% confidence intervals (CI). [25] Also, since kappa is sensitive to imbalances
187 in prevalence and bias (e.g. if a 50/50 distribution of positive and negative tests cannot be
188 accomplished) the use of PABAK calculation is a valid supplement to the original kappa values.[30,31]
189 By definition, PABAK reflects the ideal situation, thereby accounting for variation of prevalence
190 and bias between testers (as presented in the "real" world). [32] PABAK calculation is performed by
191 adjusting for high or low prevalence by computing the average of cells a and d in a cross table,
192 substituting this value for the actual values in those cells. Similarly, an adjustment for bias is
193 achieved by substituting the mean of cells b and c for those actual cell values.[30] Finally, the
194 relationship for each tester between the individual tests and the classification (mutual
195 dependency) by self-reported shoulder problems was tested by Cohen´s kappa (*k*) coefficients and
196 the characterization of the groups was tested with Fischer's exact tests.
197 The classification system proposed by Landis and Koch was used to interpret reliability as follows:
198 0.00-0.20 (Slight); 0.21-0.40 (Fair); 0.41-0.60 (Moderate); 0.61-0.80 (Substantial) and 0.81-1.00

For peer review only

(Almost perfect).[33]

Statistical Package for the Social Sciences (SPSS inc., Chicago, IL, USA), version 22, was used for all statistical analyses, with p-value of <0.05 interpreted as significant.

For peer review only

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Table 1. Performance and evaluation of the clinical shoulder instability and laxity tests

Verbal introduction:
1. I am going to perform six clinical shoulder tests on you
2. I will ask if you experience any symptoms (apprehension and/or pain) during the three first tests.
3. I will guide you through each test

Clinical tests	Description	Placing of hands, etc.	Evaluation
Figure 1	Individuals placed supine with the shoulder being tested close to the edge of the examination table. Shoulder positioned in 90˚of abduction, elbow flexed to 90˚. Examiner moves the shoulder into maximal external rotation.	One hand around the wrist of the individual with the other hand gently placed in front of the shoulder. Elbow supported at the examiners thigh.	Subjective or objective presence of apprehension and/or pain? Rated as either positive or negative. (Nominal, dichotome data)
Figure 2	From the end position of the apprehension test the humeral head is gently forced posteriorly	Examiners fifth finger placed close to the lateral part of the acromion with the wrist positioned anteriorly at the humeral head.	Relief of apprehension and/or pain? Rated as either positive or negative. (Nominal, dichotome data)
Figure 3	From end position of the relocation test the posteriorly directed force at the humeral head is quickly removed.	Removal of examiners wrist from the anterior part of the shoulder.	Subjective or objective reproduction of apprehension and/or pain? Rated as either positive or

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

			negative. (Nominal, dichotome data)
Figure 4 Figure 5	Individual placed supine with scapula resting at the examination table. Humeral head is loaded gently into the glenoid through axial pressure at the elbow. *Anterior load-and-shift test:* Shoulder positioned in the scapular plane in 90˚ of abduction with elbow flexed. Humeral head gently shifted in anterior direction. *Posterior load-and-shift test:* Shoulder positioned in the scapular plane in 20˚ of abduction with elbow flexed. Humeral head gently shifted in posterior direction.	Examiners one hand placed at the olecranon with the individual's hand positioned between the examiners torso and elbow. *Anterior load-and-shift:* Examiners hand placed on top of the shoulder with the fingers on the backside of the glenohumeral head to move it anteriorly. *Posterior load-and-shift:* Examiners wrist placed at the anterior part of the humeral head to move it posteriorly.	Humeral head movement evaluated by the use of a four-level laxity scale. 0 = little to almost no movement 1 = humeral head moves up onto the glenoid 2 = humeral head moves beyond the glenoid, but relocates spontaneously once pressure is released 3 = humeral head moves beyond the glenoid and remains dislocated Rated as positive when scored 2 or 3. (Nominal, dichotomous data)
Figure 6	Individuals sitting upright. Shoulder in neutral position (0 degree rotation).	One hand placed above the epicondyles of humerus.	Rated as positive with measurements exceeding 1 centimeter.

