Inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low: a systematic review

Emiel van Trijffel[1], Rachel J van de Pol[2], Rob AB Oostendorp[3] and Cees Lucas[1]

[1]University of Amsterdam, Academic Medical Centre, [2]Private Practice for Physiotherapy, The Hague, [3]Radboud University Nijmegen Medical Centre The Netherlands

Question: What is the inter-rater reliability for measurements of passive physiological or accessory movements in lower extremity joints? **Design:** Systematic review of studies of inter-rater reliability. **Participants:** Individuals with and without lower extremity disorders. **Outcome measures:** Range of motion and end-feel using methods feasible in daily practice. **Results:** 17 studies were included of which 5 demonstrated acceptable inter-rater reliability. Reliability of measurements of physiological range of motion ranged from Kappa −0.02 for measuring knee extension using a goniometer to ICC 0.97 for measuring knee flexion using vision. Measuring range of knee flexion consistently yielded acceptable reliability using either vision or instruments. Measurements of end-feel were unreliable for all hip and knee movements. Two studies satisfied all criteria for internal validity while reporting acceptable reliability for measuring physiological range of knee flexion and extension. Overall, however, methodological quality of included studies was poor. **Conclusion:** Inter-rater reliability of measurement of passive movements in lower extremity joints is generally low. We provide specific recommendations for the conduct and reporting of future research. Awaiting new evidence, clinicians should be cautious when relying on results from measurements of passive movements in joints for making decisions about patients with lower extremity disorders. **[van Trijffel E, van de Pol RJ, Oostendorp RAB, Lucas C (2010) Inter-rater reliability for measurement of passive physiological movements in lower extremity joints is generally low: a systematic review. *Journal of Physiotherapy* 56: 223–235]**

Key words: Lower extremity, Reliability, Reproducibility of results, Physical examination, Range of motion, Passive motion, Physiotherapy

Introduction

Physiotherapists commonly assess and treat patients with lower extremity joint disorders. Despite varying levels of evidence, a growing number of studies have shown that manual joint mobilisations or manipulations are effective in certain disorders such as hip and knee osteoarthritis, patellofemoral pain syndrome, ankle inversion sprain, plantar fasciitis, metatarsalgia, and hallux limitus/rigidus (Brantingham et al 2009). 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, Kaltenborn 2002). Passive physiological range of motion may be measured using vision or instruments such as goniometers or 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 measurements (De Vet et al 2006, Streiner and Norman 2008). High inter-rater reliability for measurements of lower extremity joints is a prerequisite for valid and uniform clinical decisions about joint restrictions and related disorders (Bartko and Carpenter 1976).

Several reviews have systematically summarised and appraised the evidence with respect to the inter-rater reliability of passive movements of human joints. Seven systematic reviews have been published on passive spinal and pelvic movement including segmental intervertebral motion assessment (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 low methodological quality. A recent systematic review showed better inter-rater reliability for measurements of passive physiological range of motion in upper extremity joints using instruments compared to measurements using vision and compared to measurements of end-feel or accessory range of motion (Van de Pol et al 2010). To date, no systematic appraisal of studies on inter-rater reliability of measurement of passive movements in lower 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 lower extremity joints?

Method

Identification and selection of studies

MEDLINE, EMBASE, and CINAHL were searched for studies published up to 1 March 2010. Search terms included all lower extremity joints and all synonyms for *reliability* and *rater* (see Appendix 1 on the eAddenda for the detailed search strategy for MEDLINE). The titles and abstracts were screened for eligibility by two reviewers (EvT, RJvdP) independently. When necessary, full text articles were retrieved. Reference lists of all

retrieved papers were hand searched for relevant studies. A supplemental hand search of 13 journals relevant to the field of physiotherapy from 1 January 2005 to 1 March 2010 (see Appendix 2 on the eAddenda for journals) was performed by one reviewer (EvT). Finally, four experts in lower extremity musculoskeletal research were approached to ask if they could provide any additional published studies. Additionally retrieved papers were checked for eligibility by a second reviewer (RJvdP).

Studies were included if they met all inclusion criteria (Box 1). No restrictions were imposed on language or date of publication. Studies were excluded if they were abstracts and documents that were anecdotal, speculative, or editorial in nature. Studies were also excluded if they investigated: active movement or restriction in passive movement due to pain or ligament instability; people with neurological conditions in which abnormal muscle tone may interfere with joint movement; people after arthroplasty; animals or cadavers. Study selection was performed by two reviewers (EvT, RJvdP) independently. Disagreements on eligibility were first resolved by discussion between the two reviewers and decided by a third reviewer (CL) if disagreement persisted.

Box 1. Inclusion criteria.

Design
- Repeated measures between raters

Participants
- Symptomatic and asymptomatic adults

Measurement procedure
- Performed passive (ie, manual) physiological or accessory movements in any of the joints of the hip, knee, or ankle–foot–toes
- Reported range of motion or end-feel
- Used methods feasible in daily 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, participant position, movement performed, method of measurement, outcomes reported), and inter-rater reliability (point estimates, estimates of precision). Two reviewers (EvT, RJvdP) extracted data independently and were not blind to journal, authors, or results. When disagreement between the two reviewers could not be resolved by discussion, a third reviewer (CL) made the final decision.

