Physiotherapy 96 (2010) 179–190

Systematic Review

Reliability of physical examination tests used in the assessment of patients with shoulder problems: a systematic review

Stephen May*, Ken Chance-Larsen, Chris Littlewood, Dave Lomas, Mahmoud Saad

Faculty of Health and Wellbeing, Sheffield Hallam University, Sheffield S10 2BP, UK

Abstract

Background Shoulder pain is a common clinical problem, and numerous tests are used to diagnose structural pathology.
Objectives To systematically review the reliability of physical examination procedures used in the clinical examination of patients with shoulder pain.
Data sources MEDLINE, PEDro, AMED, PsychInfo, Cochrane Library (2009) and CINAHL were searched from the earliest record on the database to June 2009.
Study eligibility criteria Reliability studies that included any patients with shoulder pain were analysed for their quality and reliability results.
Study appraisal and synthesis methods Pre-established criteria were used to judge the quality of the studies (high quality >60% methods score) and satisfactory levels of reliability (kappa or intraclass correlation coefficient ≥0.85, sensitivity analysis 0.70). A qualitative synthesis was performed based on levels of evidence.
Results Thirty-six studies were included with a mean methods score of 57%. Seventeen studies were deemed to be of high quality; high-quality studies were less likely to meet the pre-agreed level of reliability. The majority of studies indicated poor reliability for all procedures investigated.
Limitations Overall, the evidence regarding reliability was contradictory.
Conclusions and implications There is no consistent evidence that any examination procedure used in shoulder assessments has acceptable levels of reliability. Alternate methods of classification which are reliable should be used to classify patients with shoulder problems.

Keywords: Shoulder; Physical examination; Tests; Reliability; Systematic review

Introduction

Shoulder pain is a common musculoskeletal problem. A review found that point prevalence varied between 7% and 27% of the adult population, 1-year prevalence varied between 5% and 47%, and lifetime prevalence varied between 7% and 67% [1]. It is reasonable to conclude that the burden of shoulder pain in the general population is substantial, and it probably represents the most common musculoskeletal problem after back and neck pain [2]. Shoulder pain is commonly encountered in clinical practice, with an incidence of approximately 10 per 1000 in primary care [3,4], and a prevalence of 12% in a primary-care-based physiotherapy department [5]. Mean-related costs in the first 6 months of an episode of shoulder pain in primary care have been estimated at €690 [6].

Diagnostic labels are commonly applied to patients with shoulder pain, such as capsulitis, bursitis and subacromial impingement syndrome [7], but the pathophysiology of shoulder disorders is still unclear [8–12]. A large number of tests are used for examination of the shoulder that are purported to be indicative of these diagnostic labels, but there are a number of problems underlying these tests. The reliability of tests used to establish a diagnosis has generally been shown to be limited [13–15], the diagnostic validity of some

* Corresponding author.
E-mail address: s.may@shu.ac.uk (S. May).

0031-9406/$ – see front matter © 2010 Chartered Society of Physiotherapy. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.physio.2009.12.002

of the tests has been shown to be moderate at best [16,17], and the anatomical basis for most tests has not been validated [18].

Especially relating to validity, a number of recent systematic reviews have seriously challenged the ability of physical examination tests to make a diagnosis of any shoulder pathology [17], rotator cuff pathology [19,20] or labral lesions [21–23]. This poor diagnostic accuracy may be explained by the lack of anatomical validity of most shoulder tests [18].

Whilst test validity is a vital component of any diagnostic test, it is also important that a test demonstrates reliability. Links between validity and reliability have been little explored in the literature, but it has been suggested that reliable assessment–diagnosis/classification–treatment links are an essential building block to improving outcomes [24]. Reliability places an upper limit on validity; therefore, when reliability is high, the maximum possible validity is higher [25]. Reliability refers to the ability of multiple clinicians to come to the same conclusion from the performance of a test, or of a single clinician to do so when performing the test on two separate occasions [25]. The formal mathematical definition of reliability is subject variability divided by subject variability plus measurement error [25].

If management decisions are to be based on the test result, this measure of reproducibility is an essential component of a test and the diagnostic-treatment process. As far as the authors are aware, no previous systematic review has considered the reliability of physical examination procedures used in examination of the shoulder. This systematic review explored the reliability of physical examination tests used in examination of the shoulder.

Methods

A review protocol was developed and approved by all authors, but was not accessible on a website.

Search

Searches of MEDLINE (January 1966 to June 2009), PEDro (June 2009), AMED (1985 to June 2009), PsychInfo (1960 to June 2009), the Cochrane Library (2009) and CINAHL (1982 to June 2009) were conducted using the search terms in Box 1 , grouped into three subject areas: shoulder problems, reliability and physical examination. Initially, terms were searched individually, and then terms from separate subject areas (shoulder problems, reliability and physical examination) were searched using Boolean logic (OR). Finally, these three search areas were combined using Boolean logic (AND). This was supplemented by hand searching the reference lists of the articles found from the electronic search. Included articles had to meet the following criteria:

Box 1: Search terms

Shoulder problems: shoulder, glenohumeral, acromio$, impingement syndrome, rotator cuff, frozen shoulder, capsulitis, labral tear, glenoid labrum, bursitis and shoulder, tendon$ and shoulder, instability and shoulder.

Reliability: reliability, reproducibility, inter examiner, inter-examiner, inter tester, inter-tester, inter observer, inter-observer, intra tester, intra-tester, kappa, intra class correlation, intra-class correlation, ICC.

Physical examination: assessment, physical examination, physical tests, clinical examination, manual examination.

- Results published as full reports before June 2009 – abstracts were not included.
- The study involved physical examination procedures used in shoulder examination.
- The study involved human subjects with any sort of shoulder pain.
- The study had to be an intra- and/or inter-examiner reliability design.
- The study had to be available in English.
- The study did not involve a mechanical device, but simple tape measures were accepted as commonly available.
- The study did not involve subjects with non-musculoskeletal conditions.
- The study did not involve asymptomatic volunteers alone, although studies including a mix of symptomatic and asymptomatic participants were included.

Initially, abstracts were screened by one reviewer (SM) who discarded any that were clearly not relevant; the remaining abstracts were reviewed by two reviewers who independently decided which studies went forward. Studies went through to the next stage if they were clearly reliability studies involving the shoulder or if there was not enough detail in the abstract to determine if this was the case. Full studies were then obtained and a meeting was convened to decide which studies should be included. At least two reviewers judged each paper, and decisions were reached by majority, with a third reviewer, if there were any disagreements. There were no disagreements at this stage. Discussion and clarification about the criteria checklist and items for data extraction was also made at this meeting. A pilot study using two pairs of reviewers to review two papers was used to identify any problems with the quality criteria and to investigate reliability between two reviewers. The pairs of reviewers recorded kappa values of 0.79 and 0.86, which was deemed acceptable. Quality assessment and data extraction of results was performed independently by two reviewers for each paper; a third reviewer resolved any disagreements by majority decision. There were disagreements on six papers that required a majority decision.

