

Intra- and inter-rater reliability of anterior and posterior drawer tests for the assessment of people with shoulder instability

Clinical Rehabilitation
2025, Vol. 39(7) 914–922
© The Author(s) 2025



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02692155251339380
journals.sagepub.com/home/cre



Rocio Aldon-Villegas¹ , Gema Chamorro-Moriana¹ ,
Patricio Lopez-Tarrida²  and
Maria-Luisa Benitez-Lugo¹ 

Abstract

Background: The well-known drawer tests to assess glenohumeral laxity and instability have shown appropriate reliability, although analysed mainly in healthy subjects.

Objective: To evaluate the intra- and inter-rater reliability of anterior and posterior drawer tests in subjects with symptoms of shoulder instability.

Design: Clinometric study of intra- and inter-rater reliability of drawer tests was carried out following COSMIN recommendations and GRRAS checklist.

Setting: Centres with equipped facilities for assessments.

Participants: There were 105 participants (69 male/36 female) aged 18 to 60 years with instability symptoms in at least one shoulder. Each participant underwent bilateral assessments. The sample consists of 210 shoulders, unstable and healthy.

Intervention: Anterior and posterior drawer tests.

Main measures: Humeral translations were assessed using drawer tests and graded with Hawkins scale, modified Hawkins and dichotomising (positive/negative). Two sessions were performed (seven to fourteen-day washout period): Each patient was evaluated by two examiners in the first session and by one of them in the second. Weighted Kappa analysed the reliability.

Results: The intra-rater reliability of the anterior and posterior drawer tests was *excellent* (*weighted Kappa* = 1) with the Hawkins scale. Inter-rater reliability was *good* for the anterior drawer: *weighted Kappa* = 0.76 (95%confidence interval: 0.67–0.85) with the Hawkins scale, *weighted Kappa* = 0.78 (95%confidence interval: 0.69–0.87) with modified Hawkins, and *weighted Kappa* = 0.80 (95%confidence interval: 0.71–0.89) dichotomising; and for the posterior drawer: *weighted Kappa* = 0.62 (95%confidence interval: 0.52–0.72), *weighted Kappa* = 0.67 (95%confidence interval: 0.57–0.78), and *weighted Kappa* = 0.70 (95%confidence interval: 0.59–0.80), respectively.

¹Department of Physiotherapy, Research Group “Area of Physiotherapy” CTS-305, University of Seville, Seville, Spain

²Department of Sport and Computer Science, University of Pablo de Olavide, Seville, Spain

Corresponding author:

Gema Chamorro-Moriana, Department of Physiotherapy, Research Group “Area of Physiotherapy” CTS-305, University of Seville, Avicena St., 41009 Seville, Spain.
Email: gchamorro@us.es

Conclusion: Drawer tests demonstrated *excellent* intra-rater and *good* inter-rater reliability in subjects with symptoms of shoulder instability.

Keywords

Shoulder instability, glenohumeral laxity, drawer test, reliability, functional assessment

Received November 13, 2024; accepted April 15, 2025

Introduction

Glenohumeral instability, as well as its assessment, represents a challenge in the clinical and research setting.¹ Such instability is related to higher grades of shoulder laxity.¹ Laxity is considered a risk factor for multidirectional glenohumeral instability.^{2,3} A relationship between generalised ligament laxity and traumatic shoulder instability has even been proven.⁴ Consequently, assessment and/or diagnosis of glenohumeral laxity is highly recommended to prevent or diagnose instability.¹

The diagnosis of shoulder instability is mostly based on clinical history and manual glenohumeral laxity tests,⁵ which sometimes, are combined with imaging scans to assess the integrity of musculoskeletal structure.⁶ These are: x-ray imaging, standard or stress imaging,⁷ ultrasound,⁶ magnetic resonance imaging⁸ and even magnetic resonance arthrography after intra-articular injection of contrast, which has shown great diagnostic accuracy in terms of labral injuries,⁹ frequent in patients with instability. However, these methods enable to discover the state of the structures, but not the functionality. In addition, they are costly,¹⁰ are often not immediately available to clinicians and may carry risks, that is radiation emitted by X-rays.¹¹