Quality: No validated instrument was 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 movement (Bossuyt et al 2003a, Bossuyt et al 2003b, Van Trijffel et al 2005, Whiting et al 2003). Criteria 1 to 4 assess external validity, Criteria 5 to 9 assess internal validity, and Criterion 10 assesses statistical methods (Box 2). Criteria were rated as 'yes', 'no', or 'unclear' where insufficient information was provided. 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 (EvT, RJvdP) independently assessed methodological quality of all included studies and were not blind to journal, authors, and results. If discrepancy between reviewers persisted, a decisive judgement was passed by a third reviewer (CL).

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 assessment procedure possible?
4. Was clinical information from participants available to raters and comparable to daily practice?
5. Were participants' characteristics under study stable during research?
6. Were raters' characteristics under study stable during research?
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?

Data analysis

Data were analysed by examining ICC and Kappa (95% CI). If at least 75% of a study's ICC or Kappa values were above 0.75, the study was considered to have shown acceptable 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 characteristics of the studies (participants' clinical characteristics, raters' profession and training, movement performed, method of measurement) and methodological quality. 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 rated 'unknown' because insufficient information was provided, no statement was made regarding the presence or direction of potential bias. Finally, clinical and methodological characteristics of included studies were examined for homogeneity in order to judge the possibility of statistically summarising results by calculating pooled estimates of reliability.

Results

Flow of studies through the review

Searching MEDLINE yielded 199 citations, of which 29 papers were retrieved in full text. After removing double citations, EMBASE (196 citations) provided another three potentially relevant studies. CINAHL (98 citations) then yielded no additional relevant articles. Hand searching of reference lists identified another 14 potentially eligible studies. Of these 46, 31 studies were excluded (see Appendix 3 on the eAddenda for excluded studies). Hand searching

of journals yielded one eligible study while one expert provided another. In total, 17 studies fulfilled all inclusion criteria (Figure 1).

Description of studies

The included studies are summarised in Table 1. Seven studies investigated inter-rater reliability of measurement of passive hip movements (Aalto et al 2005, Chevillotte et al 2009, Cibere et al 2008, Croft et al 1996, Currier et al 2007, Sutlive et al 2008, Van Gheluwe et al 2002), seven investigated knee movements (Cibere et al 2004, Cleffken et al 2007, Currier et al 2007, Fritz et al 1998, Hayes & Petersen 2001, Rothstein et al 1983, Watkins et al 1991), five investigated ankle movements (Diamond et al 1989, Elveru et al 1988, Erichsen et al 2006, Smith-Oricchio & Harris 1990, Van Gheluwe et al 2002), and one investigated first ray movements (Van Gheluwe et al 2002). In 11 studies physiotherapists acted as raters. There were no disagreements between reviewers on selection of studies.

