An assessment of the interexaminer reliability of tests for shoulder instability

Anthony Tzannes, BSc(Med), MBBS, Anastasios Paxinos, MBBS, FRACS, Mary Callanan, MBBS, and George A. C. Murrell, MBBS, DPhil, Sydney, Australia

Accurate noninvasive clinical tests of shoulder instability are important in assessing and planning treatment for glenohumeral joint instability. An interexaminer agreement trial was undertaken to estimate the reliability of commonly used clinical tests for shoulder instability. Thirteen patients with a history suggestive of instability, who had been referred to a shoulder specialist for treatment of their symptomatic shoulders, were examined by four examiners of differing experience. Good to excellent interexaminer agreement was found for most variations of the load-and-shift test, with the best agreement in the 90° abducted position for the anterior direction (intraclass correlation coefficient [ICC] = 0.72) and in the 0° abducted position for the posterior (ICC = 0.68) and inferior (ICC = 0.79) directions. Fair to good interexaminer reliability was found for the sulcus sign (ICC = 0.60). With regard to the provocative tests, agreement was best when apprehension was used as the criterion for a positive test and was better for the relocation (ICC = 0.71) and release tests (ICC = 0.63) than for the apprehension (ICC = 0.47) and augmentations tests (ICC = 0.48). Reliability was poor (ICC < 0.31) when pain was used as the criterion for a positive test. These results indicate that the load-and-shift, sulcus, and provocative tests (apprehension, augmentation, relocation, and release) are reliable clinical tests for instability in symptomatic patients when care is taken with respect to arm positioning and if apprehension is used as the criterion for a positive provocative test. (J Shoulder Elbow Surg 2004;13:18-23.)

The glenohumeral joint is the most commonly dislocated joint. Dislocation most frequently occurs in the anterior direction when the arm is forcibly abducted and externally rotated, and it is usually painful and disabling. Dislocations can also occur posteriorly, inferiorly, or in multiple directions and often require surgical intervention. Excess glenohumeral joint translation, without frank dislocation, can be asymptomatic but can also cause symptoms and impair performance. A number of clinical tests have been described to assess glenohumeral joint translation and to reproduce or provoke symptoms of glenohumeral joint instability. If the results of these tests are to be valuable for clinical prediction and screening, they must be both valid and reliable. Validity (lack of systematic error bias) is defined as the extent to which a measure correctly evaluates the desired trait. Validity was not addressed in this study. Reliability is defined as the extent to which a measure is repeatable and, therefore, represents the degree of nonsystematic error. Reliability can be estimated from measurements made by different raters on the same material (agreement). Two types of agreement can be distinguished according to whether one rater makes two or more measurements on the same material (intraexaminer) or whether each of several raters independently measures the same material (interexaminer).

The purpose of this study was to determine the intraexaminer reliability of 13 clinical tests for shoulder instability.

MATERIALS AND METHODS

The subjects included in this study were (1) patients whose shoulder was sufficiently symptomatic to warrant referral to an orthopaedic shoulder specialist and (2) those whose history was suggestive of or suspicious for glenohumeral joint instability. Patients were excluded if (1) there was a history of previous surgery or (2) they had significant loss of glenohumeral joint motion (ie, they had less than 30° external rotation, less than 90° forward flexion, or less than 90° abduction or could internally rotate their arm to no greater than L5).†

There were 25 patients who met the inclusion criteria for this study. Of these, 12 were excluded because the shoulder was too painful to be examined by all 4 examiners or...
the patient presented on a day when not all of the examiners were present to examine him or her, leaving 13 patients in the study. There were 10 male and 3 female patients with 9 right and 4 left shoulders. The median age at examination was 31 years, with an SD of 13 years. The age range was from 16 to 56 years. Eleven of the patients had a presumptive diagnosis of shoulder instability, and two had a presumptive diagnosis of impingement.