Criteria checklist

There are no established or commonly used criteria for judging the quality of reliability studies. A criteria checklist consisting of three categories – study population, test procedure and test results – had been devised previously [26]. Modifications to this set of criteria, including consideration of examiners and data analysis, have been used in recent reviews [27,28], and other reviewers have used similar lists [29] although different lists of criteria have been developed [30]. For judging study quality, the weighted criteria used in a previous systematic review was used [27]. One amendment was made; it was decided relative to Point 8 that as long as some attempt was made to describe the profession and years of practice of the clinicians, this was sufficient to meet this criterion. It was considered that, for instance, an attempt to compare student and experienced clinicians should not receive a lower score, provided that the status of the examiners was clear. The criteria are provided as a footnote to Table 2. The maximum score was 100 points; a trial was considered to be of higher quality if it scored over 60%, as reported in a previous systematic review, and these studies are shown in bold in Table 2 [27].

Data analysis

Kappa is used as the reliability coefficient with nominal data, weighted kappa for ordinal data and intra-class correlation coefficient (ICC) or the Bland–Altman test for continuous data [25]. Percentage agreement is generally considered to be inappropriate as this can be heavily affected by chance agreement. Both kappa and ICC are presented as numerical values between 0.00 and 1.00. Kappa was interpreted as follows: 0.00 to 0.20, poor or slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial or good agreement; and 0.81 to 1.00, very good or almost perfect agreement [31]. Although this interpretation is commonly used, several authorities have suggested that minimal values for a useful instrument should be 0.75 or even 0.85 [25,32].

Similarly with ICC, the closer to 1.00, the greater the reliability. This has been interpreted as follows: <0.40, poor reliability; 0.40 to 0.75, fair to good reliability; and >0.90, excellent reliability [33]. As with kappa, higher values have been recommended, especially when individuals are being considered rather than groups; and coefficients of 0.85 or 0.90 are appropriate [32–35]. It was decided therefore that the predetermined criteria for satisfactory reliability would be 0.85 for both kappa and ICC. Given that such cut-off points are necessarily somewhat arbitrary, it was determined that a sensitivity analysis would be conducted by lowering the cut-off point to 0.70.

A meta-analysis was not attempted due to the heterogeneity of tests, patients and analyses, and as direct comparison of reliability studies was deemed inappropriate [25]. A qualitative, levels of evidence approach, adapted from van Tulder *et al.* [36], was used to synthesise data as shown in Table 1.

Table 1
Levels of evidence.

Level of evidence	
Strong	Consistent findings from three or more high-quality studies
Moderate	Consistent findings from at least one high-quality study and a number of low-quality studies
Limited	Consistent findings in one or more low-quality studies
Conflicting	Inconsistent findings irrespective of study quality
No evidence	No studies found

Results

A large number of studies were initially identified through the search strategy, many of which were not deemed to be relevant after review of abstracts or full articles. Finally, 36 studies [37–72] that met the inclusion criteria and had been identified by the search strategy were included. Fig. 1 provides a flow diagram of the selection process.

Of the 36 studies included, 24 involved physical therapists alone, including students in one study, six included medical professionals alone, and six included medical and other health professionals. Thirty studies investigated inter-tester reliability, 16 investigated intra-tester reliability and nine studies investigated both. A large number of physical examination procedures were investigated. The mean quality score was 57%, and 17 studies were deemed to be of high quality (score >60%) (Table 2). There was a tendency for lower-quality scores to be associated with reliability coefficient statistics over 0.85. In 78 investigations of physical examination procedures with low-quality scores, 15 were reliable (19%), whereas in 61 high-quality investigations, only seven were reliable (12%).

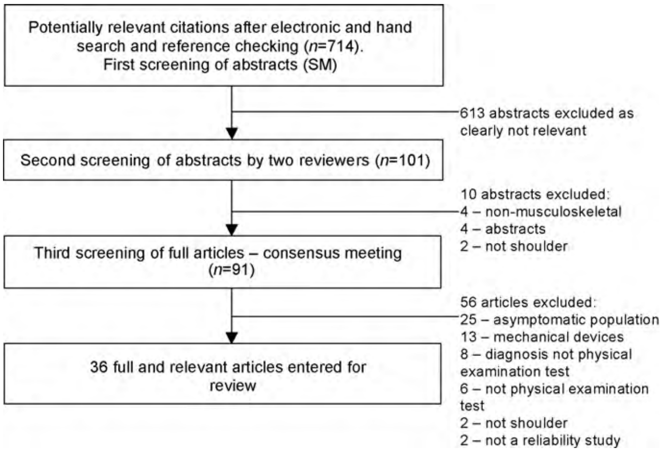
Potentially relevant citations after electronic and hand search and reference checking (n=714). First screening of abstracts (SM)
613 abstracts excluded as clearly not relevant
Second screening of abstracts by two reviewers (n=101)
10 abstracts excluded:
4 – non-musculoskeletal
4 – abstracts
2 – not shoulder
Third screening of full articles – consensus meeting (n=91)
56 articles excluded:
25 – asymptomatic population
13 – mechanical devices
8 – diagnosis not physical examination test
6 – not physical examination test
2 – not shoulder
2 – not a reliability study
36 full and relevant articles entered for review

Fig. 1. Flow diagram of selection process of studies.

Table 2
Quality scores.