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Inter-tester reliability of clinical shoulder instability and laxity tests

	Examiner pulls the distal part of the humerus in a caudal direction. Distance from the top of the humeral head and the acromion is evaluated with a ruler.	Other hand is used to evaluate the subacromial distance with a ruler.	(Nominal, dichotome data)
Figure 7	Individuals sitting upright. The shoulder girdle is stabilized by examiners forearm preventing the shoulder girdle to elevate while the individuals arm is passively moved into end range in horizontal abduction. A mirror in front of the individual is used to evaluate the shoulder abduction angle.	Forearm placed on top of the shoulder girdle with the other hand placed around the elbow joint.	Rated as positive with abduction exceeding 105 degrees. (Nominal, dichotomous data)

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

Results

Characteristics of the participating individuals are presented in Table 2. Demographics showed no difference between the individuals with affected shoulders (n=13) and normal shoulders (n=27). Furthermore, both groups (92 and 74%; p=0.18) were relatively active with a weekly participation in sports-related activity for more than four hours per week. However, as expected due to the design, affected shoulders had significantly higher pain during activity (4.23 vs. 1.44; p=0.02), higher frequency of shoulder injury ever (62% vs. <1%; p<0.001), higher subjective shoulder instability (69 vs. 11%; p<0.001) and worse total WOSI score (506 vs. 136; p=0.001) (Table 2).

Table 2. Study phase. Participant characteristics.

	Affected shoulders (n=13)	**Normal shoulders (n=27)**	***P-value***
Sex (women/men)	8/5	21/6	0.28
Age (years) mean (SD)	28 (9)	29 (7)	0.72
Weight (kg)* mean (SD)	71.0 (12.8)	74.9 (23.4)	0.59
Height (cm) mean (SD)	174.0 (8.6)	173.4 (7.9)	0.82
Pain, rest (NRS 0-10) mean (SD)	1.08 (1.44)	0.41 (1.15)	0.12
Pain, activity (NRS 0-10) mean (SD)	4.23 (2.92)	1.44 (2.12)	<0.05
Shoulder injury ever, n (%)	8 (62)	1 (4)	<0.001
Subjective shoulder instability, n (%)	9 (69)	3 (11)	<0.001
Sports-related activity (>4 hours/week), n (%)	12 (92)	20 (74)	0.18
WOSI domains, mean (SD)			
- Physical symptoms (0-1000)	225 (165)	60 (78)	<0.05
- Sports, recreation, work (0-400)	103 (93)	24 (47)	<0.05
- Lifestyle (0-400)	58 (57)	13 (21)	<0.05
- Emotions (0-300)	121 (94)	39 (49)	<0.05
WOSI total score (0-2100) mean (SD)	506 (362)	136 (174)	<0.001

SD **Standard deviation;** ***kg*** **kilo; * significance level p < 0.05;** ***cm*** **centimeter** ***NRS*** **Numeric Rating Scale**

For peer review only

Prevalence of positive tests was especially low for the load-and-shift test (table 3), and significant between-tester differences were found for relocation and sulcus sign tests (p=0.021) (not shown in tables).

Table 3. Contingency tables with findings from tester A and B.

Apprehension		**A** Yes	**A** No
B	Yes	14	4
	No	3	19

Relocation		**A** Yes	**A** No
B	Yes	6	2
	No	8	24

Surprise		**A** Yes	**A** No
B	Yes	14	4
	No	3	19

Load-and-shift		**A** Yes	**A** No
B	Yes	1	0
	No	2	37

Sulcus		**A** Yes	**A** No
B	Yes	7	1
	No	9	23

Gagey		**A** Yes	**A** No
B	Yes	8	3
	No	1	28

Reliability varied between *k*: 0.39-0.73 (95% CI: 0.00-1.00), indicating fair (relocation; *k* 0.39), moderate (load-and-shift, sulcus sign; *k* 0.43 and 0.48) and substantial (apprehension, surprise, Gagey; *k* 0.65-0.73) reliability (Table 4). The prevalence index of all six tests ranged from 0.05-0.44, (lowest for load-and-shift, relocation and sulcus; 0.05, 0.28 and 0.30), whereas the bias index ranged from and 0.03-0.20 (highest for relocation and sulcus). PABAK improved reliability for relocation, load-and-shift, sulcus sign and Gagey test, now corresponding to moderate (relocation and sulcus sign; *k* 0.50), substantial (Gagey; *k*: 0.80) and almost perfect (load and shift; *k*: 0.90) reliability (Table 4).