Each patient was assessed by 4 examiners (an orthopaedic shoulder specialist, an orthopaedic fellow, a sports medicine registrar, and a medical student) who were blinded to the results of previous examiners. All 4 examinations were performed on the same day, usually within a 30-minute period, to minimize day-to-day variation in the patient’s symptoms and signs among examiners. The examination technique and grading were standardized across the 4 examiners.

The glenohumeral joint requires a certain level of laxity to be able to maintain its range of movement. Laxity of a joint has been defined as a capacity for symptom-less translation. Shoulder instability has been defined as a pathologic condition that manifests as pain associated with excessive translation of the humeral head on the glenoid during active shoulder motion.

**Laxity tests**

The goal of the load-and-shift test is to load the humeral head into the glenoid and then to translate the humeral head anteriorly, posteriorly, and inferiorly. Several variations of this test have been described, with two main test positions—the patient seated or the patient supine—with minor variations in arm position for each position. Three variations of the test were used in this study—one in which the patient was seated with the arm resting at the side and two in which the patient was in the supine position with the arm in 20° or 90° abduction (Figure 1) in the scapular plane.

In the seated position, the examiner placed one hand over the patient’s shoulder with the fingers and thumb resting on the anterior and posterior joint line to stabilize the scapula while being in a position to feel any glenohumeral translation. The other hand then grasped the humeral head, loading it into the glenoid to ensure that it was centered before testing laxity in the anterior, posterior, and inferior directions.

In the supine position, the patient lay with the center of the scapula of the examined limb on the edge of the examining table to allow free movement of the glenohumeral joint while providing a degree of fixation of the scapulothoracic articulation. The examiner sat on a stool by the head of the patient and grasped the patient’s elbow with his corresponding hand. The other hand grasped the patient’s upper arm (Figure 1). The examiner then positioned the patient’s arm in 20° or 90° abduction in the scapular plane in neutral rotation and centered the patient’s humeral head on the glenoid by applying a load along the axis of the humerus with the hand that was grasping the patient’s elbow. The examiner then attempted to shift the patient’s humeral head off the glenoid in the anterior, posterior, and inferior directions.

The results for all versions of the test were graded as follows:

- grade 0, little to no movement of the humeral head;
- grade 1, the humeral head could be shifted so that it started to ride up onto the glenoid labrum;
- grade 2, the humeral head could be shifted off the glenoid but spontaneously relocated once the pressure had been eased; and
- grade 3, the humeral head could be shifted off the glenoid and remained dislocated once the pressure had been eased.

**Figure 1** Photograph of the load-and-shift test performed at 90° of abduction in the supine position. The patient was positioned in such a manner that the scapula was on the examining table, with the humeral head off the table, free to be translated anteriorly and posteriorly. The examiner held the patient’s elbow with one hand and gently loaded the humeral head into the glenoid. The examiner’s other hand held the patient’s upper arm and attempted to shift the patient’s humeral head off the glenoid in the anterior, posterior, and inferior directions. The results for all versions of the test were graded as follows:

---

**Prophetic tests**

- The apprehension test was carried out with the patient supine, in a position similar to that described for the supine variation of the load-and-shift test. The arm was then positioned in 90° abduction and external rotation. The arm was then further externally rotated and extended to the end range of motion (or until the patient’s request for the examiner to stop). The patient was then asked if he or she had pain or apprehension (that the shoulder would dislocate) (Figure 3, A).

---

- The relocation test was carried out immediately after the apprehension test. While the patient was still in the position that evoked symptoms, the humeral head was depressed in the posterior direction. In a positive test, the
patient’s symptoms would be reduced by this maneuver (Figure 3, B).

When the patient was still in the position of apprehension, an anteriorly directed force, augmentation,12,13 was applied to the posterior humeral head (Figure 3, C). The test was positive if the patient’s symptoms were made worse by the test.