Study	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total*
Al-Shenqiti 2005 [37]	4	0	0	6	5	5	0	10	5	0	0	5	10	0	50
Bertlison 2003 [38]	4	4	7	10	5	5	10	10	5	0	0	5	10	10	**85**
Borstad 2007 [39]	4	0	0	6	0	0	0	10	0	0	0	5	0	10	35
Bron 2007 [40]	4	4	7	3	5	5	10	10	5	10	5	5	10	0	**83**
Chesworth 1998 [41]	4	4	0	10	5	5	10	10	0	0	0	5	0	10	**63**
Croft 1994 [42]	0	0	0	0	5	5	0	0	5	10	5	0	0	0	30
Hayes 2001 [43]	4	4	0	3	5	5	10	10	5	0	0	5	10	10	**71**
Hayes 2001 [44]	4	4	0	3	5	0	10	10	5	0	0	5	10	10	**66**
Hayes 2001 [45]	4	0	7	0	5	5	10	0	5	10	5	5	0	10	**66**
Hayes 2002 [46]	4	0	0	0	5	5	0	0	5	10	5	0	0	10	44
Hickey 2007 [47]	4	4	0	0	5	5	10	10	0	10	5	5	10	10	**78**
Johansson 2008 [48]	4	4	7	3	5	5	0	10	5	0	0	0	10	0	53
Kibler 2002 [49]	4	4	0	3	5	5	0	0	5	10	0	5	0	0	41
Kim 1999 [50]	4	0	7	6	5	0	0	0	0	0	0	5	10	0	37
Kim 2007 [51]	4	4	0	6	5	0	10	0	0	0	0	5	10	0	44
Lewis 2007 [52]	4	4	0	10	5	5	10	0	0	0	0	5	0	10	53
Lewis 2008 [53]	4	4	0	10	5	5	10	10	0	0	0	5	0	10	**63**
Lo 2004 [54]	4	4	0	3	5	5	10	0	5	10	5	5	0	10	**66**
McClure 2009 [55]	4	0	0	10	5	5	0	10	0	10	5	5	0	10	**64**
Nanda 2008 [56]	4	4	0	6	5	5	10	0	0	0	0	5	10	0	49
Nijs 2005 [57]	4	4	0	3	5	5	10	0	5	0	0	0	0	0	36
Nomden 2008 [58]	4	4	7	10	5	5	10	10	5	0	0	0	0	0	60
Odom 2001 [59]	4	4	0	3	5	5	0	10	5	10	0	0	10	10	**66**
Ostor 2004 [60]	4	4	7	10	5	5	0	0	5	10	0	5	10	0	**65**
Palmer 2000 [61]	4	4	7	10	5	5	0	0	5	0	0	5	10	10	**65**
Rabin 2006 [62]	4	4	0	3	5	5	10	10	5	10	0	5	10	0	**71**
Razmjou 2004 [63]	4	4	7	10	5	5	10	10	5	0	0	5	10	10	**85**
Terwee 2005 [64]	4	4	7	10	5	5	0	10	5	0	0	5	10	10	**75**
Tzannes 2004 [65]	4	0	0	0	5	5	10	0	0	10	5	5	0	0	44
Valentine 2006 [66]	4	4	0	3	5	5	10	10	0	0	0	5	0	10	56
van Duijn 2001 [67]	4	0	0	0	5	5	10	10	5	10	5	0	10	0	**64**
Wadsworth 1987 [68]	0	0	0	0	5	0	10	10	0	0	0	0	0	0	25
Walker-Bone 2002 [69]	4	0	0	10	5	0	10	0	0	0	0	0	10	10	49
Walsworth 2008 [70]	4	0	7	6	5	5	10	0	5	0	0	5	0	10	57
Westerberg 1996 [71]	4	0	0	0	0	0	0	0	0	10	0	5	10	0	29
Yang 2006 [72]	4	4	0	10	5	0	10	10	0	0	0	5	0	10	58
Maximum	**4**	**4**	**7**	**10**	**5**	**5**	**10**	**10**	**5**	**10**	**5**	**5**	**10**	**10**	

1, adequate description of study population (0/4); 2, representative of clinical practice (0/4); 3, subjects selected randomly or consecutively (0/7); 4, number of subjects (<25 = 0, >25 = 3, >50 = 6; >75 or sample size calculation); 5, procedure clearly described and reproducible (0/5); 6, procedure executed in uniform manner (0/5); 7, adequate measures to reduce bias (0/10); 8, adequate description of examiners (0/10); 9, consensus procedure prior to testing or pilot study (0/5); 10, more than one pair of examiners tested (0/10); 11, multiple testing between examiners; 12, standardised measure of test outcome (0/5); 13, frequencies of outcome and agreement reported (0/10); 14, appropriate inferential statistics and measure of variance (0/10).
Bold results indicate >60%/high quality.

Most commonly used physical examination procedures for shoulder examination were investigated for inter-tester reliability, mainly in multiple studies (Table 3). In Table 3, the results are grouped together as they relate to certain anatomical or pathological entities or to certain physical examination procedures. For all procedures, there were inconsistent findings from high- and low-quality studies. There were 50 investigations for tests for impingement, rotator cuff and resisted tests; six had values >0.85 and a further four had a range of values that included more than 0.70 to 0.85. There were nine investigations for scapular positions and scapular movement dysfunction; two had values >0.85 and a further three had a range of values that included more than 0.70 to 0.85. End feel, accessory movements, tests for the acromioclavicular joint, tenderness and trigger points had 23 investigations; five had a range of values that included 0.70 to 0.85. There were 27 investigations of shoulder instability and glenoid labral tests; none had values >0.85, but seven had a range of values between 0.70 and 0.85. Movement diagrams had a range that included values >0.85 in three investigations, but this was from a single study (high quality). There were 27 investigations into range of movement and pain response on movement; 10 had values >0.85 and another seven had a range of values from 0.70 to 0.85. This was the nearest any procedure came to suggesting reasonable inter-tester reliability, but according to the pre-established criteria, the conclusion for range of movement was inconsistent findings.

Intra-tester results tended to be slightly better, but overall conclusions were still conflicting (Table 4). As in Table 3, the results in Table 4 are grouped together as they relate to cer-

Table 3
Inter-tester reliability studies.

Item	Reference	Statistic[a]	Variance[b]
Impingement/rotator cuff			
Painful arc	Nanda 2008 [56]	0.48	
	Palmer 2000 [61]	**0.93**	
	Walker-Bone 2002 [69]	0.47	
Active	Nomden 2008 [58]	0.46	
Start/end arc	Nomden 2008 [58]	ICC 0.72/0.57	
	Ostor 2004 [60]	0.49/0.62	
Passive	Nomden 2008 [58]	0.52	
Start/end arc	Nomden 2008 [58]	ICC 0.54/0.72	
Hawkins-Kennedy	Nanda 2008 [56]	0.55	
	Razmjou 2004 [63]	0.29	
	Ostor 2004 [60]	0.18 to 0.43	
	Johansson 2008 [48]	**0.91**	
Neer's sign	Nanda 2008 [56]	0.10	
	Nomden 2008 [58]	0.62	
	Johansson 2008 [48]	**1.00**	
	Razmjou 2004 [63]	0.51	
Patte manoeuvre	Johansson 2008 [48]	**1.00**	
Drop arm	Nanda 2008 [56]	0.35	
	Ostor 2004 [60]	0.28 to 0.66	
Supraspinatus			
Resisted abduction	Nanda 2008 [56]	0.30	
Jobe's test	Nanda 2008 [56]	0.44	
	Johansson 2008 [48]	**0.94**	
Empty can	Ostor 2004 [60]	0.44 to 0.49	
Infraspinatus			
Resisted external rotation	Nanda 2008 [56]	0.44	
	Ostor 2004 [60]	0.18 to 0.45	
Subscapularis			
Gerber's lift-off test	Nanda 2008 [56]	0.48	
	Ostor 2004 [60]	0.28 to 0.30	
Biceps			
Yergason's test	Nanda 2008 [56]	0.28	
	Ostor 2004 [60]	0.28	
Speed's test	Nanda 2008 [56]	0.44	
	Ostor 2004 [60]	0.17 to 0.32	
Resisted tests			
Multiple tests	Bertilson 2003 [38]	0.27/0.69[c]	
Elbow flexion	Palmer 2000 [61]	0.83	
	Walker-Bone 2002 [69]	0.66	
	Hayes 2001 [44]	0.25	−0.30 to 0.81
Elbow extension	Hayes 2001 [44]	0.44	−0.13 to **1.00**
External rotation	Palmer 2000 [61]	**0.90**	
	Walker-Bone 2002 [69]	0.44	
	Hayes 2001 [44]	0.37	0.01 to 0.74
Internal rotation	Palmer 2000 [61]	0.54	
	Walker-Bone 2002 [69]	−0.02	
	Hayes 2001 [44]	0.00	−0.58 to 0.58
Abduction	Palmer 2000 [61]	0.81	
	Walker-Bone 2002 [69]	0.29	
	Hayes 2001 [44]	0.45	0.09 to 0.81
Adduction	Hayes 2001 [44]	0.36	−0.30 to **1.00**
Oxford scale			
Elevation	Hayes 2002 [46]	ICC 0.72	0.38 to **0.93**
External rotation	Hayes 2002 [46]	ICC 0.55	0.17 to **0.88**
Internal rotation	Hayes 2002 [46]	ICC 0.61	0.26 to **0.89**
Gerber lift-off	Hayes 2002 [46]	ICC 0.38	0.02 to 0.81