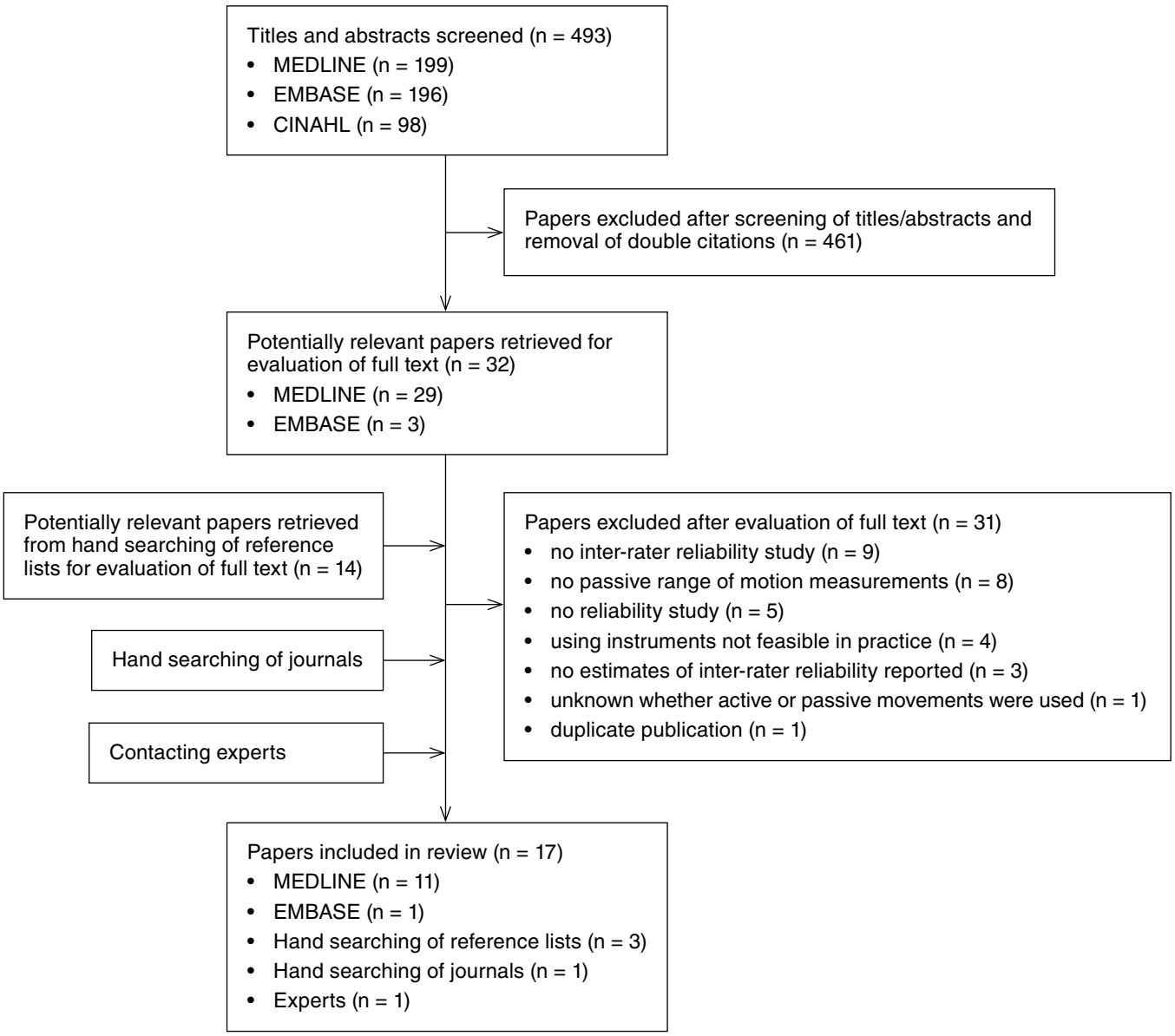

Figure 1. Flow of studies through the review.

Table 1. Summary of included studies (n = 17).

Study	Participants	Raters	Joints and movement directions	Position	Movement performed	Method	Outcome reported	Reliability statistic
Aalto et al (2005)	n = 20 Age = mean 23.3 yr (range 18–45) Condition = normal	n = 2 Profession = PT Training = ?	Hip • IR	Seated	Physiological	Goniometer	ROM	ICC
Chevillotte et al (2009)	n = 33 Age = mean 62.8 yr (SD 16.1) Condition = preoperative hip OA	n = 5 Profession = 2 hip surgeons, 2 orthopaedic surgery residents, 1 physician assistant Training = N	Hip • F • Abd • Add • IR • ER • E	Supine Supine Hip 90° F Lateral decubitus	Physiological	Vision	ROM	ICC
Cibere et al (2004)	n = 6 Age = median 62 yr (range 44–74) Condition = knee OA	n = 6 Profession = rheumatologist Training = Y	Knee • E	Unknown	Physiological	Goniometer	ROM	PABAK
Cibere et al (2008)	n = 6 Age = median 63 yr (range 49–65) Condition = hip OA	n = 6 Profession = 4 rheumatologists, 2 orthopaedic surgeons Training = Y	Hip • F • Abd • Add • IR • ER • IR • ER • E	Supine Supine Hip 90° F Seated Lateral decubitus	Physiological	Goniometer	ROM	R
Cleffken et al (2007)	n = 42 Age = mean 22.1 yr (range 19–27) Condition = normal	n = 2 Profession = ? Training = Y	Knee • F	Supine	Physiological	Inclinometer	ROM	Pearson's r
Croft et al (1996)	n = 6 Age = ? Condition = hip OA	n = 6 Profession = 5 general practitioner, 1 hospital physician Training = Y	Hip • F • IR • ER	Supine Seated	Physiological	Plurimeter	ROM	ICC

Study	Participants	Raters	Joints and movement directions	Position	Movement performed	Method	Outcome reported	Reliability statistic
Currier et al (2007)	n = 25 Age = ? Condition = knee OA	n = 2 Profession = PT doctoral student Training = Y	Hip • F • E • Abd • Add • Distraction • Patrick's test • IR • ER Knee • F • E	Supine Prone Knee 90° F Supine	Physiological	Goniometer Inclinometer	ROM End-feel	ICC (2,1) Kappa
Diamond et al (1989)	n = 31 Age = mean 59 yr (SD 12) Condition = diabetes mellitus	n = 2 Profession = PT Training = Y	Ankle • DF • INV • EV	Prone Knee 0° F	Physiological	Goniometer	ROM	ICC (2,1)
Elveru et al (1988)	n = 43 Age = mean 35.9 yr (SD 15.6) Condition = general orthopaedic disorders	n =14 Profession = PT Training = Y	Ankle • DF • PLF • INV • EV	Prone Knee 0° F	Physiological	Goniometer	ROM	ICC (1,1)
Erichsen et al (2006)	n = 27 Age = range 20–45 yr Condition = ankle pathology, normal	n = 2 Profession = PT Training = Y	Ankle • PLF • INV–EV • Med–lat talus glide	Supine	Physiological Accessory	Vision	ROM	Kappa
Fritz et al (1998)	n = 35 Age = ? Condition = knee dysfunction	n = 9 Profession = PT Training = N	Knee • F	Supine	Physiological	Vision	ROM	ICC (2,1)
Hayes & Petersen (2001)	n = 17 Age = mean 31.8 yr (SD 9.5) Condition = knee pain	n = 2 Profession = PT Training = Y	Knee • F • E	Supine	Physiological	Manual	End-feel	Kappa
Rothstein et al (1983)	n = 12 Age = ? Condition = knee pathology	n = 12 Profession = PT Training = ?	Knee • F • E	Unknown	Physiological	Goniometer	ROM	ICC