In relation to the specific physical assessment tests for glenohumeral laxity, the well-known anterior and posterior drawer tests used in the assessment of humeral translation and described for the first time by Gerber et al.¹² stand out. They are frequently used by clinicians and researchers because they are easy to perform, accessible and practical.¹³ In addition, they offer positive evidence at the levels of sensitivity,^{14–16}

specificity^{14–16} and inter-rater reliability^{17,18} in pathological groups (such as rotator cuff tears,^{15,16} impingement syndrome^{15,16}). However, reliability studies^{19–22} were conducted mainly with healthy participants and with smaller sample sizes than recommended.^{23,24}

Thus, this study aimed to evaluate the intra- and inter-rater reliability of the anterior and the posterior drawer tests in subjects with symptoms of shoulder instability.

Methods

The study design consisted of a clinometric analysis of the intra- and inter-rater reliability of anterior and posterior drawer tests. A flow diagram is shown in Supplementary Material. The research was based on the consensus-based standards for the selection of health measurement instruments (COSMIN).²⁵ It was approved by the Ethics Committee of the Virgen Macarena-Virgen del Rocío Hospitals of the Andalusian Public Health System (No. 1267-N-21) in accordance with the Helsinki Declaration.²⁶

The Guidelines for Reporting Reliability and Agreement Studies (GRRAS)²⁷ checklist was considered.

Participants were chosen by non-random convenience sampling from University, clinical, and sports centres in Seville and Cadiz. Inclusion criteria were: (a) persons with symptoms of instability in at least one shoulder with or without a clinical diagnosis, although both shoulders were always assessed (see next paragraph); (b) aged between 18 and 55 years.²⁸ Exclusion criteria: (a) subjects with musculoskeletal shoulder pathologies not associated with possible instability; (b) cognitive

impairment that affected following the clinician's instructions.

The study considered the two shoulders of each participant as a sample, that is, it included asymptomatic shoulders in order to cover all grades of the laxity scales used (range zero to three, where grades zero and one are associated with no laxity).

The assessment of the glenohumeral laxity, specifically of humeral translation, was carried out by means of the anterior and posterior drawer tests. The patient is placed in the supine position and the physical therapist stands on the side of the shoulder to be evaluated.¹² Figure 1 shows the execution of the anterior drawer test and Figure 2 shows the execution of the posterior drawer test. The displacement of the humerus over the scapula can be easily appreciated and graded.¹²

The degrees of translation were recorded according to: the Hawkins scale^{19,20,29} and the modified Hawkins scale.^{19,20,30} The Hawkins scale ranges between zero and three: grade zero, no or minimal translation; grade one, translation of the humeral head to the glenoid but not over the rim; grade two, humeral head translates over the glenoid rim but does not lock; and grade three, humeral head locks out over the rim.^{29,31} The modified Hawkins^{1,30} equates grade zero with grade one without affecting clinical assessment,³⁰ and improves intra- and inter-rater reproducibility.¹⁹ In addition, the results of both drawer tests were dichotomised based on the reliability study of clinical shoulder tests (e.g. load-and-shift) by Eshoj et al.³²: the grade zero and one of Hawkins scale as negative and grade two and three as positive.

Fieldwork was carried out in equipped facilities in Seville and Cadiz, that is, acclimatised assessment room, with treatment couches, ergonomic cushions, portable dividing screens to safeguard the privacy of the participants, tables, chairs and adjoining dressing rooms. After reading the information sheet and signing the informed consent form, descriptive data on affiliation, age, weight, height, body mass index, clinical diagnosis, symptomatology and surgical intervention were collected.

Subsequently, the anterior and posterior drawer tests were performed on the assessment couches

with cushions to keep the patient in supine decubitus with knees and hips semi-flexed. Participants had to undress from the waist up except for a bra or top, leaving the shoulder girdle visible.