Table 3 (*Continued*)

Item	Reference	Statistic[a]	Variance[b]
Scapular positioning			
Acromion – table distance	Nijs 2005 [57]	**0.88 to 0.94**	
Scapular – spine distance	Nijs 2005 [57]	0.50 to 0.80	
Scapular slide test	Nijs 2005 [57]	0.70 to **0.96**	
	Rabin 2006 [62]	0.53/0.62[d]	
With symptoms	Odom 2001 [59]	ICC 0.45 to 0.79	−0.38 to **0.91**
Without symptoms	Odom 2001 [59]	ICC 0.43 to 0.74	−0.29 to **0.88**
Scapular movement dysfunction			
Scapular dyskinesis system	Kibler 2002 [49]	0.31 to 0.42	
Scapular dyskinesis – live	McClure 2009 [55]	0.55 to 0.58	0.32 to 0.79
Scapular dyskinesis – videotape	McClure 2009 [55]	0.48 to 0.61	0.29 to 0.79
Range of movement and pain response			
Hand in neck (ROM/pain)	Nomden 2008 [58]	0.52/0.52	
	Westerberg 1996 [71]	**0.86**	
	Yang 2006 [72]	**0.90**	0.69 to **0.96**
Hand in back (ROM/pain)	Nomden 2008 [58]	0.73/0.35	
To scapula	Yang 2006 [72]	**0.90**	0.69 to **0.94**
To opposite scapula	Yang 2006 [72]	0.83	0.75 to **0.90**
Hand behind back	Hayes 2001 [45]	ICC 0.26 to 0.39	−0.01 to 0.77
Pour out of a pot	Westerberg 1996 [71]	**0.89**	
Passive flexion	Hayes 2001 [45]	ICC 0.70	0.42 to **0.92**
Stand and reach	Hayes 2001 [45]	ICC 0.74	0.45 to **0.94**
Active abduction (ROM/pain)	Nomden 2008 [58]	ICC **0.96**/0.65	
(ROM/pain)	Bertilson 2003 [38]	0.62/0.77[c]	
(ROM)	Terwee 2005 [64]	ICC 0.76 to **0.88**	0.67 to **0.91**
(ROM)	Croft 1994 [42]	ICC **0.99**	
Passive abduction (ROM/pain)	Nomden 2008 [58]	ICC **0.96**/0.69	
(ROM)	Terwee 2005 [64]	ICC 0.73 to **0.87**	0.66 to **0.90**
(ROM/pain)	Croft 1994 [42]	ICC **0.95**/0.84	
(ROM)	Hayes 2001 [45]	ICC 0.66	0.37 to **0.90**
Passive horizontal abduction (ROM)	Terwee 2005 [64]	ICC 0.15 to 0.70	0.02 to 0.81
External rotation (ROM/pain)	Nomden 2008 [58]	ICC 0.70/0.50	
(ROM)	Croft 1994 [42]	ICC 0.37	
Passive external rotation (ROM)	Terwee 2005 [64]	ICC 0.34 to 0.77	0.00 to **0.88**
(ROM)	Croft 1994 [42]	ICC 0.43	
(ROM)	Hayes 2001 [45]	ICC 0.57	0.26 to **0.87**
Internal rotation (ROM)	Borstad 2007 [39]		
Passive horizontal adduction (ROM)	Terwee 2005 [64]	ICC 0.1 to 0.49	0.04 to 0.60
Aberration in shoulder movement patterns	Hickey 2007 [47]	0.23	0.19 to 0.27
End feel			
Type of end feel	Chesworth 1998 [41]	0.62 to 0.76	
Full abduction	Hayes 2001 [43]	0.26	−0.16 to 0.68
Glenohumeral abduction	Hayes 2001 [43]	0.70	0.31 to **1.00**
Horizontal adduction	Hayes 2001 [43]	0.40	0.01 to 0.79
External rotation	Hayes 2001 [43]	0.47	0.08 to **0.86**
Internal rotation	Hayes 2001 [43]	0.41	0.03 to 0.80
Pain/resistance sequence			
Full abduction	Hayes 2001 [43]	0.51	0.37 to 0.65
Glenohumeral abduction	Hayes 2001 [43]	0.13	
Horizontal adduction	Hayes 2001 [43]	0.62	0.54 to 0.70
External rotation	Hayes 2001 [43]	0.51	
Internal rotation	Hayes 2001 [43]	0.51	0.37 to 0.65
Accessory movements			
First rib (ROM/stiff/pain)	Nomden 2008 [58]	0.26/0.09/0.66	
Inferior glide GHJ	van Duijn 2001 [67]	ICC 0.52	
Movement diagrams			
Onset pain	Chesworth 1998 [41]	0.59 to **0.91**	0.35 to **0.95**
Onset resistance	Chesworth 1998 [41]	0.34 to **0.91**	0.05 to **0.95**
Lateral rotation ROM	Chesworth 1998 [41]	0.83 to **0.90**	0.70 to **0.95**

Table 3 (*Continued*)

