
Journal of **PHYSIOTHERAPY**

VOLUME 56 • NO 1 • 2010

Official Journal of the Australian Physiotherapy Association



©Daniel Beeron, Alison Murray and John Murray, 2009. Licensed by Girringun Aboriginal Corporation, 2010

Editorial

5 A new journal name for a new decade

Research

- 7 Measurement of passive joint movements
- 19 Adolescent breast support education
- 27 Therapeutic ultrasound and sinusitis
- 33 New PEP device for COPD
- 41 Exercise adherence in people with osteoarthritis
- 49 Lower limb pain in Tibet

Appraisal

- 57 Critically Appraised Papers
- 62 Clinimetrics
- 64 Clinical Practice Guidelines



AUSTRALIAN PHYSIOTHERAPY ASSOCIATION

Inter-rater reliability for measurement of passive physiological range of motion of upper extremity joints is better if instruments are used: a systematic review

Rachel J van de Pol¹, Emiel van Trijffel² and Cees Lucas²

¹Private Practice Physiotherapy, The Hague, ²University of Amsterdam, Academic Medical Centre, Amsterdam The Netherlands

Question: What is the inter-rater reliability for measurements of passive physiological or accessory movements in upper extremity joints? **Design:** Systematic review of studies of inter-rater reliability. **Participants:** Individuals with and without upper extremity disorders. **Outcome measures:** Range of motion and end-feel using methods feasible in clinical practice. **Results:** Twenty-one studies were included of which 11 demonstrated acceptable inter-rater reliability. Two studies satisfied all criteria for internal validity while reporting almost perfect reliability. Overall, the methodological quality of studies was poor. ICC ranged from 0.26 (95% CI -0.01 to 0.69) for measuring the physiological range of shoulder internal rotation using vision to 0.99 (95% CI 0.98 to 1.0) for the physiological range of finger and thumb flexion/extension using a goniometer. Measurements of physiological range of motion using instruments were more reliable than using vision. Measurements of physiological range of motion were also more reliable than measurements of end-feel or of accessory range of motion. **Conclusions:** Inter-rater reliability for the measurement of passive movements of upper extremity joints varies with the method of measurement. In order to make reliable decisions about joint restrictions in clinical practice, we recommend that clinicians measure passive physiological range of motion using goniometers or inclinometers. [van de Pol RJ, van Trijffel E, Lucas C (2010) Inter-rater reliability for measurement of passive physiological range of motion of upper extremity joints is better if instruments are used: a systematic review. *Journal of Physiotherapy* 56: 7–17]

Key words: Upper extremity; Reliability; Physical examination; Range of motion

Introduction

Physiotherapists commonly assess and treat upper extremity disorders. Passive joint mobilisation or manipulation has been shown to be effective in disorders such as adhesive shoulder capsulitis, non-specific shoulder pain or dysfunction (Ho et al 2009), shoulder impingement syndrome (Kromer et al 2009), lateral epicondylalgia (Bisset et al 2005), and carpal tunnel syndrome (O'Connor et al 2003). Measurement of passive movement is indicated in order to assess joint restrictions and to help diagnose these disorders. Passive movement, either physiological or accessory, can be reported as range of motion, end-feel, or pain and is an indication of the integrity of joint structures (Cyriax 1982, Hengeveld and Banks 2005). Passive physiological range of motion may be measured using vision or instruments such as goniometers and inclinometers.

An essential requirement of clinical measures is that they are valid and reliable so that they can be used to discriminate between individuals (Streiner and Norman 2008). Inter-rater reliability is a component of reproducibility along with agreement and refers to the relative measurement error, ie, the variation between patients as measured by different raters in relation to the total variance of the measures (Streiner and Norman 2008). Agreement, on the other hand, provides insight into the ability of a clinical measure to yield the same value on multiple occasions and reflects absolute measurement error (De Vet et al 2006). High inter-rater reliability for measurements of upper extremity joints is a prerequisite for valid and uniform decisions about joint restrictions (Bartko and Carpenter 1976).

Many studies investigating the reliability of passive movements of human joints have been conducted. However, relatively few reviews have summarised and appraised the evidence. For example, seven systematic reviews have been published on passive spinal movement (Haneline et al 2008, Hestbæk and Leboeuf-Yde 2000, May et al 2006, Seffinger et al 2004, Stochkendahl et al 2006, Van Trijffel et al 2005, Van der Wurff et al 2000). In general, inter-rater reliability was found to be poor and studies were of poor methodological quality. To date, no systematic appraisal of studies on inter-rater reliability of measurement of passive movement in upper extremity joints has been conducted. Therefore, the research question for this systematic review was:

What is the inter-rater reliability for measurements of passive physiological or accessory movements in upper extremity joints?

Method

Identification and selection of studies

MEDLINE (PubMed) was searched by two reviewers (RJvdP, EvT) independently for studies published between January 1 1966 and July 1 2009. Search terms included all relevant upper extremity joints and all synonyms for *reliability* and *rater* (see Appendix 1 on eAddenda for detailed search strategy). Additional searches in CINAHL (1982 to July 1 2009) and EMBASE (1996 to July 1 2009) were performed by one reviewer (RJvdP). In addition, reference lists of all retrieved papers were hand searched for relevant studies.

The titles and abstracts were screened by two reviewers (RJvdP, EvT) independently. When relevant, full text papers were retrieved. Studies were included if they met all inclusion criteria (Box 1). No restrictions were imposed on language or date of publication. Abstracts and documents that were anecdotal, speculative, or editorial in nature, were not included. Studies investigating active movement or restriction in passive movement due to pain or ligament instability as well as animal or cadaver studies were not considered for inclusion. Studies of people with neurological conditions in which abnormal muscle tone may interfere with joint movement, or of people after arthroplasty were also excluded. Disagreements on eligibility were first resolved by discussion and decided by a third reviewer (CL) if disagreement persisted.