Table 1. Summary of included studies (n = 17) – continued

Study	Participants	Raters	Joints and movement directions	Position	Movement performed	Method	Outcome reported	Reliability statistic
Smith–Oricchio & Harris (1990)	n = 20 Age = range 18–53 yr Condition = ankle pathology	n = 3 Profession = PT Training = N	Ankle • INV • EV	Prone Knee 0° F	Physiological	Goniometer	ROM	ICC (3,1)
Sutlive et al (2008)	n = 30 Age = ? Condition = hip pain	n = ? Profession = PT doctoral student Training = Y	Hip • IR • ER • F • Scour test • Patrick's test	Prone Knee 90° F Supine	Physiological	Inclinometer	ROM End-feel	ICC (2,1) Kappa
Van Gheluwe et al (2002)	n = 30 Age = mean 24.8 yr Condition = normal	n = 5 Profession = podiatric physician Training = N	Hip • IR • ER Ankle • DF • INV • EV First ray • DF • PLF	Unknown	Physiological	Goniometer	ROM	Two-way ICC
Watkins et al (1991)	n = 43 Age = mean 39.5 yr (SD 15) Condition = knee pathology	n = 14 Profession = PT Training = N	Knee • F • E	Unknown	Physiological	Vision Goniometer	ROM	ICC (1,1)

Abd = abduction, Add = adduction, DF = dorsiflexion, EV = eversion, ER = external rotation, E = extension, F = flexion, IR = internal rotation, INV = inversion, OA = osteoarthritis, PLF = plantar flexion, PABAK = prevalence–adjusted bias–adjusted Kappa, PT = physiotherapist, ROM = range of motion, THA = total hip arthroplasty

Table 2. Methodological quality scores of included studies (n = 17).

Study	External validity				Internal validity					Statistical methods
	1	2	3	4	5	6	7	8	9	10
Aalto et al (2005)	N	U	Y	N	N	U	U	Y	Y	U
Chevillotte et al (2009)	Y	U	N	U	N	N	U	Y	N	U
Cibere et al (2004)	Y	U	N	N	Y	Y	Y	Y	U	Y
Cibere et al (2008)	Y	Y	Y	U	U	Y	Y	Y	U	Y
Cleffken et al (2007)	N	U	Y	U	Y	U	U	Y	N	N
Croft et al (1996)	U	U	Y	U	U	Y	U	Y	U	U
Currier et al (2007)	N	N	N	N	U	U	Y	Y	U	Y
Diamond et al (1989)	Y	U	Y	U	Y	N	Y	Y	N	Y
Elveru et al (1988)	Y	Y	Y	U	Y	Y	U	Y	N	Y
Erichsen et al (2006)	Y	N	Y	U	U	N	Y	Y	N	Y
Fritz et al (1998)	Y	U	Y	U	Y	N	Y	Y	U	Y
Hayes & Petersen (2001)	N	Y	Y	Y	U	U	Y	Y	N	U
Rothstein et al (1983)	U	Y	U	N	U	U	Y	Y	N	U
Smith-Oricchio & Harris (1990)	Y	Y	Y	Y	Y	U	Y	Y	U	Y
Sutlive et al (2008)	N	N	Y	U	U	U	Y	Y	U	Y
Van Gheluwe et al (2002)	N	Y	Y	U	U	U	U	Y	N	Y
Watkins et al (1991)	Y	Y	N	Y	Y	Y	Y	Y	N	Y

N = No, U = unclear because insufficient information provided, Y = Yes

Quality of studies

The methodological quality of included studies is presented in Table 2. One study (Smith-Oricchio & Harris 1990) fulfilled all four criteria for external validity and four studies (Cibere et al 2008, Elveru et al 1988, Hayes and Petersen 2001, Watkins et al 1991) satisfied three criteria. Two studies (Cibere et al 2004, Watkins et al 1991) fulfilled all three criteria for internal validity representing a low risk of bias, while five studies (Cibere et al 2008, Diamond et al 1989, Elveru et al 1988, Fritz et al 1998, Smith-Oricchio and Harris 1990) satisfied two criteria. Items on external and internal validity could not be scored on 48/153 (31%) occasions because of insufficient reporting. On methodological quality scores, 12/170 (7%) 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 and for physiological end-feel in Table 4. Because of clinical and methodological heterogeneity between studies, we did not attempt to calculate pooled estimates of reliability.