Two assessment sessions were performed, leaving a washout period of 7 to 14 days, so as not to cause substantial changes in the joint.³³ In addition, participants confirmed that their shoulder condition had not changed in the second session.

In the first session, each patient was evaluated by two examiners (RA and MB) – both physiotherapists with clinical experience in shoulders and prior rigorous training in the physical tests employed. The execution and results obtained by one examiner could not be known by the other. The anterior drawer test was performed first and then the posterior drawer test.

In the second session, in order to obtain data for intra-rater reliability, RA examiner repeated his assessments following the same guidelines.

After the fieldwork, all the data collected – descriptive data and from the physical tests – were transferred to an Excel matrix for subsequent analysis.

Statistical analysis

Sample size influences the accuracy of reliability.²⁴ Thus, De Vet et al.²³ recommend including at least 50 patients to complete a two-by-two table. Moreover, considering the sample size (shoulders), for a 95% confidence level, that is a 5% alpha error and a precision of 3.5%, at least 203 shoulders are required.³⁴ This study doubled the suggested minimum sample size in both participants and affected shoulders.

As for the sample descriptives, absolute (*N*) and relative (%) frequencies were considered for qualitative variables. For quantitative variables, normality was assessed using the Kolmogorov–Smirnov test, taking the mean and standard deviation for parametric variables, and the median and interquartile range for nonparametric variables.

The observed proportions and the expected proportions by chance were calculated for the anterior and posterior drawer tests.^{35,36} The Kappa index

was used to assess the level of agreement between the examiners in both tests.³⁷

The observed and the expected proportions by chance, as well as the level of agreement, were calculated for both the weighted and the unweighted forms.^{35,36}

Intra- and inter-rater reliability was obtained using weighted Kappa to take into account the different levels of disagreement between categories.³⁸ In this study, the categories are zero to three in Hawkins and one to three in modified Hawkins. The interpretation of weighted Kappa was: 0, no reliability; 0.01–0.20, poor; 0.21–0.40, fair; 0.41–0.60,

moderate; 0.61–0.80, good; and 0.81–1.00, excellent.³⁹ Statistical analysis was performed using IBM SPSS STATISTICS software version 29.

Results

The sample consisted of 210 shoulders from 105 participants, 69 men and 36 women, with symptoms of instability in at least one shoulder (see inclusion criteria). The descriptions of the participants (age, weight, height, body mass index) and the shoulder sample (clinical diagnosis, symptomatic and surgical treatment) are shown in Table 1.

Table 1. Descriptive characteristics of the participants and the assessed shoulders.

		Men (n = 69)	Women (n = 36)	Total (n = 105)
Age (years)	$\bar{x}$	28.3	30.6	29.1
	95% CI	26.9–29.6	28.1–33.1	27.8–30.3
	SD	8.3	10.6	9.2
	Med	25	28	26
	IQR	22.5–31.5	22–34.8	22–32.5
Weight (kg)	$\bar{x}$	81.6	62.9	75.2
	95% CI	79.8–83.3	60.8–65.0	73.4–76.9
	SD	10.2	8.9	13.2
	Med	80	60.5	75
	IQR	74–85.6	57–67	65–83
Height (cm)	$\bar{x}$	178.8	165.5	174.2
	95% CI	177.8–179.7	163.9–165.2	173.0–175.4
	SD	5.6	6.9	8.9
	Med	179	165	175
	IQR	175–182.5	159.3–169	168–181
BMI	$\bar{x}$	25.5	23.0	24.6
	95% CI	25.0–25.9	22.3–23.7	24.3–25.0
	SD	2.68	3.23	3.09
	Med	25	22.44	24.46
	IQR	23.46–26.57	21.31–24.30	22.62–26.04
Clinical diagnosis^a	N	43	27	70
	%	31.2	37.5	33.3
	95% CI	23.3–39.0	26–49	26.9–39.8
	N	102	52	154
Symptomatic^a	%	73.9	72.2	73.3
	95% CI	66.5–81.3	61.6–82.8	67.3–79.4
Surgical treatment^a	N	17	6	23
	%	12.32	8.33	11
	95% CI	6.8–17.8	2.0–14.7	6.7–15.1

BMI: body mass index; CI: confidence interval; Med: median; IQR: interquartile range.