The release test13 was carried out immediately after a positive relocation test. The examiner suddenly released the pressure pushing the humeral head posteriorly (while still holding the patient’s arm in the position of apprehension). A positive test was recorded if the patient’s symptoms worsened at the time of removal of the relocating pressure (Figure 3, D).

For each variation and direction of the load-and-shift test, sulcus sign, apprehension test, relocation test, release test, and augmentation test, the results of the 4 examiners were compared by use of 2-way mixed-effect intraclass correlation coefficients (ICCs)4,8,11 on SPSS for Windows V9 (SPSS Inc, Chicago, IL). This coefficient is derived from an analysis of variance model that incorporates both patient and examiner effects. The particular model used assumes that the examiners were a random sample of all possible examiners (random-effects model). Although the examiners were, in fact, selected by convenience, and the fixed-effects model should therefore apply, estimates of ICCs and other quantities obtained from this model are slightly more conservative than those obtained in the fixed-effects model. Values of ICC range from zero to unity. High levels of agreement (ICC approximately equal to 1) arise when the amount of variation resulting from the readers is smaller than the variability resulting from the sum of all sources of variation (subject, reader, and residual error). An important assumption in this model is that the measurements have consistent within-subject and within-reader variances.

RESULTS

Load-and-shift test

More translation in the anterior direction was found when the shoulder was abducted 20° and 90° than when tested in 0° abduction and with the patient sitting. The greatest agreement among examiners was obtained when the shoulder was abducted 90° (ICC = 0.72) (Table I).

Far less translation was noted by all examiners when the humeral head was translated posteriorly than when translated anteriorly. Most translation was noted with the arm at 0° and 90°. All examiners graded the translation as 0 when the arm was positioned at 20° abduction. Of the two positions in which posterior translation was noted, the best agreement (ICC = 0.68) was found when the arm was at 0° and the patient was seated (Table I).

Very little inferior translation was noted by all but one examiner, irrespective of arm position. The best agreement was found at 0° and 20° abduction (ICC = 0.74) (Table I). The mean sulcus noted was 0.5 cm (range, 0-1.5 cm) with fair to good agreement among examiners (ICC = 0.6, P < .0001) (Table I).

Provocative tests

For all of the provocative tests, the best agreement among examiners was found when apprehension was the criterion for a positive test. There was poor agreement when pain alone was the criterion and intermediate agreement when pain and/or apprehension were the criteria for a positive test (Table II).

For the apprehension test, the greatest correlation (ICC = 0.47, P < .0002) was found when the test was performed with apprehension only considered as a positive test. There was poor agreement when pain alone was the criterion (ICC = 0.31) and intermediate agreement when pain and/or apprehension were the criteria for a positive test (ICC = 0.44) (Table II).

For the relocation test, the greatest correlation (ICC = 0.71, P < .0001) was found when the test was performed with apprehension only considered as a positive test. There was poor agreement when pain alone was the criterion (ICC = 0.31) and intermediate agreement when pain and/or apprehension were the criteria for a positive test (ICC = 0.44) (Table II).

For the augmentation test, the greatest correlation...
(ICC = 0.48, P = .0003) was found when the test was performed with apprehension only considered as a positive test. There was very poor agreement when pain alone was the criterion (ICC = 0.09) and intermediate agreement when pain and/or apprehension were the criteria for a positive test (ICC = 0.33) (Table II).

For the release test, the greatest correlation (ICC = 0.63, P < .0001) was found when the test was performed with apprehension only considered as a positive test. There was poor agreement when pain alone was the criterion (ICC = 0.31) and intermediate agreement when pain and/or apprehension were the criteria for a positive test (ICC = 0.45) (Table II).

**DISCUSSION**

The major conclusion of this study was that, in this population of patients with a symptomatic shoulder and histories consistent with or suspicious for gleno-humeral joint instability, fair to excellent interexaminer reliability was obtained; however, arm position in the load-and-shift tests and the criteria for a positive provocative test were important.