Item	Reference	Statistic[a]	Variance[b]
Acromio-clavicular joint			
Stress test (pain)	Nomden 2008 [58]	0.51	
	Palmer 2000 [61]	0.80	
	Walker-Bone 2002 [69]	0.30	
Horizontal adduction	Ostor 2004 [60]	0.03 to 0.38	
Shoulder instability			
Load and shift			
Anterior			
Seated	Tzannes 2004 [65]	ICC 0.53	
20° abduction	Tzannes 2004 [65]	ICC 0.60	
90° abduction	Tzannes 2004 [65]	ICC 0.72	
Posterior			
Seated	Tzannes 2004 [65]	ICC 0.68	
20° abduction	Tzannes 2004 [65]	No variance	
90° abduction	Tzannes 2004 [65]	ICC 0.42	
Inferior			
Seated	Tzannes 2004 [65]	ICC 0.79	
20° abduction	Tzannes 2004 [65]	ICC 0.79	
90° abduction	Tzannes 2004 [65]	ICC 0.65	
Sulcus sign	Tzannes 2004 [65]	ICC 0.60	
Apprehension test			
Pain	Tzannes 2004 [65]	ICC 0.31	
Apprehension	Tzannes 2004 [65]	ICC 0.47	
Either	Tzannes 2004 [65]	ICC 0.44	
Relocation test	Tzannes 2004 [65]		
Pain	Tzannes 2004 [65]	ICC 0.31	
Apprehension	Tzannes 2004 [65]	ICC 0.71	
Either	Tzannes 2004 [65]	ICC 0.44	
Augmentation test	Tzannes 2004 [65]		
Pain	Tzannes 2004 [65]	ICC 0.09	
Apprehension	Tzannes 2004 [65]	ICC 0.48	
Either	Tzannes 2004 [65]	ICC 0.33	
Release test	Tzannes 2004 [65]		
Pain	Tzannes 2004 [65]	ICC 0.31	
Apprehension	Tzannes 2004 [65]	ICC 0.63	
Either	Tzannes 2004 [65]	ICC 0.45	
Apprehension, relocation and surprise tests	Lo 2004 [54]	ICC 0.83	0.76 to **0.92** (range)
Glenoid labral tears			
Active compression test	Walsworth 2008 [70]	0.24	−0.02 to 0.50
Anterior slide test	Walsworth 2008 [70]	0.21	−0.05 to 0.46
Crank test	Walsworth 2008 [70]	0.20	−0.05 to 0.46
Biceps load test	Kim 1999 [50]	0.85	
Passive compression test	Kim 2007 [51]	0.77	
Tenderness			
Scapula	Bertilson 2003 [38]	0.33/0.17[c]	
Shoulder	Bertilson 2003 [38]	0.38/0.52[c]	
Myofascial trigger points			
Palpation nodule in taut band	Bron 2007 [40]	0.11 to 0.75	
Palpation referred pain	Bron 2007 [40]	−0.13 to 0.64	
Palpation local twitch response	Bron 2007 [40]	−0.05 to 0.45	
Palpation jump sign	Bron 2007 [40]	0.02 to 0.68	

ROM, range of movement; ICC, intraclass correlation coefficient; GHJ, glenohumeral joint.

[a] Kappa or weighted kappa except where stated.

[b] 95% confidence intervals.

[c] With and without knowledge of history.

[d] In different planes.

Bold results indicate >0.85 kappa/ICC.

Table 4
Intra-tester reliability studies.

Item	Reference	Statistic[a]	Variance[b]
Impingement tests			
Neer's sign	Johansson 2008 [48]	1.00	
Hawkins-Kennedy	Johansson 2008 [48]	1.00	
Patte	Johansson 2008 [48]	1.00	
Jobe's test	Johansson 2008 [48]	1.00	
Scapular positioning			
Scapular slide test			
With symptoms	Odom 2001 [59]	ICC 0.57 to **0.86**	
Without symptoms	Odom 2001 [59]	ICC 0.75 to 0.80	
Angular measurements	Lewis 2008 [53]	ICC 0.84 to **0.98**	0.72 to **0.99**
Linear measurements	Lewis 2008 [53]	ICC 0.61 to **0.98**	0.38 to **0.99**
Scapular movement dysfunction			
Scapular dyskinesis system	Kibler 2002 [49]	0.49 to 0.59	
Accessory movements			
Inferior glide GHJ	van Duijn 2001 [67]	ICC 0.53 to **0.88**	
Movement diagram			
Onset pain	Chesworth 1998 [41]	0.58 to **0.87**	0.35 to **0.93**
Onset resistance	Chesworth 1998 [41]	0.69 to **0.89**	0.48 to **0.94**
Lateral rotation ROM	Chesworth 1998 [41]	0.74 to 0.86	0.57 to **0.92**
Range of movement and pain response			
Passive flexion	Hayes 2001 [45]	ICC 0.59	0.28 to 0.85
Stand and reach	Hayes 2001 [45]	ICC 0.49	0.18 to 0.80
Passive abduction	Hayes 2001 [45]	ICC 0.60	0.30 to **0.86**
Hand in neck (ROM)	Yang 2006 [72]	0.80	0.63 t0 **0.93**
Hand to scapula	Yang 2006 [72]	**0.90**	0.72 to **0.92**
Hand to opposite scapula	Yang 2006 [72]	**0.86**	0.65 to **0.90**
Hand behind back	Hayes 2001 [45]	ICC 0.14 to 0.39	−0.11 to 0.75
External rotation (ROM)	Valentine 2006 [66]	0.85 to **0.93**	0.74 to **0.96**
Passive external rotation	Hayes 2001 [45]	ICC 0.67	0.38 to **0.89**
Internal rotation (ROM)	Borstad 2007 [39]		
With symptoms		ICC 0.67	0.45 to 0.82
Without symptoms		ICC 0.79	0.55 to **0.91**
(ROM)	Valentine 2006 [66]	ICC 0.83 to **0.98**	0.71 to **0.99**
Side-lying adduction (ROM)	Borstad 2007 [39]		
With symptoms		ICC 0.40	0.09 to 0.64
Without symptoms		ICC 0.63	0.29 to 0.83
Supine adduction (ROM)	Borstad 2007 [39]		
With symptoms		ICC 0.79	0.63 to **0.89**
Without symptoms		ICC 0.74	0.47 to **0.88**
Pectoralis minor			
Length test	Lewis 2007 [53]	**0.90 to 0.93**	0.81 to **0.96**
End feel			
Type of end feel	Chesworth 1998 [41]	0.48 to 0.59	
Full abduction	Hayes 2001 [43]	0.84	0.64 to **1.00**
Glenohumeral abduction	Hayes 2001 [43]	0.65	0.32 to **0.97**
Horizontal adduction	Hayes 2001 [43]	**0.92**	0.77 to **1.00**
External rotation	Hayes 2001 [43]	**0.87**	0.69 to **1.00**
Internal rotation	Hayes 2001 [43]	**0.86**	0.67 to **1.00**
Pain/resistance sequence			
Full abduction	Hayes 2001 [43]	0.87	0.80 to **0.94**
Glenohumeral abduction	Hayes 2001 [43]	0.83	0.75 to **0.90**
Horizontal adduction	Hayes 2001 [43]	0.59	0.46 to 0.72
External rotation	Hayes 2001 [43]	0.81	0.71 to **0.91**
Internal rotation	Hayes 2001 [43]	0.82	0.73 to **0.92**

Table 4 (*Continued*)