Table 4. Reliability of six clinical shoulder instability and laxity tests.

	Observed agreement	Expected agreement	Prevalence index	Bias index	Kappa (95% CI)	PABAK
Apprehension	0.83	0.51	0.44	0.03	0.65 (0.38; 0.85)	0.65
Relocation*	0.75	0.59	0.28	0.15	0.39 (0.07; 0.68)	0.50
Surprise	0.83	0.51	0.44	0.03	0.65 (0.38; 0.85)	0.65
Load and Shift	0.95	0.90	0.05	0.05	0.48 (0.00; 1.00)	0.90
Sulcus sign*	0.75	0.56	0.30	0.20	0.43 (0.17; 0.72)	0.50
Gagey	0.90	0.62	0.40	0.05	0.73 (0.46; 0.94)	0.80

*****_Significant inter-tester differences_, **CI** Confidence intervals, **PABAK** _prevalence-and-bias-adjusted kappa_

The *k* values for mutual dependency indicate that apprehension, relocation and surprise tests for both examiners were the most frequently used tests for characterizing self-reported shoulder problems (Table 5). This was further confirmed by the significant group difference in the presence of positive tests.

Table 5. Kappa statistics for mutual dependency of the individual tests and self-reported shoulder problems for each tester.

	Observed agreement	Expected agreement	Prevalence index	Kappa	p-value (AS/NS)
Apprehension					
Examiner A	0.75	0.53	0.38	0.47	0.003
Examiner B	0.68	0.52	0.40	0.33	0.04
Relocation*					
Examiner A	0.83	0.55	0.35	0.61	<0.001
Examiner B	0.73	0.63	0.25	0.27	0.08
Surprise					
Examiner A	0.75	0.53	0.38	0.47	0.003
Examiner B	0.68	0.52	0.40	0.33	0.04
Load and Shift					
Examiner A	0.75	0.65	0.20	0.29	0.03
Examiner B	0.70	0.67	0.18	0.10	0.33
Sulcus sign*					
Examiner A	0.63	0.61	0.28	0.05	0.52
Examiner B	0.63	0.54	0.38	0.19	0.19
Gagey					
Examiner A	0.70	0.60	0.28	0.26	0.10
Examiner B	0.70	0.58	0.30	0.29	0.08

AS affected shoulders; **NS** normal shoulder; *****_Significant inter-tester differences_

For peer review only

Discussion

The inter-tester reliability across the selected six clinical shoulder instability and laxity tests ranged from fair to substantial. Use of PABAK calculations improved inter-tester reliability to substantial and almost perfect across most tests, except for the relocation and sulcus sign tests. The tests most often used to characterize self-reported shoulder instability and laxity (mutual dependency) were apprehension, relocation and surprise tests.

The inter-tester reliability for the apprehension, relocation and surprise was higher than, or equivalent, to previously reported results of these tests using the same diagnostic procedures (apprehension and/or pain).[23] Specifically for the apprehension and surprise test, the present *k* values were somewhat higher than previously reported (0.65 vs. 0.44-0.45). The reason for this may be that the current study included both affected and normal shoulder individuals as opposed to only including symptomatic subjects.[23] This may have increased subject variation, known to affect reliability positively. Also, PABAK calculations did not affect the overall reliability of the apprehension and surprise tests, probably due to an optimal prevalence index of positive and negative tests (close to 0.50). For the relocation test, the existing inter-tester reliability was almost similar to previously reported (*k* 0.39 vs. 0.44),[23] however lower. Apparently, the primary reason for the current poor reliability in relocation was presence of systematic bias between testers, as indicated by the actual raw data (contingency tables) and the statistical significant inter-examiner difference. Likewise, systematic bias between testers was also found for the sulcus sign test in the present

For peer review only

study. Hypothetically, this may be explained by inter-tester variability in the force produced to translate the humeral head in posterior (relocation test) or inferior (sulcus sign test) direction, in the current study. This is, however, only speculative and further studies are needed to standardize these tests.