The load-and-shift tests were more reliable with the patient in the supine position rather than sitting upright. In the anterior direction, the 90° abducted position was the most reliable (ICC = 0.72). In the posterior direction, the 20° abducted position was the
 assessments were made on 13 patients by 4 examiners.

Table II  ICCs for three variations of the anterior apprehension, relocation, augmentation, and release tests and the chance that the results occurred by chance (P).

<table>
<thead>
<tr>
<th>Test</th>
<th>Correlation (ICC)</th>
<th>Significance (P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apprehension test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>0.31</td>
<td>0.0082</td>
</tr>
<tr>
<td>Apprehension</td>
<td>0.47</td>
<td>0.0002</td>
</tr>
<tr>
<td>Pain and/or apprehension</td>
<td>0.44</td>
<td>0.0004</td>
</tr>
<tr>
<td>Relocation test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>0.31</td>
<td>0.0147</td>
</tr>
<tr>
<td>Apprehension</td>
<td>0.71</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Pain and/or apprehension</td>
<td>0.44</td>
<td>0.0004</td>
</tr>
<tr>
<td>Augmentation test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>0.09</td>
<td>0.23</td>
</tr>
<tr>
<td>Apprehension</td>
<td>0.48</td>
<td>0.0003</td>
</tr>
<tr>
<td>Pain and/or apprehension</td>
<td>0.33</td>
<td>0.0058</td>
</tr>
<tr>
<td>Release test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>0.31</td>
<td>0.0119</td>
</tr>
<tr>
<td>Apprehension</td>
<td>0.63</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Pain and/or apprehension</td>
<td>0.45</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Assessments were made on 13 patients by 4 examiners.

most reliable, in that there was no variance among the examiners’ results (ICC = 1.0). The 0° abducted position was the next most reliable test position (ICC = 0.68). In the inferior direction, all positions had fair to good reliability (ICC ≥ 0.65); however, similar to the load-and-shift test in the posterior direction at 20° abduction, only a few patients had a laxity greater than 0. The accuracy of the ICC score in this situation is questionable, as a lack of variability between test findings makes the statistical analysis of reliability difficult.

Levy et al7 used a k calculation to determine the level of interexaminer agreement. This measure is sensitive to differing values of pretest prevalence. As no data were published with regard to the actual test results, no comparisons can be made between the two studies.

The sulcus sign was found to have fair to good reliability (ICC = 0.60) in this study. In another study, we have found that the sulcus sign, when greater than 2 cm, is predictive for multidirectional instability.15

For the provocative tests of anterior instability, changes in apprehension alone were the most reliable interpretation. Overall, the clinical tests in which an increase in a symptom was measured (apprehension, release, and augmentation) had a lower reliability than those tests in which there was a decrease in symptoms (relocation). The clinical tests in which the increase was slow (apprehension and augmentation, as the increasing pressure on the joint was applied slowly, as opposed to a relocating force suddenly removed) were less reliable than that in which the movement (and hence change in symptoms) was rapid (release test).

When pain alone was interpreted as a positive test, the reliability was consistently the lowest across the four clinical tests. This could be the result of the poor validity of this test interpretation. There is some controversy in the literature as to whether pain should be accepted as a positive test.1,2,6,13 The results of this study suggest that if the findings of the provocative tests are to be compared among examiners, the examiners should focus on apprehension rather than pain. Similarly, Speer et al14 showed that the criterion of apprehension is also more predictive of anterior instability than pain alone.

In summary, the combined results of this and other studies suggest that the load-and-shift test for anterior instability is most valid and reliable when performed with the patient in the supine position and the arm abducted to 90°, whereas posterior instability testing is best performed at 20° of abduction. The sulcus sign is a valid and reliable test for multidirectional instability. The provocative clinical tests for anterior instability (apprehension test, relocation test, augmentation test, and release test) are valid and reliable if apprehension rather than pain is used as the criterion for a positive test.15

We are grateful to the patients who took part in this study.

REFERENCES