Item	Reference	Statistic[a]	Variance[b]
Resisted tests			
Elbow flexion	Hayes 2001 [44]	0.48	0.07 to **0.89**
Elbow extension	Hayes 2001 [44]	0.52	0.00 to **1.00**
External rotation	Hayes 2001 [44]	0.60	0.32 to **0.88**
Internal rotation	Hayes 2001 [44]	0.63	0.34 to **0.92**
Abduction	Hayes 2001 [44]	0.67	0.40 to **0.93**
	Wadsworth 1987 [68]	PCC **0.98**	
Adduction	Hayes 2001 [44]	0.44	0.04 to 0.84
Oxford scale			
Elevation	Hayes 2002 [46]	ICC 0.79	0.51 to **0.94**
External rotation	Hayes 2002 [46]	ICC **0.86**	0.66 to **0.96**
Internal rotation	Hayes 2002 [46]	ICC **1.00**	**1.00 to 1.00**
Gerber lift-off	Hayes 2002 [46]	ICC 0.29	−0.08 to 0.71
Myofascial trigger point detection			
Taut band	Al-Shenqiti 2005 [37]	1.00	
Spot tenderness	Al-Shenqiti 2005 [37]	1.00	
Jump sign	Al-Shenqiti 2005 [37]	1.00	
Pain recognition	Al-Shenqiti 2005 [37]	1.00	
Referred pain	Al-Shenqiti 2005 [37]	0.79 to 0.88	
Local twitch response	Al-Shenqiti 2005 [37]	0.75 to 1.00	

ROM, range of movement; ICC, intraclass correlation coefficient; PCC, Pearson correlation coefficient; GHJ, glenohumeral joint.

[a] Kappa or weighted kappa except where stated.

[b] 95% confidence intervals, unless stated.

tain anatomical or pathological entities or to certain physical examination procedures. Out of 54 intra-tester investigations, 25 had values >0.85 and a further 11 had a range of values from 0.70 to 0.85.

Discussion

The reliability of commonly used shoulder examination procedures was examined systematically. According to the study criteria, none of the tests demonstrated acceptable levels of reliability, with overall evidence being contradictory. The studies had to include patients with symptoms, although asymptomatic volunteers could also be included, as it was considered that this most closely matched the real clinical situation. Approximately half of the studies were of high quality, with high-quality studies less likely to meet the criteria for acceptable reliability. A high standard was set for an acceptable level of reliability coefficient (0.85), but a sensitivity analysis was conducted with a lower level (0.70). The interpretation of reliability coefficients that have traditionally been used has suggested that much lower levels of reliability are acceptable, such as more than 0.40 for moderate reliability [31]. However, a number of authorities have challenged the acceptability of these levels of reliability, and suggested that levels need to be much higher to be acceptable [25,32–35]. Furthermore, the authors suggest that if the point of these tests is to make a diagnosis, which in turn will direct treatment, a reasonably high level of reliability is necessary to ensure that this clinical reasoning process is reproducible between clinicians. Unfortunately, this review did not find this to be the case.

Validity and reliability

To the authors' knowledge, no previous systematic review of shoulder physical examination procedures has been published, although a number of systematic reviews of the validity of these tests have been published [17–23]. These highlighted the generally poor levels of sensitivity and/or specificity of many tests. In fact, this review found that the evidence about reliability was contradictory for all tests, even those for which these reviews found some evidence. One review suggested that a rotator cuff tear might be suspected by combined Hawkins'/painful arc/infraspinatus tests, lift-off, belly press or drop-arm tests [20]; another suggested that Neer's and Hawkins' tests had reasonable sensitivity but poor specificity [19]. One review suggested that some tests had evidence of their anatomical validity, namely lift-off and Hawkins' tests, and possibly the active compression (O'Brien's) test and shoulder quadrant [18]. There was no indication that any of these tests with moderate validity were any more reliable than other tests, and there were no reliability studies for the belly press test and the quadrant position.

Reliability of physical examination procedures in general

There have been a number of systematic reviews of physical examination procedures for the sacro-iliac joint, lumbar and cervical spines, and trigger point diagnosis [26–30,73–75]. The conclusions from some of these reviews were very similar to the conclusion from the present review, with identification of poor levels of reliability in general [27–30,73,74]. However, some of the reviews also identified stronger levels of reliability in procedures that used symptom response rather than movement or palpation [26–30,73–75]. The present review did not find that procedures that were based on symptom response were more reliable.

Limitations

The strengths of the present review are its systematic nature, its use of multiple reviewers, the use of studies that included patients with symptoms, and the use of a high standard for reliability. However, as in all systematic reviews, the results are dependent on the studies included. Only studies in English were included, and it is not known if relevant articles in other languages were missed. Additional studies would be unlikely to make a significant difference to the overall conclusion, as it was contradictory. When quality scoring the studies, some of the items were rather ambivalent and produced some debate between the reviewers; the final decision was based on a majority of two out of three reviewers for six papers. Reliability studies mainly used kappa or ICC, and although these are commonly used and recommended [25], they do suffer from weaknesses. These relate to prevalence and bias indexes, in which a high prevalence or a homogeneous population would deflate coefficient values and high observer expectation bias would inflate values, and vice versa [76]. From this, it must be recognised that reliability is a relative and not an absolute property.

Clinical and research implications

It should be noted that tests are frequently used in conjunction with each other and to support clinical presentations. It might be that when used in this way, these tests are more reliable; however, this cannot be concluded from the present literature. Further research would be needed to confirm if this was the case. At present, given the findings from this review and previous systematic reviews into the validity of these tests, it is clear that due to their poor validity and reliability, they cannot be used to make the clinical diagnoses for which they were intended. The present authors concur with a previous suggestion that these should be abandoned in conservative care, because of the lack of uniformity, validity and reliability of diagnostic labels; instead, subgroups should be identified using reliable clinical characteristics [77]. The implications are that researchers need to start identifying clinical characteristics that have management and prognostic implications, and that clinicians should abandon the diagnostic pathological model which is based on tests that lack validity and reliability.

Conclusion

In conclusion, 36 reliability studies investigating physical examination procedures used in the assessment of patients with shoulder pain were identified. There was conflicting evidence about their reliability, and most fell below the predetermined levels of acceptable reliability. Using these procedures to make their associated diagnoses is an invalid and unreproducible process.

Ethical approval: Faculty of Health and Wellbeing Ethics Committee, Sheffield Hallam University, UK (WEC Reference No. 2009/28).

Conflict of interest: None declared.

References

[1] Luime JJ, Koes BW, Hendriksen IJM, Burdorf A, Verhagen AP, Miedema HS, *et al.* Prevalence and incidence of shoulder pain in the general population: a systematic review. Scand J Rheum 2004;33:73–81.

[2] Picavet HSJ, Schouten JSAG. Musculoskeletal pain in the Netherlands: prevalences, consequences and risk groups, the DMC_3-study. Pain 2003;102:167–78.

[3] van der Windt DAWM, Koes BW, de Jong BA, Bouter LM. Shoulder disorders in general practice: incidence, patient characteristics, and management. Ann Rheum Dis 1995;54:959–64.

[4] Ostor AJK, Richards CA, Prevost AT, Speed CA, Hazleman BL. Diagnosis and relation to general health of shoulder disorders presenting to primary care. Rheumatism 2005;44:800–5.