^aNote: sample comprising shoulders.

Table 2. Results of anterior and posterior drawer tests.

Examiner	Grades ^a	Anterior drawer test		Posterior drawer test	
		Frequency	95% CI	Frequency	95% CI
RA	I	53 (25.2%)	19.4–31.1	157 (74.8%)	68.9–80.6
	II	157 (74.8%)	68.9–80.6	52 (24.8%)	18.9–30.6
	III			1 (0.5%)	0–1.4
MB	0	2 (1%)	0–2.3	7 (3.3%)	0.9–5.8
	I	60 (28.6%)	22.5–34.7	138 (65.7%)	59.3–72.1
	II	146 (69.5%)	63.3–75.8	63 (30%)	23.8–36.2
	III	2 (1%)	0–2.3	2 (1%)	0–2.3

CI: confidence interval.

^aGrades of Hawkins scale.²⁹**Table 3.** Inter-rater reliability analysed thought weighted Kappa.

Scales	Anterior drawer test	Posterior drawer test
Hawkins	0.76 (0.67–0.85)	0.62 (0.52–0.72)
Modified Hawkins	0.78 (0.69–0.87)	0.67 (0.57–0.78)
Dichotomous	0.80 (0.71–0.89)	0.70 (0.59–0.80)

The values in brackets correspond to the 95% confidence interval.

The intra-rater reliability of both the anterior and posterior drawer tests, based on one examiner's ratings (RA), was *excellent* (*weighted Kappa* = 1; which did not enable the calculation of the confidence interval) when using the Hawkins scale. Therefore, the same result was obtained with the modified Hawkins and dichotomising.

Table 2 shows the results of anterior and posterior drawer tests of both examiners.

The inter-rater reliability was *good* for the anterior and posterior drawer tests, being slightly higher in the anterior drawer test. In this case, the weighted Kappa values increased from 0.76 (95% confidence interval: 0.67–0.85) with the Hawkins scale to 0.80 (95% confidence interval: 0.71–0.89) when the scale was dichotomised. For the posterior drawer test, the values increased from 0.62 (95% confidence interval: 0.52–0.72) to 0.70 (95% confidence interval: 0.59–0.80) (Table 3).

Additionally, Table 4 shows the observed proportions, the expected proportions by chance, and the Kappa index values, both weighted and unweighted, for the anterior and posterior drawer tests at the first assessment session. The anterior drawer test achieved *good* results with both the weighted Kappa (*weighted Kappa* = 0.76) and the unweighted Kappa (*Kappa* = 0.75). The posterior drawer test also showed *good* results, although slightly lower values (*weighted Kappa* = 0.62 and *Kappa* = 0.61, respectively).

Discussion

This study analysed the intra- and inter-rater reliability of anterior and posterior drawer tests for assessing glenohumeral laxity in subjects with symptoms of instability of at least one shoulder with or without a clinical diagnosis, whether or not they had undergone surgery. The main findings showed *excellent* intra-rater (*weighted Kappa* = 1) for both drawer tests, even with the Hawkins scale; and *good* inter-rater reliability.

As for intra-rater reliability, Morita et al.²⁰ found a similar result for the anterior drawer with a rater (*weighted Kappa* = 0.861, calculated by us); and lower than our results for the posterior drawer. A rater had *good* (*weighted Kappa* = 0.796) and *excellent* (*weighted Kappa* = 0.867) reliability with Hawkins and its modification²⁰ respectively, while other less experienced ones had *moderate*

Table 4. Observed proportion, expected proportion by chance, and Kappa Index, weighted and unweighted.