Box 1. Inclusion criteria

Design
• Repeated measures between raters
Participants
• Symptomatic and asymptomatic individuals
Measurement procedure
• Performed passive (ie, manual) physiological or accessory movements in any of the joints of the shoulder, elbow, or wrist-hand-fingers
• Reported range of motion or end-feel
• Used methods feasible in clinical practice (considering instruments, costs, amount of training required)
Outcomes
• Estimates of inter-rater reliability

Assessment of characteristics of the studies

Description: We extracted data on participants (number, age, clinical characteristics), raters (number, profession, training), measurements (joints and movement direction, position, movement performed, method, outcomes reported), and inter-rater reliability (point estimates, estimates of precision). Two reviewers (RJvdP and EvT) extracted data independently and were not blind to journal, authors, or results. When disagreement between reviewers could not be resolved by discussion, a third reviewer (CL) made the final decision.

Quality: No validated instrument is available for assessing methodological quality of inter-rater reliability studies. Therefore, a list of criteria for quality was compiled derived from the QUADAS tool, the STARD Statement, and criteria used for assessing studies on reliability of measuring passive spinal movements (Bossuyt et al 2003a, Bossuyt et al 2003b, Van Trijffel et al 2005, Whiting et al 2003). Criteria were rated ‘yes’, ‘no’, or ‘unknown’ where insufficient information was provided (Box 2). Criteria 1 to 4 assess external validity, Criteria 5 to 9 assess internal validity, and Criterion 10 assesses statistical methods. External validity was considered sufficient if Criteria 1 to 4 were rated ‘yes’. With respect to internal validity, Criteria 5, 6, and 7 were assumed to be decisive in determining risk of bias. A study was considered to have a low risk of bias if Criteria 5, 6, and 7 were all rated ‘yes’, a moderate risk if two of these criteria were rated ‘yes’, and a high risk if none or only one of these criteria were rated ‘yes’. After training, two reviewers (RJvdP, EvT) independently assessed methodological quality of all included studies and were not blind to journal,

Box 2. Criteria for assessing methodological quality

1. Was a representative sample of participants used?
2. Was a representative sample of raters used?
3. Is replication of the measurement procedure possible?
4. Was clinical information from participants available to raters and comparable to clinical practice?
5. Were participants’ characteristics stable during the study?
6. Were raters’ characteristics stable during the study?
7. Were raters blinded to each other’s results?
8. Can non-random loss to follow-up be ruled out?
9. Was an estimate of intra-rater reliability validly determined and was it above 0.80?
10. Were appropriate measures (Kappa, ICC) used for calculating reliability?

authors, and results. If discrepancy between reviewers persisted after discussion, a decisive judgement was passed by the third reviewer (CL).

Data analysis

Data were analysed by examining ICC and Kappa (95% CI). ICC > 0.75 indicated an acceptable level of reliability (Burdock et al 1963, cited by Kramer and Feinstein 1981). Corresponding Kappa levels were used as assigned by Landis and Koch (1977) where <0.00 = poor, 0.00–0.20 = slight, 0.21–0.40 = fair, 0.41–0.60 = moderate, 0.61–0.80 = substantial, and 0.81–1.00 = almost perfect reliability. In addition, reliability was analysed relating it to methodological quality and risk of bias. Reliability from studies not fulfilling Criteria 5 or 6 could have been underestimated, while reliability from studies not fulfilling Criterion 7 could have been overestimated. Negative scores on combinations of Criteria 5–7 could have led to bias in an unknown direction. Where one or more of these three criteria were unknown, no statement was made regarding the presence or direction of potential bias. Finally, because of clinical and methodological heterogeneity between studies, we did not attempt to statistically summarise data by calculating pooled estimates of reliability.

Results

Flow of studies through the review

Searching MEDLINE yielded 326 citations of which 26 papers were retrieved in full text. CINAHL (95 citations) and EMBASE (34) yielded no additional relevant articles. Hand searching supplied another 20 potentially relevant studies. Of these 46, 25 studies were excluded (see Appendix 2 on eAddenda for excluded studies). In total, 21 studies fulfilled all inclusion criteria (Figure 1).

Description of studies

The included studies are summarised in Table 1. Thirteen studies investigated inter-rater reliability of measurement of passive shoulder movements (Awan et al 2002, Chesworth et al 1998, De Winter et al 2004, Hayes et al 2001, Hayes and Petersen 2001, Heemskerk et al 1997, Lin and Yang 2006, MacDermid et al 1999, Nomden et al 2009, Riddle et al 1987, Terwee et al 2005, Tyler et al 1999, Van Duijn and Jensen 2001), two investigated elbow movements (Patla and Paris 1993, Rothstein et al 1983), four investigated wrist movements (Bovens et al 1990, Horger 1990, LaStayo