[5] May S. An outcome audit for musculoskeletal patients in primary care. Physiother Theory Pract 2003;19:189–98.

[6] Kuijpers T, van Tulder MW, van der Heijden GJMG, Bouter LM, van der Windt DAWM. Costs of shoulder pain in primary care consulters: a prospective cohort study in the Netherlands. BMC Musculoskel Dis 2006;7:83.

[7] Cyriax J. Textbook of orthopaedic medicine. Volume 1: diagnosis of soft tissue lesions. 8th edn. London: Bailliere Tindall; 1982.

[8] Carette S. Adhesive capsulitis—research advances frozen in time? J Rheum 2000;27:1329–31.

[9] Chard MD, Cawston TE, Riley GP, Gresham GA, Hazleman BL. Rotator cuff degeneration and lateral epicondylitis: a comparative histological study. Ann Rheum Dis 1994;53:3034.

[10] Khan KM, Cook JL, Taunton JE, Bonar F. Overuse tendinosis, not tendonitis: Part 1: a new paradigm for a difficult clinical problem. Phys Sports Med 2000;28:38–46.

[11] Lewis JS, Green AS, Dekel S. The aetiology of subacromial impingement syndrome. Phys Ther 2001;87:458–69.

[12] Tytherleigh-Strong G, Hirahara A, Miniaci A. Rotator cuff disease. Curr Opinion Rheum 2001;13:135–45.

[13] Liesdek C, van der Windt DA, Koes BW, Bouter LM. Soft-tissue disorders of the shoulder: a study of inter-observer agreement between general practitioners and physiotherapists and an overview of physiotherapeutic treatment. Physiotherapy 1997;83:12–7.

[14] De Winter AF, Jans MP, Scholten RJ, Deville W, Schaardenburg D, Bouter LM. Diagnostic classification of shoulder disorders: interobserver agreement and determinants of disagreement. Ann Rheum Dis 1999;58:272–7.

[15] Hanchard N, Cummins J, Jeffries C. Evidence-based clinical guidelines for the diagnosis, assessment and physiotherapy management of shoulder impingement syndrome. London: Chartered Society of Physiotherapy; 2004.

[16] Calis M, Akgun K, Birtane M, Karacan I, Calis H, Tuzun F. Diagnostic values of clinical diagnostic tests in subacromial impingement syndrome. Ann Rheum Dis 2000;59:44–7.

[17] Hegedus EJ, Goode A, Campbell S, Morin A, Tamaddoni M, Moorman CT, *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.

[18] Green R, Shanley K, Taylor NF, Perrott M. The anatomical basis for clinical tests assessing musculoskeletal function of the shoulder. Phys Ther Rev 2008;13:17–24.

[19] Beaudreuil J, Nizard R, Thomas T, Peyre M, Liotard JP, Boileau P, *et al.* Contribution of clinical tests to the diagnosis of rotator cuff disease: a systematic literature review. Joint Bone Spine 2009;76:15–9.

[20] Hughes PC, Taylor NF, Green RA. Most clinical tests cannot accurately diagnose rotator cuff pathology: a systematic review. Aus J Physiother 2008;54:159–70.

[21] Dessau WA, Magarey ME. Diagnostic accuracy of clinical tests for superior labral anterior posterior lesions: a systematic review. J Orthop Sports Phys Ther 2008;38:341–52.

[22] Mirkovic M, Green R, Tayor N, Perrott M. Accuracy of clinical tests to diagnose superior labral anterior and posterior (SLAP) lesions. Phys Ther Rev 2005;10:5–14.

[23] Munro W, Healy R. The validity and accuracy of clinical tests used to detect labral pathology of the shoulder—a systematic review. Man Ther 2009;14:119–30.

[24] Spratt K. Statistical relevance. In: Orthopaedic knowledge update spine. 2nd edn. American Academy of Orthopaedic Surgeons; 2002. p. 497–505.

[25] Streiner DL, Norman GR. Health measurement scales. 3rd edn Oxford: Oxford University Press; 2003.

[26] van der Wurff P, Hagmeijer RHM, Meyne W. Clinical tests of the sacroiliac joint. A systematic methodological review. Part 1. Reliability. Man Ther 2000;5:30–6.

[27] May S, Littlewood C, Bishop A. Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. Aust J Physiother 2006;52:91–102.

[28] Seffinger MA, Najm WI, Mishra SI, Adams A, Dickerson VM, Murphy LS, *et al.* Reliability of spinal palpation for diagnosis of back and neck pain. A systematic review of the literature. Spine 2004;29:E413–25.

[29] Stochkendahl MJ, Christensen HW, Hartvigsen J, Vach W, Haas M, Hestbaek L, *et al.* Manual examination of the spine: a systematic critical literature review of reproducibility. J Manip Physiol Ther 2006;29:475–85.

[30] van Trijffel E, Anderegg Q, Bossuyt PMM, Lucas C. Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. Man Ther 2005;10: 256–69.

[31] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

[32] McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. 2nd edn. New York: Oxford University Press; 1987.

[33] Fleiss J. The design and analysis of clinical experiments. New York: John Wiley; 1986.

[34] Ware JE, Brook RH, Davies AR, Lohr KN. Choosing measures of health status for individuals in general populations. Am J Public Health 1981;71:620–5.

[35] Weiner EA, Stewart BJ. Assessing individuals. Boston: Little Brown; 1984.

[36] van Tulder M, Furlan A, Bombardier C, Bouter L. Updated method guidelines for systematic reviews in the Cochrane Collaboration Back Review Group. In: The Cochrane Library, Issue 4. Oxford: Update Software; 2003.

[37] Al-Shenqiti AM, Oldham JA. Test–retest reliability of myofascial trigger point detection in patients with rotator cuff tendonitis. Clin Rehabil 2005;19:482–7.

[38] Bertilson BC, Grunnesjo M, Strender L. Reliability of clinical tests in the assessment of patients with neck/shoulder problems—impact of history. Spine 2003;28:2222–31.

[39] Borstad JD, Mathiowetz KM, Minday LE, Pradhu B, Christopherson DE, Ludewig PM. Clinical measurement of posterior shoulder flexibility. Man Ther 2007;12:386–9.

[40] Bron C, Franssen J, Wensing M, Oostendorp RAB. Interrater reliability of palpation of myofascial trigger points in three shoulder muscles. J Man Manip Ther 2007;15:203–15.

[41] Chesworth BM, MacDermid JC, Roth JH, Patterson SD. Movement diagram and "end-feel" reliability when measuring passive lateral rotation of the shoulder in patients with shoulder pathology. Phys Ther 1998;78:593–601.

[42] Croft P, Pope D, Boswell R, Rigby A, Silman A. Observer variability in measuring elevation and external rotation of the shoulder. Br J Rheum 1994;33:942–6.

[43] Hayes KW, Petersen CM. Reliability of assessing end-feel and pain and resistance sequence in subjects with painful shoulder and knees. J Orthop Sports Phys Ther 2001;31:432–45.