Reliability for the present sulcus sign test was slightly lower than previously reported (*k* 0.39 vs. >0.50).[22,23] The discrepancy in reliability observed may be due to the use of different test positions with participants in the current study sitting upright[29] as opposed to a previous lying test position.[22] However, due to the presence of systematic bias in both the relocation and sulcus sign test, PABAK did not affect the overall reliability much.

For the load-and-shift test, reliability was relatively low (including wide CI). This may be due to the current low prevalence index below 50%, which is the optimum prevalence in reliability studies. [25] However, the present dichotomous rating of the load-and-shift test (meaning that only individuals that could either subluxate or dislocate the shoulder during testing was deemed positive) may have influenced the prevalence of positive tests largely. Therefore, using PABAK, reliability of the load-and-shift test improved considerably (from moderate to almost perfect). Nevertheless, different statistics (kappa vs. Intra-class-Correlation Coefficients), different scoring systems (dichotomous rating (positive yes/no) vs. four point grading scale (0-3) [23] and inclusion of shoulder asymptomatic athletes only[22] make comparison across studies difficult.

Finally, reliability of the Gagey test was substantial and PABAK did not affect reliability much due to a nearly optimal prevalence and low bias between testers. Unfortunately, there is no other study to compare with.

For peer review only

289 Although the current study was designed to investigate reliability, and not diagnostic accuracy, the mutual dependency between the individual tests and self-reported shoulder problems was analysed. It revealed that the tests most often used to characterize those with and without self-reported shoulder instability and laxity (mutual dependency) proved to be the apprehension, relocation and surprise tests. This may indicate a relationship between these tests, which may come as no surprise, since these tests are a continuum of the apprehension test and, thus, closely related.[9] Nevertheless, for clinicians it is of interest to specify the clinical characteristics of patients with self-reported shoulder problems. Thus, the current prevalence of positive tests may mirror these characteristics of the included patients and should be taken into consideration in the management of such musculoskeletal conditions. It is recommended to develop and test the clinimetric properties of a more comprehensive test battery for evaluating such self-reported shoulder problems. No prior studies were found addressing mutual dependency of the current tests for shoulder instability and laxity, which hampers comparison.

304 The present study has several limitations. Firstly, the lack of standardized measurement of the amount of force exerted by the two testers during especially the relocation and sulcus sign test may have limited the current inter-tester reliability. Further standardization in both performance and interpretation is therefore needed. Also, the current study did not randomize the order of the clinical tests. However, we do not believe this to have biased the reliability of the data, since the same order was used for both testers.

311 Secondly, no valid gold standard for classifying shoulder instability and laxity was used.

For peer review only

To compensate for this, self-reported confirmation of shoulder-related problems was applied, but this was not reflected in the current WOSI scores, which were relatively low. Lack of a more objective gold standard may have decreased diagnostic accuracy, however not reliability, which was the primary objective of the present study. Also, in the group with normal shoulders, one individual reported to have had a previous shoulder injury and three individuals reported subjective shoulder instability, which does not comply with the inclusion criteria for being regarded as shoulder healthy in the current study. At the clinical session, a self-reported questionnaire was completed regarding demographic data and historical information. Apparently, in the baseline questionnaire three shoulder healthy individuals answered yes to perceiving instability in their shoulder and one had had a previous shoulder injury, even though they all had reported no shoulder trouble during the telephone inclusion interview. However, as depicted in table 2, WOSI and pain scores in the group with normal shoulders seem not to be influenced severely by these four individuals. Also, re-calculations of demographic data and mutual dependency with the revised classification into affected/normal shoulders did not change the mutual dependency of the most frequently used tests for classification into affected/normal shoulders, and neither was kappa and demographics affected (data not shown).