Hip (n = 7): None of the studies fulfilled all criteria for external or internal validity. In two studies (Aalto et al 2005, Cibere et al 2008), acceptable reliability was reached. Inter-rater reliability (ICC) of measurements of passive physiological range of motion ranged from 0.12 (95% CI 0.00 to 0.35), for surgeons and a physician assistant using vision to measure extension in preoperative patients with hip osteoarthritis (Chevillotte et al 2009), to 0.91, for physiotherapists using a goniometer to measure internal rotation in non-symptomatic participants (Aalto et al 2005). Chevillotte and colleagues (2009) found unacceptable reliability for measurements of all physiological hip movements. However, their estimates could have been underestimated due to instability of characteristics of participants as well as of raters. Cibere et al (2008) found acceptable reliability for measuring range of flexion, abduction, and internal rotation using a goniometer by trained rheumatologists and orthopaedic surgeons in patients with hip osteoarthritis. No specific movement direction or method of measurement was consistently associated with high or low reliability. Inter-rater reliability (Kappa) of measurements of physiological end-feel ranged from poor (–0.13, 95% CI –0.48 to 0.22) for extension (Currier et al 2007) to moderate (0.52, 95% CI 0.08 to 0.96) for the Scour test (Sutlive et al 2008). Both studies investigating reliability

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

Method of measurement	Study	Inter-rater reliability
Goniometer		
Hip		
Flexion	Cibere et al (2008)	R = 0.91, 0.91
Extension	Cibere et al (2008)	R = 0.66
Internal rotation	Aalto et al (2005)	ICC = 0.75 to 0.91
	Cibere et al (2008)	R = 0.87 to 0.95
	Van Gheluwe et al (2002)	ICC = 0.41 (lower limit 0.26) to 0.51 (lower limit 0.35)
External rotation	Cibere et al (2008)	R = 0.55 to 0.87
	Van Gheluwe et al (2002)	ICC = 0.35 (lower limit 0.20) to 0.37 (lower limit 0.21)
Abduction	Cibere et al (2008)	R = 0.88, 0.91
	Currier et al (2007)	ICC = 0.54 (0.19 to 0.76)
Adduction	Cibere et al (2008)	R = 0.56, 0.72
	Currier et al (2007)	ICC = 0.37 (–0.03 to 0.67)
Knee		
Flexion	Currier et al (2007)	ICC = 0.87 (0.73 to 0.94)
	Rothstein et al (1983)	ICC = 0.84 to 0.93
	Watkins et al (1991)	ICC = 0.90
Extension	Cibere et al (2004)	PABAK = –0.02, 0.88
	Currier et al (2007)	ICC = 0.69 (0.41 to 0.85)
	Rothstein et al (1983)	ICC = 0.59 to 0.80
	Watkins et al (1991)	ICC = 0.86
Ankle		
Dorsiflexion	Diamond et al (1989)	ICC = 0.74, 0.87
	Elveru et al (1988)	ICC = 0.00
	Van Gheluwe et al (2002)	ICC = 0.26 (lower limit 0.12), 0.31 (lower limit 0.17)
Plantar flexion	Elveru et al (1988)	ICC = 0.74
Inversion	Diamond et al (1989)	ICC = 0.86, 0.88
	Elveru et al (1988)	ICC = 0.30
	Smith-Oricchio & Harris (1990)	ICC = 0.42
	Van Gheluwe et al (2002)	ICC = 0.28 (lower limit 0.14), 0.40 (lower limit 0.22)
Eversion	Diamond et al (1989)	ICC = 0.78, 0.79
	Elveru et al (1988)	ICC = 0.22
	Smith-Oricchio & Harris (1990)	ICC = 0.25
	Van Gheluwe et al (2002)	ICC = 0.46 (lower limit 0.30), 0.49 (lower limit 0.32)
First ray		
Dorsiflexion	Van Gheluwe et al (2002)	ICC = 0.14 (lower limit 0.04), 0.16 (lower limit 0.06)
Plantar flexion	Van Gheluwe et al (2002)	ICC = 0.19 (lower limit 0.07), 0.21 (lower limit 0.09)
Vision		
Hip		
Flexion	Chevillotte et al (2009)	ICC = 0.56 (0.37 to 0.75)
Extension	Chevillotte et al (2009)	ICC = 0.12 (0.00 to 0.35)
Internal rotation	Chevillotte et al (2009)	ICC = 0.50 (0.30 to 0.70)
External rotation	Chevillotte et al (2009)	ICC = 0.37 (0.19 to 0.60)
Abduction	Chevillotte et al (2009)	ICC = 0.49 (0.29 to 0.70)
Adduction	Chevillotte et al (2009)	ICC = 0.39 (0.20 to 0.62)

Table 3 – continued

Method of measurement	Study	Inter-rater reliability
Knee		
Flexion	Fritz et al (1998)	ICC = 0.97
	Watkins et al (1991)	ICC = 0.83
Extension	Watkins et al (1991)	ICC = 0.82
Ankle		
Plantar flexion	Erichsen et al (2006)	K = 0.20 (–0.22 to 0.63), 0.47 (0.13 to 0.81)
Inversion–eversion	Erichsen et al (2006)	K = 0.37 (–0.03 to 0.77), 0.37 (–0.03 to 0.77)
Inclinometer		
Hip		
Flexion	Currier et al (2007)	ICC = 0.56 (0.21 to 0.78)
Extension	Currier et al (2007)	ICC = 0.20 (–0.22 to 0.55)
Internal rotation	Currier et al (2007)	ICC = 0.76 (0.53 to 0.89)
	Sutlive et al (2008)	ICC = 0.88 (0.74 to 0.94)
External rotation	Currier et al (2007)	ICC = 0.29 (–0.12 to 0.62)
	Sutlive et al (2008)	ICC = 0.77 (0.53 to 0.89)
Patrick's test	Currier et al (2007)	ICC = 0.57 (0.23 to 0.79)
Knee		
Flexion	Cleffken et al (2007)	Pearson's r = 0.83 to 0.87
Plurimeter		
Hip		
Flexion	Croft et al (1996)	ICC = 0.87
Internal rotation	Croft et al (1996)	ICC = 0.48
External rotation	Croft et al (1996)	ICC = 0.43

Table 4. Inter-rater reliability (95% CI) for measurement of passive physiological end-feel by joint and movement direction.