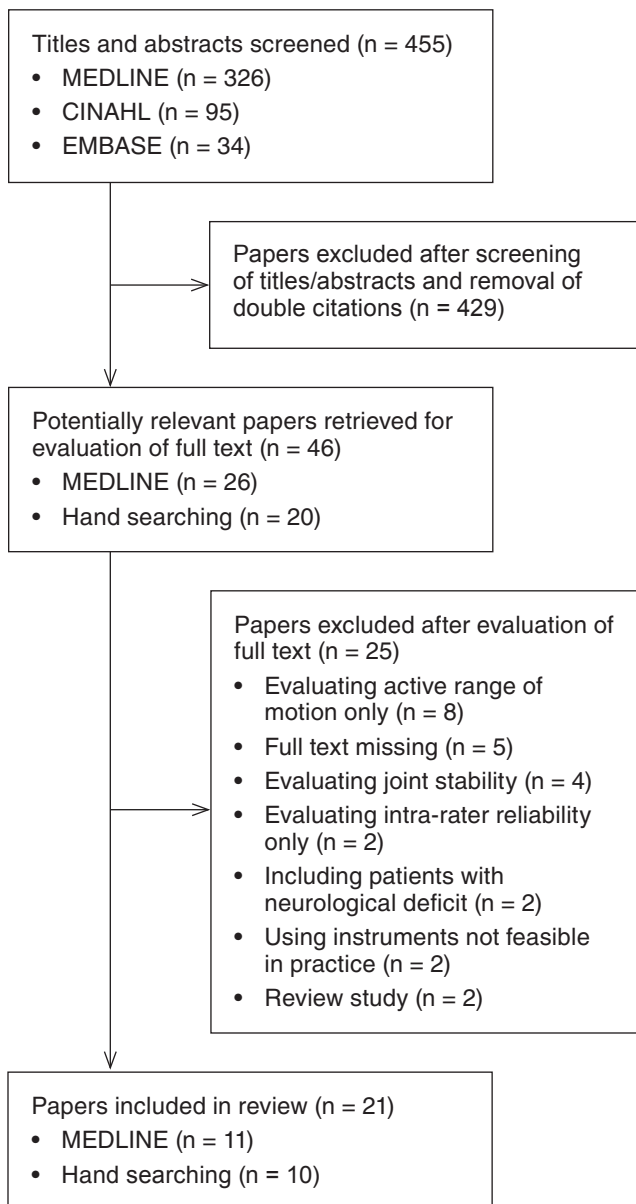


Figure 1. Flow of studies through the review

and Wheeler 1994, Staes et al 2009), one investigated phalangeal joint movements (Glasgow et al 2003), and one investigated thumb movements (De Kraker et al 2009). In all except two studies (Bovens et al 1990, De Kraker et al 2009), physiotherapists acted as raters. There were no disagreements between reviewers on selection of studies.

Quality of studies

The methodological quality of included studies is presented in Table 2. One study (MacDermid et al 1999) fulfilled all four criteria for external validity and four studies satisfied three criteria. Two studies (Glasgow et al 2003, Nomden et al 2009) fulfilled all three criteria for internal validity representing a low risk of bias, while six studies satisfied two criteria. Criteria on internal and external validity could not be scored on 52 (28%) occasions because of insufficient reporting. Twenty (10%) disagreements occurred between reviewers which were all resolved by discussion.

Inter-rater reliability by joint

The inter-rater reliability for measurement of physiological range of motion is presented in Table 3, accessory range of motion in Table 4 and physiological end-feel in Table 5.

Shoulder (n = 13): One study (MacDermid et al 1999) fulfilled all criteria for external validity and another (Nomden et al 2009) fulfilled all criteria for internal validity. ICC for measurement of physiological range of motion using vision ranged from 0.26 (95% CI -0.01 to 0.69) for internal rotation (Hayes et al 2001) to 0.96 for abduction (Nomden et al 2009). In seven studies (Chesworth et al 1998, De Winter et al 2004, Heemskerk et al 1997, Lin and Yang 2006, MacDermid et al 1999, Nomden et al 2009, Tyler et al 1999) acceptable reliability (ICC > 0.75) was reached. The highest reliability occurred in Nomden et al (2009) and was associated with a low risk of bias for patients with shoulder pathology using trained, experienced physiotherapists of which one was a specialist in manual therapy. In general, measuring passive physiological range of motion using instruments, such as goniometers or inclinometers, resulted in higher reliability than using vision. Of the four studies classified as having a moderate risk of bias (Awan et al 2002, De Winter et al 2004, Terwee et al 2005, Van Duijn and Jensen 2001), one (De Winter et al 2004) reported acceptable reliability for measuring abduction (ICC 0.83) and external rotation (ICC 0.90) using an inclinometer. The externally valid study by MacDermid et al (1999) reported acceptable reliability (ICC 0.86, 95% CI 0.72 to 0.92 and ICC 0.85, 95% CI 0.73 to 0.91) for measuring external rotation in symptomatic individuals by two experienced physiotherapists with advanced manual therapy training. In the one study investigating accessory range of motion of the glenohumeral joint (inferior gliding), reliability was found to be unacceptable (ICC 0.52) (Van Duijn and Jensen 2001). Overall, measurements of range of motion were more reliable than measurements of end-feel. Kappa for end-feel ranged from 0.26 (95% CI -0.16 to 0.68) in full shoulder abduction to 0.70 (95% CI 0.31 to 1.0) in abduction with scapula stabilisation (Hayes and Petersen 2001). No specific movement direction was consistently associated with high or low reliability.

Elbow (n = 2): Neither of the studies fulfilled all criteria for external or internal validity. Rothstein et al (1983) demonstrated acceptable reliability for measuring range of flexion (ICC from 0.85 to 0.97) and extension (0.92 to 0.95) using different types of goniometers in patients with elbow pathology. The reliability of measurements of physiological range of motion reported by Rothstein et al (1983) was substantially higher than the reliability of measurements of end-feel of flexion (Kappa 0.40) and extension (Kappa 0.73) reported by Patla and Paris (1993).

Wrist-hand-fingers (n = 6): One study (Glasgow et al 2003) satisfied all criteria for internal validity. Almost perfect reliability (ICC 0.99, 95% CI 0.98 to 1.0), associated with a low risk of bias, was reported for measurements of passive torque-controlled physiological range of finger and thumb flexion/extension using a goniometer in patients with a traumatic hand injury (Glasgow et al 2003). Three studies (Bovens et al 1990, Horger 1990, LaStayo and Wheeler 1994) investigated the reliability of measurements of physiological range of motion at the wrist of which the latter two reported acceptable ICC values for wrist extension (ICC 0.80 to 0.84) and flexion (ICC 0.86 to 0.93) using goniometers. In contrast, Bovens et al (1990) reported poor reliability for

Table 1. Summary of included studies (n = 21).

Study	Participants	Raters	Joints Movement direction	Position	Movement performed	Method	Outcome reported	Reliability statistic
Awan et al (2002)	n = 56 Age = range 13–18 yr Condition = normal	n = 4 Profession = 2 physiatrists, 1 PT, 1 resident doctor Training = Y	Shoulder • IR • ER	Supine Sh 90° Abd	Physiological	Digital inclinometer Vision	ROM	ICC
Bovens et al (1990)	n = 148 Age = mean 48 yr (SD 7) Condition = normal	n = 3 Profession = physician Training = Y	Wrist-hand-fingers • Wrist F • Wrist E	Palms together Hands together	Physiological	Vision	ROM	R
Chesworth et al (1998)	n = 34 Age = mean 55 yr (SD 18.5) Condition = shoulder pathology, post-surgery	n = 2 Profession = PT/MT Training = N	Shoulder • ER	Supine Sh 20° Abd Elbow 90° F	Physiological	Vision Manual	ROM End-feel	ICC (2,1)
De Kraker et al (2009)	n = 25 Age = mean 30 yr (SD 7) Condition = normal	n = 2 Profession = 1 HT, 1 trainee plastic and reconstructive surgery Training = N	Wrist-hand-fingers • Thumb Abd	Seated Elbow 90° F Wrist neutral	Physiological	Pollexograph Goniometer	ROM	ICC
De Winter et al (2004)	n = 155 Age = mean 47 yr (SD 12.6) Condition = shoulder path	n = 2 Profession = PT Training = Y	Shoulder • Abd • ER	Seated Supine	Physiological	Digital inclinometer	ROM	ICC
Glasgow et al (2003)	n = 10 Age = mean 39.7 yr (SD 13.5) Condition = traumatic hand injuries	n = 2 Profession = unknown Training = N	Hand-wrist-fingers • IP F • IP E • MCP F • MCP E • Thumb F • Thumb E	Unknown	Physiological	Goniometer	ROM	ICC (2,1)
Hayes et al (2001)	n = 8 Age = mean 66 yr (SD 5.7) Condition = shoulder pathology, post-surgery	n = 4 Profession = 2 PT, 1 orthopaedic surgeon, 1 sports physician Training = Y	Shoulder • F • Abd • ER • IR	Seated	Physiological	Vision	ROM	ICC (2,1)