Thirdly, due to a relative short recruitment period besides difficulties in recruiting subjects with shoulder instability and laxity only thirteen subjects with an affected shoulder were included. Naturally, this also affected the prevalence of positive and negative test findings meaning that the prevalence of 0.50, as recommended in reliability studies,[25] in all six tests was not accomplished. However, to overcome this,

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

PABAK calculations was used and reported along with kappa, to show transparently how data would have been with equal distributions of positive and negative test results. Nevertheless, future studies should use inclusion criteria of more established shoulder instability and laxity conditions, and, if possible, verified by objective criteria as surrogate for a gold standard of shoulder instability and laxity. This may optimize prevalence as well as diagnostic accuracy in studies where this is a further aim.

The strengths of the study are the use of standardized procedures (including blinding to patient status and the use of a three-phased protocol for conducting reliability studies). Also, presentation of raw data, using contingency tables, along with kappa and PABAK values, increases data transparency and improves interpretation of the reliability study.

Conclusions

This study showed acceptable inter-tester reliability for four of six clinical shoulder instability and laxity tests in relatively sports active individuals with and without self-reported shoulder problems. However, relocation and sulcus sign tests need further standardisation before being recommended for use in clinical practice. Based on the frequency and mutual dependency of the current tests, especially apprehension and surprise tests seem important in the characterisation of self-reported shoulder problems. Future research on the validity of tests for shoulder instability and laxity is needed.

For peer review only

Author statement: HE, KGI, CML and BJK conceived and designed the study. HE and BHK recruited study participants and collected data. HE performed the statistical analysis. HE, KGI, CML, BJK interpreted the results. HE drafted the manuscript with KGI, CML, BHK and BJK contributing to the manuscript. All authors have read and approved the final manuscript. HE is the guarantor.

Consent for publication: The depicted individuals in Table 1 have provided written consent for publication of the pictures.

Funding: This work was supported by Region of Southern Denmark's Research fund and The Danish Rheumatism Association.

Competing interest: None declared

Data sharing statement: No additional data are available

For peer review only

References

1. Luime, J. J., B. W. Koes, I. J. Hendriksen, A. Burdorf, A. P. Verhagen, H. S. Miedema, and J. A. Verhaar. 2004. 'Prevalence and incidence of shoulder pain in the general population; a systematic review', *Scandinavian Journal of Rheumatology*, 33: 73-81.

2. Farrar, N. G., J. J. Malal, J. Fischer, and M. Waseem. 2013. 'An overview of shoulder instability and its management', *Open Orthopaedics Journal*, 7: 338-46.

3. Goss, T. P. 1988. 'Anterior glenohumeral instability', *Orthopedics*, 11: 87-95.

4. Cameron, K. L., M. L. Duffey, T. M. DeBerardino, P. D. Stoneman, C. J. Jones, and B. D. Owens. 2010. 'Association of generalized joint hypermobility with a history of glenohumeral joint instability', *J Athl Train*, 45: 253-8.

5. Chahal, J., J. Leiter, M. D. McKee, and D. B. Whelan. 2010. 'Generalized ligamentous laxity as a predisposing factor for primary traumatic anterior shoulder dislocation', *J Shoulder Elbow Surg*, 19: 1238-42.

6. Dodson, C. C., and F. A. Cordasco. 2008. 'Anterior glenohumeral joint dislocations', *Orthop Clin North Am*, 39: 507-18, vii.

7. Jaggi, A., and S. Lambert. 2010. 'Rehabilitation for shoulder instability', *Br J Sports Med*, 44: 333-40.

8. Wilk, K. E., L. C. Macrina, and M. M. Reinold. 2006. 'Non-operative rehabilitation for traumatic and atraumatic glenohumeral instability', *N Am J Sports Phys Ther*, 1: 16-31.

9. Hegedus, E. J., A. P. Goode, C. E. Cook, L. Michener, C. A. Myer, D. M. Myer, and A. A. Wright. 2012. 'Which physical examination tests provide clinicians with the most value when examining the shoulder? Update of a systematic review with meta-analysis of individual tests', *Br J Sports Med*, 46: 964-78.