End-feel	Study	Inter-rater reliability Kappa (95% CI)
Hip		
Flexion	Currier et al (2007)	0.41 (0.14 to 0.68)
	Sutlive et al (2008)	0.21 (–0.22 to 0.64)
Extension	Currier et al (2007)	–0.13 (–0.48 to 0.22)
Internal rotation	Currier et al (2007)	0.20 (–0.07 to 0.47)
	Sutlive et al (2008)	0.51 (0.19 to 0.83)
External rotation	Currier et al (2007)	–0.02 (–0.37 to 0.33)
Abduction	Currier et al (2007)	0.15 (–0.14 to 0.44)
Adduction	Currier et al (2007)	0.00 (–0.39 to 0.39)
Patrick's test	Currier et al (2007)	0.39 (0.12 to 0.66)
	Sutlive et al (2008)	0.47 (0.12 to 0.81)
Distraction	Currier et al (2007)	0.13 (–0.24 to 0.50)
Scour test	Sutlive et al (2008)	0.52 (0.08 to 0.96)
Knee		
Flexion	Currier et al (2007)	0.31 (–0.53 to 1.00)
	Hayes & Petersen (2001)	–0.01 (–0.36 to 0.35)
Extension	Currier et al (2007)	0.25 (–0.18 to 0.68)
	Hayes & Petersen (2001)	0.43 (–0.06 to 0.92)

of end-feel measurements used symptomatic participants (Currier et al 2007, Sutlive et al 2008).

Knee (n = 7): Two studies (Cibere et al 2004, Watkins et al 1991) fulfilled all criteria for internal validity. Cibere et al (2004) demonstrated almost perfect inter-rater reliability (Kappa 0.88) for rheumatologists using a goniometer to measure passive physiological range of extension in patients with knee osteoarthritis. Watkins and colleagues (1991) reported acceptable reliability for physiotherapists using either vision of a goniometer to measure physiological range of flexion and extension in symptomatic participants. In the study by Fritz and colleagues (1998), acceptable reliability was also reached. Inter-rater reliability of measurements of passive physiological range of motion ranged from Kappa –0.02 for measuring extension before standardisation training (Cibere et al 2004) to ICC 0.97 for physiotherapists using vision to measure flexion in symptomatic participants (Fritz et al 1998). Measuring physiological range of flexion in supine with the hip in 90 deg flexion consistently yielded acceptable reliability regardless of the method of measurement. Inter-rater reliability (Kappa) of measurements of physiological end-feel ranged from poor (–0.01, 95% CI –0.36 to 0.35) for flexion to moderate (0.43, 95% CI –0.06 to 0.92) for extension (Hayes & Petersen 2001). Both studies investigating reliability of end-feel measurements used symptomatic participants (Currier et al 2007, Hayes and Petersen 2001).

Ankle-foot-toes (n = 5): One study (Smith-Oricchio and Harris 1990) fulfilled all criteria for external validity. In this study, unacceptable inter-rater reliability was demonstrated by physiotherapists using a goniometer to measure passive physiological range of ankle inversion (ICC 0.42) and eversion (ICC 0.25) in symptomatic participants. In the study by Diamond and colleagues (1989), acceptable estimates of reliability were reached for measurements of physiological range of ankle dorsiflexion, inversion, and eversion in diabetic patients by well-trained physiotherapists using a goniometer. These estimates could have been underestimated due to instability of characteristics of raters. Inter-rater reliability (ICC) of measurements of passive physiological range of motion ranged from 0.00 for measuring ankle dorsiflexion in patients with orthopaedic disorders by trained physiotherapists using a goniometer (Elveru et al 1988) to 0.88 for measuring ankle inversion (Diamond et al 1989). Inter-rater reliability of measurements of physiological range of motion of the first ray in non-symptomatic participants by podiatric physicians using a goniometer was unacceptable (Van Gheluwe et al 2002). Finally, the only study in this review investigating accessory range of motion showed fair (Kappa 0.35) to moderate (Kappa 0.48) inter-rater reliability for measurements of medio-lateral talar motion by physiotherapists in symptomatic participants (Erichsen et al 2006).

Discussion

This systematic review included 17 studies investigating inter-rater reliability of passive movements in lower extremity joints. Five studies demonstrated acceptable reliability. In four of these, physiotherapists acted as raters. Reliability of measurements of physiological range of motion ranged from Kappa –0.02 for rheumatologists using a goniometer to measure knee extension in patients with knee osteoarthritis, to ICC 0.97 for physiotherapists visually estimating knee flexion in symptomatic participants. Measuring physiological range of knee flexion consistently yielded acceptable reliability using either vision or instruments. Measurements of end-feel were unreliable for all hip and knee movements. Two high-quality studies (Cibere et al 2004, Watkins et al 1991) reported acceptable reliability for measuring physiological range of knee flexion and extension. Overall, however, methodological quality of the included studies was poor.