Author(s)	n	Age	Condition	n = 2	Profession	Training	Shoulder	Standing (with and without scapular stabilisation)	Physiological	Manual	End-feel	Kappa
Hayes & Petersen (2001)	n = 18	Age = mean 34.3 yr (SD 12.9)	Condition = shoulder pain	n = 2	Profession = PT	Training = Y	Shoulder • Abd • ER • IR • Hor Add • Full Abd	Standing and without scapular stabilisation	Physiological	Manual	End-feel	Kappa
Heemskerk et al (1997)	n = 12	Age = mean 36 yr (range 25-49)	Condition = normal	n = 2	Profession = PT	Training = N	Shoulder • Abd • ER	Seated Supine	Physiological	Digital inclinometer	ROM	ICC
Horger (1990)	n = 48	Age = mean 38.8 yr (range 18-71)	Condition = wrist injuries	n = 26	Profession = 11 OT, 2 PT, 6 HT, 7 non-specialised raters	Training = N	Wrist • E • F • Abd • Add	Unknown	Physiological	Goniometer	ROM	ICC (1,1)
LaStayo & Wheeler (1994)	n = 140	Age = mean 41.5 yr (range 6-81)	Condition = wrist pathology	n = 32	Profession = 25 OT, 6 PT, 1 OT/PT (17 of which HT)	Training = N	Wrist • E • F	Unknown	Physiological	Goniometer	ROM	ICC (2,1)
Lin & Yang (2006)	n = 16	Age = mean 54.5 yr (SD 9.2)	Condition = shoulder stiffness	n = 2	Profession = PT	Training = N	Shoulder • Hor F • Hor E	Supine Sh 90° F Sh 0° adduction Scapula stabilised	Physiological	Inclinometer	ROM	ICC (3,1)
MacDermid et al (1999)	n = 34	Age = mean 55 yr (SD 18)	Condition = shoulder pathology, post-surgery	n = 2	Profession = PT/MT	Training = N	Shoulder • ER	Supine Sh 20° to 30° Abd Elbow 90° F	Physiological	Goniometer	ROM	ICC
Nomden et al (2009)	n = 91	Age = mean 48.5 yr (SD 11.8)	Condition = shoulder pathology	n = 2	Profession = 1 PT, 1 PT/MT	Training = N	Shoulder • Abd • ER	Seated Abd: Sh 0° Abd Sh ER Thumb up ER: Sh 0° F Elbow 90° F	Physiological	Vision	ROM	ICC (1,1)
Patla & Paris (1993)	n = 20	Age = ?	Condition = normal, elbow pathology	n = 2	Profession = PT	Training = Y	Elbow • E • F	Standing Sh 20° Abd Elbow 20° F	Physiological	Goniometer Manual	ROM End-feel	Kappa

Study	Participants	Raters	Joints Movement direction	Position	Movement performed	Method	Outcome reported	Reliability statistic
Riddle et al (1987)	n = 50 Age = mean 48.6 yr (SD 14.4) Condition = shoulder pathology	n = 16 Profession = PT Training = N	Shoulder • F • E • Abd • Hor Abd • Hor Add • ER • IR	Supine Prone Seated Side lying Standing	Physiological	Goniometer	ROM	ICC (1,1)
Rothstein et al (1983)	n = 12 Age = ? Condition = elbow pathology	n = 12 Profession = PT Training = N	Elbow • E • F	Unknown	Physiological	Goniometer	ROM	ICC
Staes et al (2009)	n = 30, 15 Age = mean 21.3 yr (SD 1.6), mean 38.3 yr (SD 11) Condition = normal, wrist pathology	n = 2 Profession = PT Training = Y	Wrist • Hamate • Lunate • Scaphoid • Trapezoid against Capitata	Resting position	Accessory	Vision	ROM End-feel	Weighted Kappa
Terwee et al (2005)	n = 201 Age = mean 48 yr (SD 12) Condition = shoulder pathology	n = 2 Profession = PT Training = Y	Shoulder • ELE • Abd • ER • Hor Add	Seated Sh 0° F Elbow 90° F	Physiological	Vision	ROM	ICC (2,1)
Tyler et al (1999)	n = 28 Age = mean 30 yr (SD 8.9) Condition = normal	n = 2 Profession = PT Training = N	Shoulder • Hor F	Side lying Sh 90° Abd Scapula stabilised	Physiological	Measuring tape	ROM	ICC (3,k)
Van Duijn & Jensen (2001)	n = 18 Age = mean 36.6 yr (SD 10) Condition = shoulder pathology, normal	n = 6 Profession = PT Training = N	Shoulder • Inf glide	Supine	Accessory	Vision	ROM	ICC (2,1)

Abd = abduction, Add = adduction, ELE = elevation, ER = external rotation, E = extension, F = flexion, HT = hand therapist, Hor = horizontal, INF GL = inferior glide, IP = interphalangeal, IR = internal rotation, MCP = metacarpophalangeal, MED = medial, MT = manual therapist, N = No, OT = occupational therapist, PT = physiotherapist, ROM = range of motion, Sh = shoulder, Y = Yes

Table 2. Methodological quality of included studies by joint.