10. Ann Ganestam, Mikkel L. Attrup, Per Hølmich and Kristoffer W Barfoed. 2015. 'Evaluation of Clinical Practice of Shoulder Examination amon Ten Experienced Shoulder Surgeons', *Journal of Orthopaedic Research and Physiotherapy*, 1.

11. Luime, J. J., A. P. Verhagen, H. S. Miedema, J. I. Kuiper, A. Burdorf, J. A. Verhaar, and B. W. Koes. 2004. 'Does this patient have an instability of the shoulder or a labrum lesion?', *JAMA*, 292: 1989-99.

12. Tzannes, A., and G. A. Murrell. 2002. 'Clinical examination of the unstable shoulder', *Sports Medicine*, 32: 447-57.

13. Gagey, O. J., and N. Gagey. 2001. 'The hyperabduction test', *J Bone Joint Surg Br*,

For peer review only

83: 69-74.

14. Lo, I. K., B. Nonweiler, M. Woolfrey, R. Litchfield, and A. Kirkley. 2004. 'An evaluation of the apprehension, relocation, and surprise tests for anterior shoulder instability', *Am J Sports Med*, 32: 301-7.

15. Silliman, J. F., and R. J. Hawkins. 1993. 'Classification and physical diagnosis of instability of the shoulder', *Clin Orthop Relat Res*: 7-19.

16. Cavallo, R. J., and K. P. Speer. 1998. 'Shoulder instability and impingement in throwing athletes', *Med Sci Sports Exerc*, 30: S18-25.

17. Matsen, F. A., 3rd, and J. D. Zuckerman. 1983. 'Anterior glenohumeral instability', *Clin Sports Med*, 2: 319-38.

18. Robinson, C. M., J. Howes, H. Murdoch, E. Will, and C. Graham. 2006. 'Functional outcome and risk of recurrent instability after primary traumatic anterior shoulder dislocation in young patients', *J Bone Joint Surg Am*, 88: 2326-36.

19. van Kampen, D. A., T. van den Berg, H. J. van der Woude, R. M. Castelein, C. B. Terwee, and W. J. Willems. 2013. 'Diagnostic value of patient characteristics, history, and six clinical tests for traumatic anterior shoulder instability', *J Shoulder Elbow Surg*, 22: 1310-9.

20. Warby, S. A., T. Pizzari, J. J. Ford, A. J. Hahne, and L. Watson. 2014. 'The effect of exercise-based management for multidirectional instability of the glenohumeral joint: a systematic review', *J Shoulder Elbow Surg*, 23: 128-42.

21. Gibson, K., A. Growse, L. Korda, E. Wray, and J. C. MacDermid. 2004. 'The effectiveness of rehabilitation for nonoperative management of shoulder instability: a systematic review', *J Hand Ther*, 17: 229-42.

22. Levy, A. S., S. Lintner, K. Kenter, and K. P. Speer. 1999. 'Intra- and interobserver reproducibility of the shoulder laxity examination', *Am J Sports Med*, 27: 460-3.

23. Tzannes, A., A. Paxinos, M. Callanan, and G. A. Murrell. 2004. 'An assessment of the interexaminer reliability of tests for shoulder instability', *J Shoulder Elbow Surg*, 13: 18-23.

24. Kottner, J., L. Audige, S. Brorson, A. Donner, B. J. Gajewski, A. Hrobjartsson, C. Roberts, M. Shoukri, and D. L. Streiner. 2011. 'Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed', *J Clin Epidemiol. 64 (1)*: 96-106.

25. Patijn J, Remvig L. . 2007. "Reproducibility And Validity 2007-Protocol Formats For Diagnostics Procedures In Manual/Musculoskeletal Medicine." In.

For peer review only

26. Downie, W. W., P. A. Leatham, V. M. Rhind, V. Wright, J. A. Branco, and J. A. Anderson. 1978. 'Studies with pain rating scales', *Annals of the Rheumatic Diseases*, 37: 378-81.