Inter-rater reliability for measurement of passive physiological range of motion in lower extremity joints was, overall, considerably less than that in upper extremity joints (Van de Pol et al 2010). In upper extremity joints, measuring large physiological ranges of motion like those in the shoulder, wrist, or fingers using instruments frequently yielded satisfactory reliability (Van de Pol et al 2010). This finding could only partly be confirmed for the lower extremity. For instance, measurement of physiological knee flexion using either vision or instruments indeed showed acceptable reliability, but measurements of relatively smaller ankle movements were unreliable in four out of five studies. However, inter-rater reliability for hip measurements varied widely across movements and methods of measurement. This heterogeneity in reliability could be explained by the large variation among studies in operational definitions of measurement procedures particularly with respect to participant positioning and instruction, and raters' execution of movements and handling of instruments. New research investigating inter-rater reliability for measurement of passive physiological hip movements should incorporate measurement procedures that are in accordance with international standards such as described by Clarkson (2005).

Based on the evidence of three studies (Currier et al 2007, Hayes and Petersen 2001, Sutlive et al 2008), we concluded that measurements of end-feel were unreliable for all hip and knee movements. This conclusion is similar to findings for other regions such as the shoulder and the elbow (Van de Pol et al 2010) and the spinal joints (Haneline et al 2008, Van Trijffel et al 2005). Cyriax (1982) originally described the concept of end-feel as the different sensations imparted to the hand of the rater at the extreme of the possible range of joint motion and he believed these were of great diagnostic relevance. This concept has then since long been incorporated in the various international approaches in manual therapy and subsequent educational programs (Farrell and Jensen 1992). As a consequence, manual therapists frequently use end-feel as an important indicator of spinal and extremity joint dysfunction (Abbott et al 2009, Van Ravensberg et al 2005, Van Trijffel et al 2009). The frequency of using end-feel measurements by physiotherapists for diagnosing lower extremity disorders is unknown but assumed to be high. Studies addressing the intra- and inter-rater reliability of end-feel measurements for diagnosing extremity disorders are needed, with clear and uniform criteria for classifying end-feel.

Only one of the included studies (Smith-Oricchio and Harris 1990) fulfilled all criteria for external validity implying that its results are generalisable to clinical practice. In particular, the majority of studies did not describe sufficiently whether measurements of passive movements were performed with or without clinical information from participants available to raters. In accordance with guidelines for the methodological quality assessment of diagnostic accuracy studies (Whiting et al 2003), we rated Criterion 4 in our quality assessment

list (Box 2) as positive when this information would also be available in clinical practice. Presumably measurements of passive movements of lower extremity joints usually take place after taking a history and performing one or more physical test procedures such as inspection, palpation, resistance tests, provocation tests, or measurement of active movements. Interpretation of measurements of passive movements will then inevitably be influenced by the previously gathered data. This dependence of test results on other information will alter estimates of inter-rater reliability as opposed to the ones generated by blinded single-test research. In medical test reading, providing clinical information was shown to increase diagnostic accuracy, ie, sensitivity (Whiting et al 2004). Research into the inter-rater reliability of measurements of passive movements of the extremities should therefore closely resemble clinical practice. However, no data are available on how and when physiotherapists use measurements of passive movements in relation to other diagnostic procedures within their clinical reasoning and decision-making. Identifying the role and position of a test within a diagnostic strategy can help to design studies to evaluate the diagnostic value of tests (Bossuyt et al 2006). In diagnostic research, a stepwise evaluation of tests is increasingly proposed considering not only the test's technical reliability and accuracy but also its place in the clinical pathway and, eventually, its impact on patient outcomes (Van den Bruel et al 2007). Investigating the role and position of measurements of passive movements of the extremities within clinical pathways for diagnosing disorders forms an unexplored field of research in physiotherapy and could improve the external validity of future reliability studies.

With respect to internal validity, only two studies (Cibere et al 2004, Watkins et al 1991) satisfied all three criteria, suggesting unbiased estimates of inter-rater reliability. This disappointing finding is similar to those of reviews of measurements of upper extremity movements (Van de Pol et al 2010) and spinal movement (Seffinger et al 2004, Van Trijffel et al 2005). However, in many cases, these validity criteria could not be scored due to inadequate reporting of the study protocol. In these cases, it was not possible to provide any indication of the presence and/or direction of the risk of bias. The criteria related to the stability of test circumstances, for both participants and raters, indicate underestimation of reliability if they are not met. Instability of the participants' characteristics under study – in this case the joint's mobility – may be caused by changes in the biomechanical properties of joint connective tissues as a result of natural variation over time or mobilising effects of the assessment procedure itself (Rothstein and Echternach 1993). Similarly, instability of the raters' capability of making judgments may be the result of, for example, mental fatigue. A lack of appropriate blinding of raters, on the other hand, could lead to overestimation of reliability. If several of these methodological flaws are present, the direction of risk of bias is difficult to predict. Researchers should give careful consideration to ensuring stability of participants' and raters' characteristics during research and to provide detailed information on the study protocol by following the STARD statement (Bossuyt et al 2003a, Bossuyt et al 2003b). Similar recommendations for improving the reporting of reliability studies were made in the field of medical research (Gow et al 2008).