[44] Hayes KW, Petersen CM. Reliability of classifications derived from Cyriax's resisted testing in subjects with painful shoulders and knees. J Orthop Sports Phys Ther 2001;33:235–46.

[45] Hayes K, Walton JR, Szomor ZL, Murrell GAC. Reliability of five methods for assessing shoulder range of motion. Aust J Physiother 2001;47:289–94.

[46] Hayes K, Walton JR, Szomor ZL, Murrell GAC. Reliability of 3 methods for assessing shoulder strength. J Should Elbow Surg 2002;11:33–9.

[47] Hickey BW, Milosavljevic S, Bell ML, Milburn PD. Accuracy and reliability of observational motion analysis in identifying shoulder symptoms. Man Ther 2007;12:263–70.

[48] Johansson K, Ivarson S. Intra- and interexaminer reliability of four manual shoulder maneuvers used to identify subacromial pain. Man Ther 2008;14:231–9.

[49] Kibler WB, Uhl TL, Maddux JWQ, Brooks PV, Zeller B, McMullen J. Qualitative clinical evaluation of scapular dysfunction: a reliability study. J Should Elbow Surg 2002;11:550–6.

[50] Kim S, Ha K, Han K. Biceps load test: a clinical test for superior labrum anterior and posterior lesions in shoulders with recurrent anterior dislocations. Am J Sports Med 1999;27:300–3.

[51] Kim Y, Kim J, Ha K, Coy S, Joo M, Chung Y. The passive compression test: a new clinical test for superior labral tears of the shoulder. Am J Sports Med 2007;35:1489–94.

[52] Lewis JS, Valentine RE. The pectoralis minor length test: a study of the intra-rater reliability and diagnostic accuracy in subjects with and without shoulder symptoms. BMC Musculo Dis 2007;8:64.

[53] Lewis JS, Valentine RE. Intraobserver reliability of angular and linear measurement of scapular position in subjects with and without symptoms. Arch Phys Med Rehabil 2008;89:1795–802.

[54] Lo IKY, Nonweiler B, Woolfrey M, Litchfield R, Kirkley A. An evaluation of the apprehension, relocation, and surprise tests for anterior shoulder instability. Am J Sports Med 2004;32:301–7.

[55] McClure P, Tatae AR, Kareha S, Irwin D, Zlupko E. A clinical method for identifying scapular dyskinesis. Part 1. Reliability. J Athl Train 2009;44:160–4.

[56] Nanda R, Gupta S, Kanapathipillai P, Liow RY, Rangan A. An assessment of the inter-examiner reliability of clinical tests for subacromial impingement and rotator cuff integrity. Eur J Orthop Surg Traumatol 2008;18:495–500.

[57] Nijs J, Roussel N, Vermeulen K, Souvereyns G. Scapular positioning in patients with shoulder pain: a study examining the reliability and clinical importance of 3 clinical tests. Arch Phys Med Rehabil 2005;86:1349–55.

[58] Nomden JG, Slagers AJ, Bergman GJD, Winters JC, Kropmans TJB, Dijkstra PU. Interobserver reliability of physical examination of shoulder girdle. Man Ther 2008;14:206–12.

[59] Odom CJ, Taylor AB, Hurd CE, Denegar CR. Measurement of scapular asymmetry and assessment of shoulder dysfunction using the lateral scapular slide test: a reliability and validity study. Phys Ther 2001;81:799–809.

[60] Ostor AJK, Richards CA, Prevost AT, Hazleman BL, Speed CA. Interrater reproducibility of clinical tests for rotator cuff lesions. Ann Rheum Dis 2004;63:1288–92.

[61] Palmer K, Walker-Bone K, Linaker C, Reading I, Kellingray S, Coggon D, *et al.* The Southampton examination schedule for the diagnosis of musculoskeletal disorders of the upper limb. Ann Rheum Dis 2000;59:5–11.

[62] Rabin A, Irrgang JJ, Fitzgerald GK, Eubanks A. The intertester reliability of the scapular assistance test. J Orthop Sports Phys Ther 2006;36:653–60.

[63] Razmjou H, Holtby R, Myhr T. Pain provocative shoulder tests: reliability and validity of the impingement tests. Physio Can 2004;56:229–37.

[64] Terwee CB, de Winter AF, Scholten RJ, Jans MP, Deville W, van Schaardenburg D, *et al.* Interobserver reproducibility of the visual estimation of range of motion. Arch Phys Med Rehabil 2005;86:1356–61.

[65] Tzannes A, Paxinos A, Callanan M, Murrell GAC. An assessment of the interexaminer reliability of tests for shoulder instability. J Should Elbow Surg 2004;13:18–23.

[66] Valentine RE, Lewis JS. Intraobserver reliability of 4 physiologic movements of the shoulder in subjects with and without symptoms. Arch Phys Med Rehabil 2006;87:1242–9.

[67] van Duijn AJ, Jensen RH. Reliability of inferior glide mobility testing of the glenohumeral joint. J Man Manip Ther 2001;9:109–14.

[68] Wadsworth CT, Krishnan R, Sear M, Harrold J, Nielsen DH. Intrarater reliability of manual muscle testing and hand-held dynametric muscle testing. Phys Ther 1987;67:1342–7.

[69] Walker-Bone K, Byng P, Linaker C, Reading I, Coggon D, Palmer KT, *et al.* Reliability of the Southampton examination schedule for the diagnosis of upper limb disorders in the general population. Ann Rheum Dis 2002;61:1103–6.

[70] Walsworth MK, Doukas WC, Murphy KP, Mielcarek BJ, Michener LA. Reliability and diagnostic accuracy of history and physical examination for diagnosing glenoid labral tears. Am J Sports Med 2008;36: 162–8.

[71] Westerberg C, Solem-Bertoft E, Lundh I. The reliability of three active motor tests used in painful shoulder disorders. Scand J Rehabil Med 1996;28:63–70.

[72] Yang J, Lin J. Reliability of function-related tests in patients with shoulder pathologies. J Orthop Sports Phys Ther 2006;36:572–6.

[73] Hestboek L, Leboeuf-Yde C. Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic critical literature review. J Manip Physiol Ther 2000;23:258–75.

[74] Hollerwoger D. Methodological quality and outcomes of studies address manual cervical spine examinations: a review. Man Ther 2006;11:93–8.

[75] Lucas N, Macaskill P, Irwig L, Moran R, Bogduk N. Reliability of physical examination for diagnosis of myofascial trigger points. A systematic review of the literature. Clin J Pain 2009;25:80–90.

[76] Sim J, Wright CC. The kappa statistic in reliability studies: use interpretation, and sample size requirements. Phys Ther 2005;85:257–68.

[77] Schellingerhout JM, Verhagen AP, Thomas S, Koes BW. Lack of uniformity in diagnostic labelling of shoulder pain: time for a different approach. Man Ther 2008;13:478–83.