	Weighted			Unweighted		
	Observed proportion	Expected proportion by chance	Kappa index	Observed proportion	Expected proportion by chance	Kappa index
Anterior drawer test	0.97	0.86	0.76	0.90	0.59	0.75
Posterior drawer test	0.94	0.85	0.62	0.83	0.57	0.61

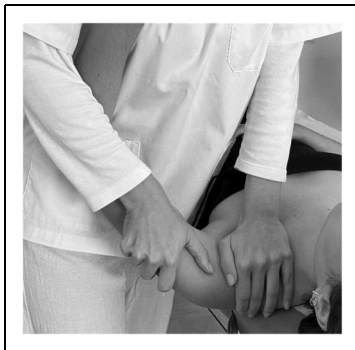


Figure 1. The anterior drawer test.¹² The examiner holds the patient's forearm with the elbow slightly flexed and relaxed. The shoulder is held between 80° and 120° of abduction, 0° and 20° of flexion, and 0° and 30° of external rotation. The examiner fixes the scapula with their medial hand. The lateral hand contacts the head of the humerus applying an anterior force that causes translation of the humerus over the scapula.

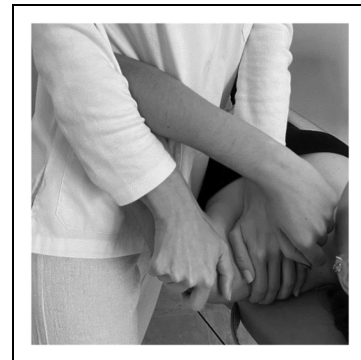


Figure 2. The posterior drawer test.¹² The medial hand stabilises the scapula. The examiner's lateral hand holds the patient's forearm with the elbow flexed to 120°. The shoulder is held between 80° and 120° abduction and 20° flexion. A posterior force is applied to the head of the humerus.

(*weighted Kappa* = 0.587) and *good* (*weighted Kappa* = 0.678). Levy et al.¹⁹ calculated intra-rater reliability with four raters, reporting lower reliability than our study, and that of Morita²⁰ using Kappa instead of weighted Kappa. As with Morita et al.²⁰ the lowest data were from the least experienced rater. The reliability was at most *moderate* (*Kappa* < 0.5) for three raters with the Hawkins scale and its modification. However, the authors noted that lab conditions could have negatively affected.¹⁹ The inter-rater reliability obtained for the drawer tests was *good*, using both the Hawkins scale (anterior drawer: *weighted Kappa* = 0.76; posterior drawer: *weighted Kappa* = 0.62)

and its modification (anterior drawer: *weighted Kappa* = 0.78; posterior drawer: *weighted Kappa* = 0.67). Similar results were reported by Morita et al.²⁰ for the anterior drawer. In contrast, for the posterior drawer with modified Hawkins was only *moderate* according to Morita et al. (*weighted Kappa* = 0.428)²⁰ and Levy et al. (*Kappa* > 0.5).¹⁹ Levy's results had more weight due to the involvement of four raters. Moreover, our study complemented the findings by dichotomising the Hawkins scale based on Eshoj et al.,³² increasing the inter-rater reliability of both drawer tests to almost *excellent* for the anterior drawer (*weighted Kappa* = 0.80).

The evidence of the reliability of the drawer tests has been analysed mainly in healthy subjects.^{19–22} However, this study included unilateral and bilateral instability symptoms, although both shoulders were always included to increase the heterogeneity of the sample, that is unstable, lax (without instability symptoms) and stable shoulders. Given the relationship between shoulder instability and glenohumeral laxity,^{1–3} and that the drawer tests assess the latter,¹² analysing their reliability in this population is crucial.

As for the Hawkins scale, a score of zero could not be expected in our population. However, the assessed asymptomatic shoulders tended to be lax except for traumatic instability, based on the relationship between shoulder instability and generalised laxity.⁴

Regarding sample sizes, unlike other studies,^{17–22} we exceeded the required and doubled De Vet et al.'s²³ recommendation of ≥ 50 participants, ensuring robust reliability evidence.