27. Kirkley, A., S. Griffin, H. McLintock, and L. Ng. 1998. 'The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI)', *Am J Sports Med*, 26: 764-72.

28. 'World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects'. 2000. *JAMA*, 284: 3043-5.

29. Bahk, M., E. Keyurapan, A. Tasaki, E. L. Sauers, and E. G. McFarland. 2007. 'Laxity testing of the shoulder: a review', *Am J Sports Med*, 35: 131-44.

30. Sim, J., and C. C. Wright. 2005. 'The kappa statistic in reliability studies: use, interpretation, and sample size requirements', *Phys Ther*, 85: 257-68.

31. Byrt, T., J. Bishop, and J. B. Carlin. 1993. 'Bias, prevalence and kappa', *J Clin Epidemiol*, 46: 423-9.

32. Hoehler, F. K. 2000. 'Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity', *J Clin Epidemiol*, 53: 499-503.

33. Landis, J. R., and G. G. Koch. 1977. 'The measurement of observer agreement for categorical data', *Biometrics*, 33: 159-74.

For peer review only

Figure legends

Figure 1. Apprehension

Figure 2. Relocation

Figure 3. Surprise

Figure 4. Load and shift – anterior direction

Figure 5. Load and shift – posterior direction

Figure 6. Sulcus sign

Figure 7. Gagey

For peer review only

Table 1, Figure 1

26x19mm (300 x 300 DPI)

For peer review only

Table 1, Figure 2

26x19mm (300 x 300 DPI)

For peer review only

Table 1, Figure 3

26x21mm (300 x 300 DPI)

For peer review only

Table 1, Figure 4

30x19mm (300 x 300 DPI)

Table 1, Figure 5

28x23mm (300 x 300 DPI)

review only

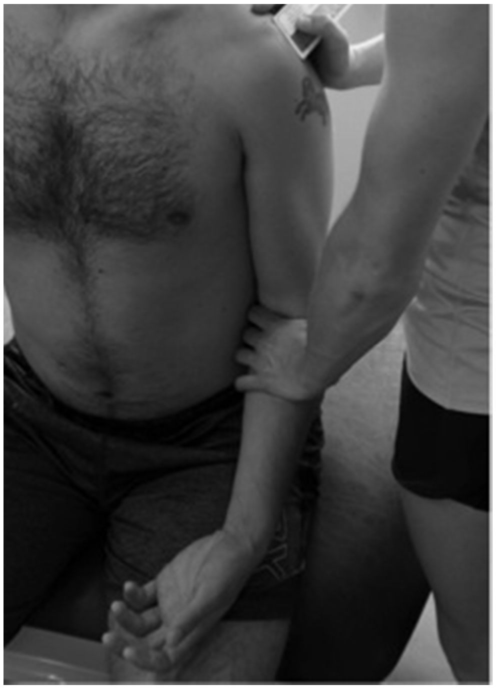


Table 1, Figure 6

20x28mm (300 x 300 DPI)

view only

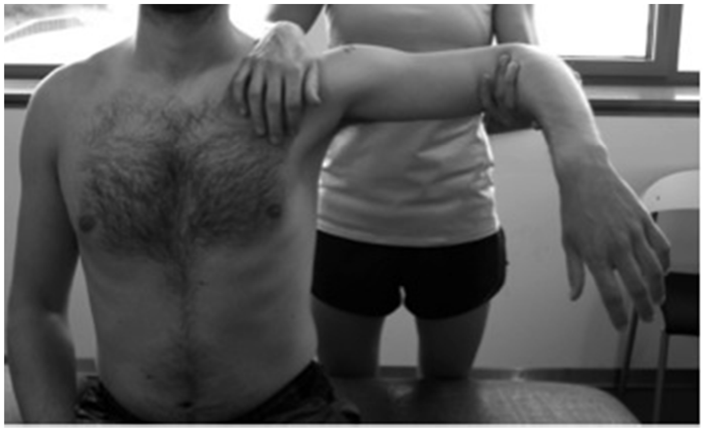


Table 1, Figure 7

29x17mm (300 x 300 DPI)

For peer review only