Study	External validity				Internal validity					Statistical methods
	1	2	3	4	5	6	7	8	9	10
Shoulder										
Awan et al (2002)	N	U	Y	N	Y	U	Y	Y	N	U
Chesworth et al (1998)	Y	N	Y	Y	N	U	Y	Y	N	Y
De Winter et al (2004)	Y	U	Y	U	Y	Y	U	N	N	U
Hayes et al (2001)	Y	U	Y	U	N	U	Y	Y	N	Y
Hayes & Petersen (2001)	N	Y	Y	Y	U	U	Y	Y	N	U
Heemskerck et al (1997)	N	U	N	N	U	N	Y	Y	N	U
Lin & Yang (2006)	Y	U	Y	U	U	U	Y	Y	N	Y
MacDermid et al (1999)	Y	Y	Y	Y	U	U	Y	Y	N	U
Nomden et al (2009)	Y	Y	Y	U	Y	Y	Y	Y	U	Y
Riddle et al (1987)	Y	Y	U	N	U	U	Y	Y	N	Y
Terwee et al (2005)	Y	Y	Y	N	Y	U	Y	Y	U	Y
Tyler et al (1999)	N	U	Y	N	Y	U	U	Y	Y	Y
Van Duijn & Jensen (2001)	N	Y	Y	N	Y	U	Y	Y	N	N
Elbow										
Patla & Paris (1993)	N	U	Y	U	Y	U	U	Y	U	Y
Rothstein et al (1983)	U	Y	U	N	U	U	Y	Y	N	U
Wrist-hand-fingers										
Bovens et al (1990)	N	U	Y	U	Y	Y	U	Y	U	U
De Kraker et al (2009)	N	U	Y	U	U	N	Y	Y	N	U
Glasgow et al (2003)	Y	U	Y	N	Y	Y	Y	N	Y	Y
Horger (1990)	Y	Y	N	U	U	U	Y	Y	N	Y
LaStayo & Wheeler (1994)	Y	Y	Y	N	U	U	Y	Y	N	Y
Staes et al (2009)	N	U	Y	N	U	Y	Y	N	N	U

U = unknown because insufficient information provided

measurements by physicians of physiological wrist extension using vision. Reliability for measuring physiological thumb abduction was reported to be higher using a pollexograph (ICC 0.59, 95% CI 0.42 to 0.89) than a goniometer (ICC 0.37, 95% CI -0.42 to 0.79). Finally, measuring accessory movements of carpal bones against the capitate bone using a 3-point scale yielded fair to moderate reliability (weighted Kappa from 0.29 to 0.42) in healthy individuals and fair to almost perfect reliability (weighted Kappa from 0.33 to 0.87) in post-operative patients (Staes et al 2009).

Discussion

This systematic review included 21 studies investigating inter-rater reliability of measurements of passive movements of upper extremity joints, of which 11 demonstrated acceptable reliability (ICC > 0.75). Reliability varied considerably with the method of measurement and ICC ranged from 0.26 (95% CI -0.01 to 0.69) for measuring the physiological range of shoulder internal rotation using vision to 0.99 (95% CI 0.98 to 1.0) for the physiological range of finger and thumb flexion/extension using a goniometer. In general, measurements of physiological range of motion using instruments were more reliable than measurements using vision. Furthermore, measurements of physiological range of motion were also more reliable than measurements of end-feel or of accessory range of motion. Overall, methodological quality of included studies was poor, although two high-quality studies reported almost perfect reliability (Glasgow et al 2003, Nomden et al 2009).

In general, reliability for measurements of passive movements of upper extremity joints were substantially higher than for measurements of passive segmental intervertebral and sacroiliac joints which rarely exceed Kappa 0.40 (Van Trijffel et al 2005, Van der Wurff et al 2000). Seffinger et al (2004) attributed these differences in reliability to differences in size of joints. We think, however, that differences may be more linked to a joint's potential physiological range of motion. For instance, measurement of large joints with limited range such as the sacroiliac joint is associated with poor reliability, whereas measurement of small joints with greater range, such as the atlantoaxial spinal segment and finger joints, has been shown to be reliable (Cleland et al 2006, Glasgow et al 2003, Ogince et al 2007, Van der Wurff et al 2000). We also found that measuring large physiological ranges of motion, like that in the shoulder and in the wrist, frequently yielded satisfactory levels of reliability and note that these levels were predominantly as a result of using goniometers or inclinometers. In addition, findings from four studies (Chesworth et al 1998, Hayes and Petersen 2001, Patla and Paris 1993, Van Duijn and Jensen 2001) indicated that measuring end-feel or accessory movements of joints with large ranges of motion was associated with lower reliability. Staes et al (2009), on the other hand, reported better reliability for end-feel assessment of accessory intercarpal motion as compared to mobility classifications. With respect to spinal movement, Haneline et al (2008) similarly found somewhat higher reliability for measurement of end-feel. We hypothesise that measuring physiological movement for joints with large ranges of motion using goniometers

Table 3. Inter-rater reliability (95% CI) for measurement of passive physiological range of motion by method of measurement, joint and movement direction.