This study and McFarland et al.³⁵ advocate the modified Hawkins scale for grading humeral translation, as equalising grade zero (rarely obtained in our sample) and grade one (frequent in the absence of instability) does not affect clinical valuation,⁴⁰ but it increases inter-rater reliability¹⁹ by avoiding confusions.³⁰ Nevertheless, we compared the original Hawkins scale with the modified one and found a minimal improvement with the latter, surpassed by Levy et al.,¹⁹ which increased from 47% to 78%.

On the other hand, some reliability studies of drawer tests^{19,20,22} employed the Kappa index to assess agreement,³⁸ whereas this study used the weighted Kappa because of the multiple response options. It reduces the error between the observed and the expected proportions by chance.³⁷ It considers the levels of disagreement between categories and the size of the differences,^{41,42} providing more consistent information for reliability.

Moreover, the *excellent* intra-rater and *good* inter-rater reliability were reinforced by other evidence⁷ on the validity of diagnostic tools for laxity and/or instability. Thus, the stress radiography obtained a significant correlation with the anterior drawer.⁷

Manual tests may be influenced by the experience, skill and sensitivity of raters,¹³ as shown by Levy et al.¹⁹ and Morita et al.²⁰ Many arthrometers were developed to measure glenohumeral laxity, but the discrepancies in the amount of force to be applied and the patient position lead to inconclusive findings.¹³ Thus, manual laxity tests are still relevant.

Clinicians and researchers use the anterior and posterior drawers due to being simple, accessible, useful with a moderate sensitivity and high specificity.^{14–16} The original tests by Gerber et al.¹² have undergone modifications and there is no consensus on their execution. As Gerber et al.,¹² we advocate for supine positions to ensure better relaxation⁴⁰ and reliability¹²; and for scapular stabilisation to avoid compensatory movements that interfere with glenohumeral translation. However, Morita et al.²⁰ do not carry out this stabilisation.

Regarding the study's limitations, intra-rater reliability was assessed by a single rater, and not two as we did for inter-rater, where we obtained more robust evidence. In addition, an even larger sample would have enabled the tests to be applied randomly to further minimise bias.

As to its strengths, the reliability of the drawer tests followed COSMIN and GRRAS checklists; considered a large sample; and the weighted Kappa instead of Kappa. The humeral translation was graded with Hawkins scale, its modification and dichotomised results, enabling comparison between studies. Laboratory conditions were optimal.

A prospective study that analyses and compares the reliability of different physical tests for glenohumeral laxity to obtain the most appropriate, together with the patient's clinical history, would help in the diagnosis of shoulder instability.

Given the *excellent* intra-rater and *good* inter-rater reliability obtained, anterior and posterior drawer tests are recommended for the assessment of glenohumeral instability and/or laxity. We suggest their use by a single clinician to assess the progress of unstable shoulders. In other cases, these tests should be complemented with other objective assessment tools.

Clinical messages

- Anterior and posterior drawer tests assess unstable and/or lax shoulders reliably and could be complemented by diagnostic imaging.
- Drawer tests are appropriate assessment tools for a single clinician to value the progressions of shoulder laxity in physiotherapeutic and/or surgical treatments.
- The modified Hawkins scale is better than the Hawkins to grade humeral translation.

Acknowledgements

The authors would like to acknowledge all the subjects for their participation in this methodological study. The authors would also like to thank the Research Group 'Area of Physiotherapy CTS-305' of the University of Seville for their collaboration.

Authors' contributions

GC and RA conceptualised the idea and designed the study. RA and MB carried out the data collection. RA and PL performed the statistical data analysis. GC and RA wrote the first version of the papers. All authors contributed to the final version. All authors have read and agreed to the published version of the manuscript.

Consent to participate

Informed consent to participate was obtained in written form from all participants.

Consent for publication

Informed consent for publication was obtained in written form from all participants.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Ethical considerations

Ethics Committee of the Virgen Macarena-Virgen del Rocío Hospitals of the Andalusian Public Health System (No. 1267-N-21).