Method of measurement	Study	Inter-rater reliability
Inclinometer		
Shoulder		
External rotation	Awan et al (2002) De Winter et al (2004) Heemskerk et al (1997)	ICC 0.41, 0.51 ICC 0.90 ICC 0.81 to 0.87
Internal rotation	Awan et al (2002)	ICC 0.50 to 0.66
Abduction	De Winter et al (2004) Heemskerk et al (1997)	ICC 0.83 ICC 0.27 to 0.84
Horizontal flexion	Lin & Yang (2006)	ICC 0.82 (0.54 to 0.94)
Horizontal extension	Lin & Yang (2006)	ICC 0.89 (0.69 to 0.96)
Goniometer		
Shoulder		
External rotation	MacDermid et al (1999) Riddle et al (1987)	ICC 0.85 (0.73 to 0.91), 0.86 (0.72 to 0.92) ICC 0.88, 0.90
Internal rotation	Riddle et al (1987)	ICC 0.43, 0.55
Abduction	Riddle et al (1987)	ICC 0.84, 0.87
Horizontal abduction	Riddle et al (1987)	ICC 0.28, 0.30
Horizontal adduction	Riddle et al (1987)	ICC 0.35, 0.41
Flexion	Riddle et al (1987)	ICC 0.87, 0.89
Extension	Riddle et al (1987)	ICC 0.26, 0.27
Elbow		
Flexion	Rothstein et al (1983)	ICC 0.85 to 0.97
Extension	Rothstein et al (1983)	ICC 0.92 to 0.95
Wrist-hand-fingers		
Wrist Flexion	Horger (1990) LaStayo & Wheeler (1994)	ICC 0.86 (0.78 lower limit) ICC 0.88 to 0.93
Wrist Extension	Horger (1990) LaStayo & Wheeler (1994)	ICC 0.84 (0.75 lower limit) ICC 0.80 to 0.84
Wrist Abduction	Horger (1990)	ICC 0.66 (0.51 lower limit)
Wrist Adduction	Horger (1990)	ICC 0.83 (0.74 lower limit)
Thumb abduction	De Kraker et al (2009)	ICC 0.37 (-0.42 to 0.79)
Finger/thumb flexion and extension	Glasgow et al (2003)	ICC 0.99 (0.98 to 1.0)
Vision		
Shoulder		
External rotation	Chesworth et al (1998) Hayes et al (2001) Nomden et al (2009) Terwee et al (2005)	ICC 0.83 (0.70 to 0.90), 0.90 (0.83 to 0.95) ICC 0.57 (0.26 to 0.87) ICC 0.70 ICC 0.73 (0.22 to 0.88)
Internal rotation	Awan et al (2002) Hayes et al (2001)	ICC 0.51, 0.65 ICC 0.26 (-0.01 to 0.69)
Abduction	Hayes et al (2001) Nomden et al (2009) Terwee et al (2005)	ICC 0.66 (0.37 to 0.90) ICC 0.96 ICC 0.67 (0.35 to 0.81)
Horizontal adduction	Terwee et al (2005)	ICC 0.36 (0.22 to 0.48)
Flexion	Hayes et al (2001)	ICC 0.70 (0.42 to 0.92)
Elevation	Terwee et al (2005)	ICC 0.87 (0.83 to 0.90)
Wrist-hand-fingers		
Wrist flexion	Bovens et al (1990)	r 0.59
Wrist extension	Bovens et al (1990)	r 0.09
Tape measure		
Shoulder		
External rotation	Tyler et al (1999)	ICC 0.80
Pollexograph		
Wrist-hand-fingers		
Thumb abduction	De Kraker et al (2009)	ICC 0.59 (0.42 to 0.89)

Table 4. Inter-rater reliability (95% CI) for measurement of passive accessory range of motion by joint and movement direction.

Accessory motion	Study	Inter-rater reliability
Shoulder		
Inferior glide	Van Duijn & Jensen (2001)	ICC 0.52
Wrist-hand-fingers		
Wrist capitate	Staes et al (2009)	K _w 0.29 to 0.42, 0.33 to 0.87

or inclinometers, and measuring end-feel for joints with limited range of motion will lead to more reliable decisions about joint restrictions in clinical practice. Since few studies have investigated reliability of measurement of end-feel or accessory movements in upper extremity joints, future research should focus on the inter-rater reliability of these measures compared with measurements of physiological movements within the same sample of participants and raters.

In this review, we found studies investigating inter-rater reliability of upper extremity joint motion examination to have been poorly conducted. Only one study satisfied all external validity criteria and only two met all internal validity criteria. None of the included studies was both externally and internally valid. This finding is no different from that of reviews of reliability of measurements of spinal movement (Seffinger et al 2004, Van Trijffel et al 2005). The majority of the studies in our review met the criterion concerning blinding procedures. However, criteria about the stability of participants' and raters' characteristics during the study were often either unmet or unknown. Instability of the participants' characteristics under investigation, in this case joint range of motion or end-feel, may be caused by changes in the biomechanical properties of connective tissues as a result of natural variation over time or the effect of the measurement procedure itself (Rothstein and Echternach 1993). Similarly, instability of the raters, in this case their consistency in making judgments, may be caused by mental fatigue. Instability of raters' or participants' characteristics can lead to underestimations of reliability, whereas a lack of appropriate blinding of raters can lead to overestimation. In the presence of all of these

methodological flaws, direction of risk of bias is difficult to predict. Factors about internal validity are closely linked to issues of generalisation of results. For instance, performing several measurements on a large number of participants in a limited time period is not only susceptible to bias but also does not reflect clinical practice. Reliability of measurements varies across populations of participants and raters (Streiner & Norman 2008). In order to better reflect clinical practice, it is preferable to measure participants who would normally have their passive movements measured as part of the physiotherapy assessment, ie, consecutive patients with musculoskeletal conditions rather than healthy volunteers, as well as allowing raters access to information from the history and physical examination (Whiting et al 2003). However, we had decided *a priori* to include studies of asymptomatic individuals because of the information on reliability they may provide. Seven of our included studies used healthy volunteers as participants.

We note that the majority of included studies calculated ICC for expressing reliability of measurement of range of motion between raters. ICC are the most appropriate parameter of reliability for continuous data reflecting the ability of raters to discriminate between individuals (De Vet et al 2006). For effect of intervention, however, insight into absolute measurement error is required and other parameters, such as the limits of agreement, are preferable for expressing agreement within raters on measurements across multiple occasions over time (Bland and Altman 1986, De Vet et al 2006). To date, such data with respect to measurement of passive movements of upper extremity joints are rarely available. Since reliable measures of passive movement do not necessarily also have low absolute measurement errors, they cannot necessarily be used to evaluate the effect of intervention.

Finally, with regard to physiological range of motion in the shoulder, we found large variation in reliability of measurement of external rotation and abduction range. Cyriax (1982) first described patterns of joint restrictions to distinguish between capsular and other causes, eg, external rotation being most limited followed by abduction followed by internal rotation indicates a capsular cause. This pattern, however, was not corroborated in patients with idiopathic loss of shoulder range of motion (Rundquist and Ludewig 2004). In addition, almost complete loss of external rotation is the pathognomonic sign of frozen shoulder (Dias et al 2005). Valid diagnosis of shoulder disorders based on pattern of passive external rotation and abduction loss of range requires further research.