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

ORCID iDs

Rocio Aldon-Villegas  <https://orcid.org/0000-0002-8831-1473>

Gema Chamorro-Moriana  <https://orcid.org/0000-0003-4592-0882>

Maria-Luisa Benitez-Lugo  <https://orcid.org/0000-0003-2058-8799>

Patricio Lopez-Tarrida  <https://orcid.org/0000-0003-1289-2096>

Supplemental material

Supplemental material for this article is available online.

References

1. Jia X, Ji JH, Petersen SA, et al. An analysis of shoulder laxity in patients undergoing shoulder surgery. *J Bone Jt Surg-Am* 2009; 91: 2144–2150.
2. Bigliani LU, Codd TP, Connor PM, et al. Shoulder motion and laxity in the professional baseball player. *Am J Sports Med* 1997; 25: 609–613.
3. Zemek MJ and Magee DJ. Comparison of glenohumeral joint laxity in elite recreational swimmers. *Clin J Sport Med* 1996; 6: 40–47.
4. Caplan J, Julien TP, Michelson J, et al. Multidirectional instability of the shoulder in elite female gymnasts. *Am J Orthop (Belle Mead NJ)* 2007; 36: 660–665.
5. Hegedus EJ, Goode A, Campbell S, et al. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. *Br J Sports Med* 2008; 42: 80–92.
6. Santiago FR, Martínez AM, Muñoz PT, et al. Imaging of shoulder instability. *Quant Imaging Med Surg* 2017; 7: 422–433.
7. Park JY, Kim Y, Oh KS, et al. Stress radiography for clinical evaluation of anterior shoulder instability. *J Shoulder Elbow Surg* 2016; 25: E339–E347.
8. Staker JL, Braman JP and Ludewig PM. Kinematics and biomechanical validity of shoulder joint laxity tests as diagnostic criteria in multidirectional instability. *Braz J Phys Ther* 2021; 25: 883–890.