Table 5. Inter-rater reliability (95% CI) for measurement of physiological end-feel by joint and movement direction.

End-feel	Study	Inter-rater reliability
Shoulder		
External rotation	Chesworth et al (1998)	ICC 0.34 (0.05 to 0.57) to 0.91 (0.84 to 0.95)
	Hayes & Petersen (2001)	K 0.47 (0.08 to 0.87)
Internal rotation	Hayes & Petersen (2001)	K 0.41 (0.03 to 0.80)
Abduction	Hayes & Petersen (2001)	K 0.70 (0.31 to 1.0)
Horizontal adduction	Hayes & Petersen (2001)	K 0.40 (0.01 to 0.79)
Full abduction	Hayes & Petersen (2001)	K 0.26 (-0.16 to 0.68)
Elbow		
Flexion	Patla & Paris (1993)	K 0.40
Extension	Patla & Paris (1993)	K 0.73

Limitations of this review

This review has limitations with respect to its search strategy, quality assessment, and analysis. Only 11 included studies originated from our electronic search. A reason for this low electronic yield may be the inconsistent terminology used in reliability research. In our experience, reliability studies were poorly indexed in databases. In addition, our search strategy may have been too specific. Although much effort was put into reference tracing and hand searching, it is possible that eligible studies were missed. Furthermore, unpublished studies were not included. Publication bias can form a real threat to internal validity of systematic reviews of reliability studies because they are more likely to report low reliability.

Additionally, quality assessment was performed by using criteria derived mainly from the quality assessment of diagnostic accuracy studies. No evidence is available on whether these items can be applied to reliability studies. Empirical evidence of bias, especially concerning blinding of raters and stability of characteristics of participants and raters, is lacking. Another method for scoring methodological quality may have resulted in different conclusions.

Finally, our analysis was based on point estimates of reliability. Including interpretation of the precision of these estimates would have provided a more detailed perspective. However, only a limited number of included studies presented 95% CI. In the majority of these cases, CI were quite wide suggesting low sample sizes. None of our included studies reported an *a priori* sample size calculation.

Recommendations

We conclude that inter-rater reliability of measurements of passive movements in upper extremity joints varies with the method of measurement. In order to make reliable decisions about joint restrictions in clinical practice, we recommend that clinicians measure passive physiological range of motion using goniometers or inclinometers. Future research should focus on comparing inter-rater reliability of end-feel and accessory movements with passive physiological range of motion assessment, using symptomatic individuals. In addition, more research is needed on the elbow and wrist joints. Careful consideration should be given to ensuring stability of participants' and raters' characteristics during the study and *a priori* sample sizes should be calculated. Following the STARD statement will also improve the quality of reporting of reliability studies (Bossuyt et al 2003a, Bossuyt et al 2003b). Finally, new intra-rater reliability studies determining the absolute measurement error (agreement) when measuring passive range of motion in upper extremity joints will provide insight into the amount of change in range needed to indicate an effect of intervention beyond this error. ■

eAddenda: Appendix 1, Appendix 2 available at JoP. physiotherapy.asn.au

Correspondence: Emiel van Trijffel, Department of Clinical Epidemiology, Biostatistics & Bioinformatics, University of Amsterdam, Academic Medical Centre, The Netherlands. Email: E.vanTrijffel@amc.uva.nl

References

- Awan R, Smith J, Boon AJ (2002) Measuring shoulder internal rotation range of motion: a comparison of 3 techniques. *Archives of Physical Medicine and Rehabilitation* 83: 1229–1234.
- Bartko JJ, Carpenter WT (1976) On the methods and theory of reliability. *The Journal of Nervous and Mental Disease* 163: 307–317.
- Bisset L, Paungmali A, Vicenzino B, Beller E (2005) A systematic review and meta-analysis of clinical trials on physical interventions for lateral epicondylalgia. *British Journal of Sports Medicine* 39: 411–422.
- Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurements. *Lancet* i: 307–310.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al (2003a) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. *Clinical Chemistry* 49: 1–6.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM et al (2003b) The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clinical Chemistry* 49: 7–18.
- Bovens AM, Van Baak MA, Vrencken JG, Wijnen JA, Verstappen FT (1990) Variability and reliability of joint measurements. *American Journal of Sports Medicine* 18: 58–63.
- Chesworth BM, MacDermid JC, Roth JH, Patterson SD (1998) Movement diagram and 'end-feel' reliability when measuring passive lateral rotation of the shoulder in patients with shoulder pathology. *Physical Therapy* 78: 593–601.
- Cleland JA, Childs JD, Fritz JM, Whitman JM (2006) Interrater reliability of the history and physical examination in patients with mechanical neck pain. *Archives of Physical Medicine and Rehabilitation* 87: 1388–1395.
- Cyriax J (1982) Textbook of orthopaedic medicine. Volume one: Diagnosis of soft tissue lesions (8th edn). London: Bailliere Tindall, Chapter 5.
- De Kraker M, Selles RW, Schreuders TAR, Stam HJ, Hovius SER (2009) Palmar abduction: Reliability of 6 measurement methods in healthy adults. *The Journal of Hand Surgery (Am)* 34: 523–530.
- De Vet HC, Terwee CB, Knol DL, Bouter LM (2006) When to use agreement versus reliability measures. *Journal of Clinical Epidemiology* 59: 1033–1039.
- De Winter AF, Heemskerk MA, Terwee CB, Jans MP, Devillé W, Van Schaardenburg DJ et al (2004) Inter-observer reproducibility of measurements of range of motion in patients with shoulder pain using a digital inclinometer. *BMC Musculoskeletal Disorders* 5: 18.
- Dias R, Cutts S, Massoud S (2005) Frozen shoulder. *BMJ* 331: 1453–1456.
- Glasgow C, Wilton J, Tooth L (2003) Optimal daily total end range time for contracture: resolution in hand splinting. *Journal of Hand Therapy* 16: 207–218.
- Haneline MT, Cooperstein R, Young M, Birkeland K (2008) Spinal motion palpation: a comparison of studies that assessed intersegmental end feel vs excursion. *Journal of Manipulative and Physiological Therapeutics* 31: 616–626.
- Hayes K, Walton JR, Szomor ZR, Murrell GA (2001) Reliability of five methods for assessing shoulder range of motion. *Australian Journal of Physiotherapy* 47: 289–294.
- Hayes KW, Petersen CM (2001) Reliability of assessing end-feel and pain and resistance sequence in subjects with painful shoulders and knees. *Journal of Orthopaedic and Sports Physical Therapy* 31: 432–445.
- Heemskerk MAMB, Van Aarst M, Van der Windt DAWM (1997) De reproduceerbaarheid van het meten van de passieve beweeglijkheid van de schouder met de EDI-320 digitale