9. Beltran J, Rosenberg ZS, Chandnani VP, et al. Glenohumeral instability: evaluation with MR arthrography. *Radiographics* 1997; 17: 657–673.
10. Mitchell JM. Utilization trends for advanced imaging procedures: evidence from individuals with private insurance coverage in California. *Med Care* 2008; 46: 460–466.
11. Lin EC. Radiation risk from medical imaging. *Mayo Clin Proc* 2010; 85: 1142–1146.
12. Gerber C and Ganz R. Clinical assessment of instability of the shoulder. With special reference to anterior and posterior drawer tests. *J Bone Joint Surg Br* 1984; 66: 551–556.
13. Gomes E, Andrade R, Valente C, et al. Inconsistency in shoulder arthrometers for measuring glenohumeral joint laxity: a systematic review. *Bioengineering* 2023; 10: 799.
14. McFarland EG. Instability and laxity. In: Kim TK (ed.) *Examination of the shoulder: the complete guide*. New York: Thieme, 2006, pp. 162–212.
15. Van Kampen DA, Van Den Berg T, Van Der Woude HJ, et al. Diagnostic value of patient characteristics, history, and six clinical tests for traumatic anterior shoulder instability. *J Shoulder Elbow Surg* 2013; 22: 1310–1319.
16. Farber AJ, Castillo R, Clough M, et al. Clinical assessment of three common tests for traumatic anterior shoulder instability. *J Bone Joint Surg Am* 2006; 88: 1467–1474.
17. McFarland EG, Torpey BM and Curl LA. Evaluation of shoulder laxity. *Sports Med* 1996; 22: 264–272.
18. Staker JL, Lelwica AE, Ludewig PM, et al. Three-dimensional kinematics of shoulder laxity examination and the relationship to clinical interpretation. *Int Biomech* 2017; 4: 77–85.
19. Levy AS, Lintner S, Kenter K, et al. Intra- and interobserver reproducibility of the shoulder laxity examination. *Am J Sports Med* 1999; 27: 460–463.
20. Morita W and Tasaki A. Intra- and inter-observer reproducibility of shoulder laxity tests: comparison of the drawer, modified drawer and load and shift tests. *J Orthop Sci* 2018; 23: 57–63.
21. McFarland EG, Kim TK, Park HB, et al. The effect of variation in definition on the diagnosis of multidirectional instability of the shoulder. *J Bone Joint Surg Am* 2003; 85: 2138–2144.
22. Ellenbecker TS, Bailie DS, Mattalino AJ, et al. Intrarater and interrater reliability of a manual technique to assess anterior humeral head translation of the glenohumeral joint. *J Shoulder Elbow Surg* 2002; 11: 470–475.
23. De Vet HCW, Terwee CB, Mokkink LB, et al. *Measurement in medicine: a practical guide*. Cambridge, UK: Cambridge University Press, 2011.
24. Apeldoorn AT, Den Arend MC, Schuitemaker R, et al. Interrater agreement and reliability of clinical tests for assessment of patients with shoulder pain in primary care. *Physiother Theory Pract* 2021; 37: 177–196.
25. Mokkink LB, Boers M, Van Der Vleuten CPM, et al. COSMIN risk of bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. *BMC Med Res Methodol* 2020; 20: 293.
26. Manzini JL. Declaración de Helsinki: principios éticos para la investigación médica sobre sujetos humanos. *Acta Bioeth* 2000; 6: 321–334.
27. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS). *J Clin Epidemiol* 2011; 64: 96–106.
28. Dawson J, Fitzpatrick R and Carr A. The assessment of shoulder instability. *J Bone Joint Surg Br* 1999; 81: 420–426.
29. Hawkins RJ, Schutte JP, Janda DH, et al. Translation of the glenohumeral joint with the patient under anesthesia. *J Shoulder Elbow Surg* 1996; 5: 286–292.
30. McFarland EG, Neira CA, Gutierrez MI, et al. Clinical significance of the arthroscopic drive-through sign in shoulder surgery. *Arthrosc J Arthrosc Relat Surg* 2001; 17: 38–43.
31. Tzannes A and Murrell GA. Clinical examination of the unstable shoulder. *Sports Med* 2002; 32: 447–457.
32. Eshoj H, Ingwersen KG, Larsen CM, et al. Intertester reliability of clinical shoulder instability and laxity tests in subjects with and without self-reported shoulder problems. *BMJ Open* 2018; 8. doi:10.1136/bmjopen-2017-018472
33. Van der Linde JA, van Kampen DA, van Beers LWAH, et al. The Oxford shoulder instability score; validation in Dutch and first-time assessment of its smallest detectable change. *J Orthop Surg* 2015; 10: 1–8.
34. Lachenbruch PA, Lwanga SK and Lemeshow S. Sample size determination in health studies: a practical manual. *J Am Stat Assoc* 1991; 86: 1149.
35. Cortés-Reyes E, Rubio-Romero JA and Gaitán-Duarte H. Métodos estadísticos de evaluación de la concordancia y la reproducibilidad de pruebas diagnósticas. *Rev Colomb Obstet Ginecol* 2010; 61: 247–255.
36. Gordillo JTT and Rodríguez VHP. Cálculo de la fiabilidad y concordancia entre codificadores de un sistema de categorías para el estudio del foro online en e-learning. *Rev Investig Educ* 2009; 27: 89–103.
37. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica* 2012; 22: 276–282.
38. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968; 70: 213–220.
39. Altman DG. *Practical statistics for medical research*. 1st ed. Chapman and Hall/CRC, 1991.
40. McFarland EG, Campbell G and McDowell J. Posterior shoulder laxity in asymptomatic athletes. *Am J Sports Med* 1996; 24: 468–471.
41. Tang W, Hu J, Zhang H, et al. Kappa coefficient: a popular measure of rater agreement. *Shanghai Arch Psychiatry* 2015; 27: 62–67.
42. O'Leary S, Lund M, Ytre-Hauge TJ, et al. Pitfalls in the use of kappa when interpreting agreement between multiple raters in reliability studies. *Physiotherapy* 2014; 100: 27–35.