- hoekmeter. *Dutch Journal of Physiotherapy* 107: 146–149. [In Dutch]
- Hengeveld E, Banks K (2005) Maitland's peripheral manipulation (4th edn). Oxford: Butterworth-Heinemann, Chapter 6.
- Hestbæk L, Leboeuf-Yde C (2000) Are chiropractic tests for the lumbo-pelvic spine reliable and valid? A systematic critical literature review. *Journal of Manipulative and Physiological Therapeutics* 23: 258–275.
- Ho CY, Sole G, Munn J (2009) The effectiveness of manual therapy in the management of musculoskeletal disorders of the shoulder: A systematic review. *Manual Therapy* 14: 463–474.
- Horger MM (1990) The reliability of goniometric measurements of active and passive wrist motions. *American Journal of Occupational Therapy* 44: 342–348.
- Kramer MS, Feinstein AR (1981) Clinical biostatistics LIV. The biostatistics of concordance. *Clinical Pharmacology and Therapeutics* 29: 111–123.
- Kromer TO, Tautenhahn UG, De Bie RA, Staal JB, Bastiaenen CHG (2009) Effects of physiotherapy in patients with shoulder impingement syndrome: a systematic review of the literature. *Journal of Rehabilitation Medicine* 41: 870–880.
- Landis JR, Koch DG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33: 159–164.
- LaStayo PC, Wheeler DL (1994) Reliability of passive wrist flexion and extension goniometric measurements: a multicenter study. *Physical Therapy* 74: 162–174.
- Lin JJ, Yang JL (2006) Reliability and validity of shoulder tightness measurement in patients with stiff shoulders. *Manual Therapy* 11: 146–152.
- MacDermid JC, Chesworth BM, Patterson S, Roth JH (1999) Intratester and intertester reliability of goniometric measurement of passive lateral shoulder rotation. *Journal of Hand Therapy* 12: 187–192.
- May S, Littlewood C, Bishop A (2006) Reliability of procedures used in the physical examination of non-specific low back pain: a systematic review. *Australian Journal of Physiotherapy* 52: 91–102.
- Nomden JG, Slagters AJ, Bergman GJ, Winters JC, Kropmans TJ, Dijkstra PU (2009) Interobserver reliability of physical examination of shoulder girdle. *Manual Therapy* 14: 152–159.
- O'Connor D, Marshall SC, Massy-Westropp N (2003) Non-surgical treatment (other than steroid injection) for carpal tunnel syndrome. *Cochrane Database of Systematic Reviews* Issue 1. Art. No.: CD003219. DOI: 10.1002/14651858.CD003219.
- Ogince M, Hall T, Robinson K, Blackmore AM (2007) The diagnostic validity of the cervical flexio-rotation test in C1/2-related cervicogenic headache. *Manual Therapy* 12: 256–262.
- Patla CE, Paris SV (1993) Reliability of interpretation of the Paris classification of normal end feel for elbow flexion and extension. *Journal of Manual and Manipulative Therapy* 1: 60–66.
- Riddle DL, Rothstein JM, Lamb RL (1987) Goniometric reliability in a clinical setting. Shoulder measurements. *Physical Therapy* 67: 668–673.
- Rothstein JM, Echtertnach JL (1993) Primer on measurement: an introductory guide to measurement issues. Alexandria, VA: American Physical Therapy Association, pp. 73–85.
- Rothstein JM, Miller PJ, Roettger RF (1983) Goniometric reliability in a clinical setting. Elbow and knee measurements. *Physical Therapy* 63: 1611–1615.
- Rundquist PJ, Ludewig PM (2004) Patterns of motion loss in subjects with idiopathic loss of shoulder range of motion. *Clinical Biomechanics* 19: 810–818.
- Seffinger MA, Najm WI, Mishra SI, Adams A, Dickerson VM, Murphy LS et al (2004) Reliability of spinal palpation for diagnosis of back and neck pain: a systematic review of the literature. *Spine* 29: E413–425.
- Staes FF, Banks KJ, De Smet L, Daniels KJ, Carels P (2009) Reliability of accessory motion testing at the carpal joints. *Manual Therapy* 14: 292–298.
- Stochkendahl MJ, Christensen HW, Hartvigsen J, Vach W, Haas M, Hestbaek L et al (2006) Manual examination of the spine: a systematic critical literature review of reproducibility. *Journal of Manipulative and Physiological Therapeutics* 29: 475–485.
- Streiner DL, Norman GR (2008). Health measurement scales. A practical guide to their development and use (4th ed.) Oxford: Oxford University Press, Chapter 8.
- Terwee CB, De Winter AF, Scholten RJ, Jans MP, Devillé W, Van Schaardenburg D et al (2005) Interobserver reproducibility of the visual estimation of range of motion of the shoulder. *Archives of Physical Medicine and Rehabilitation* 86: 1356–1361.
- Tyler TF, Roy T, Nicholas SJ, Gleim GW (1999) Reliability and validity of a new method of measuring posterior shoulder tightness. *Journal of Orthopaedic and Sports Physical Therapy* 29: 262–269.
- Van der Wurff P, Hagmeijer RH, Meyne W (2000) Clinical tests of the sacroiliac joint. A systematic methodological review. Part 1: Reliability. *Manual Therapy* 5: 30–36.
- Van Duijn AJ, Jensen RH (2001) Reliability of inferior glide mobility testing of the glenohumeral joint. *Journal of Manual and Manipulative Therapy* 9: 109–114.
- Van Trijffel E, Anderegg Q, Bossuyt PM, Lucas C (2005) Inter-examiner reliability of passive assessment of intervertebral motion in the cervical and lumbar spine: a systematic review. *Manual Therapy* 10: 256–269.
- Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PM, Kleijnen J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 3: 25.