A lack of inter-rater reliability adversely affects the accuracy of diagnostic decisions and subsequent treatment selection (Quinn 1989). This is particularly problematic when effective treatments are available and certain patients run the risk of not receiving them due to error and variation in decision-making among therapists. For instance, hip osteoarthritis is usually defined according to the clinical criteria of the American College of Rheumatology which include criteria about restrictions of physiological range of hip flexion and internal rotation (Altman et al 1991). Hoeksma and colleagues (2004) found a beneficial effect of specific manual manipulations and mobilisations of the hip joint on pain, range of motion, and activities in patients with hip osteoarthritis. However, our review did not show acceptable inter-rater reliability for measuring physiological range of hip flexion and internal rotation. In clinical practice, error and variation in diagnostic classification of hip osteoarthritis may therefore be leaving many patients undertreated. Furthermore, Cyriax's (1982) capsular pattern of gross restriction of physiological passive range of hip flexion, abduction, internal rotation and slight restriction of extension for diagnosing hip osteoarthritis was not corroborated, making diagnosis based on measurement of passive movements invalid (Bijl et al 1998, Klässbo et al 2003). Finally, another example in which treatment selection relies on measurement of passive movements is related to the finding that in patients with acute ankle sprain, manual mobilisation or manipulation has an initial beneficial effect on range of ankle dorsiflexion (Van der Wees et al 2006). Only a reliable measurement of restricted ankle dorsiflexion allows a valid decision whether or not to manually intervene. However, measuring passive physiological range of ankle dorsiflexion using a goniometer did not show acceptable reliability. Physiotherapists should incorporate a wider range of findings from their clinical assessment into their decisions about patients with lower extremity disorders and not rely too strongly on results from measurements of passive movements in joints.

Limitations of this review

This review has limitations with respect to its study identification, quality assessment, and data analysis. In our experience, reliability studies were poorly indexed in databases. Although much effort was put in reference tracing and hand searching, eligible studies may have been missed. Furthermore, unpublished studies were not included. Publication bias can threaten the internal validity of systematic reviews of reliability studies because unpublished studies are more likely to report low reliability. Quality assessment was performed by using a criteria list mainly derived from the assessment of diagnostic accuracy studies. It is not known whether these items also apply in the context of reliability. 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. We encourage further validation of the Quality Appraisal of Reliability Studies checklist (Lucas et al 2010). Also, study methods were frequently underreported in the included studies. We did not attempt to retrieve more information on study methods from the original authors. Complete information on these methods may have altered our conclusions with respect to study quality.

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 these cases, lower limits never indicated acceptable reliability and most CI were quite wide suggesting low sample sizes. None of the included studies reported an *a priori* sample size calculation.

Recommendations

We conclude that inter-rater reliability of measurement of passive physiological movements in lower extremity joints is generally low. Future research should focus on determining the role and position of measurements of passive movements in extremity joints within clinical reasoning and decision-making. In addition, the inter-rater reliability of measurements of passive physiological hip and ankle range of motion in particular and of measurements of end-feel should be further investigated. Careful consideration should be given to uniform standardisation of measurement procedures and to ensuring stability of participants' and raters' characteristics during research. Sample size calculations should be performed. Finally, following the STARD statement will also improve the quality of reporting of reliability studies (Bossuyt et al 2003a, Bossuyt et al 2003b). Awaiting new evidence, clinicians should be cautious about relying on results from measurements of passive movements in joints for making decisions about patients with lower extremity disorders. ■

eAddenda: Appendix 1, 2, and 3 available at www.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

Aalto TJ, Airaksinen O, Härkönen TM, Arokoski JP (2005) Effect of passive stretch on reproducibility of hip range of motion measurements. *Archives of Physical Medicine and Rehabilitation* 86: 549–557.

Abbott JH, Flynn TW, Fritz JM, Hing WA, Reid D, Whitman JM (2009) Manual physical assessment of spinal segmental motion: Intent and validity. *Manual Therapy* 14: 36–44.

Altman R, Alarcon G, Appelrouth D, Bloch D, Borenstein D, Brandt K (1991) The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hip. *Arthritis and Rheumatism* 34: 505–514.

Bartko JJ, Carpenter WT (1976) On the methods and theory of reliability. *The Journal of Nervous and Mental Disease* 163: 307–317.

Bijl D, Dekker J, Van Baar ME, Oostendorp RAB, Lemmens AM, Bijlsma JWJ et al (1998) Validity of Cyriax's concept of capsular pattern for the diagnosis of osteoarthritis of hip and/or knee. *Scandinavian Journal of Rheumatology* 27: 347–351.

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.

Bossuyt PM, Irwig L, Craig J, Glasziou P (2006) Comparative accuracy: assessing new tests against existing diagnostic pathways. *British Medical Journal* 332: 1089–1092.

Brantingham JW, Globe G, Pollard H, Hicks M, Korporaal C, Hoskins W (2009) Manipulative therapy for lower extremity conditions: expansion of literature review. *Journal of Manipulative and Physiological Therapeutics* 32: 53–71.

Chevillotte CJ, Ali MH, Trousdale RT, Pagnano MW (2009) Variability in hip range of motion on clinical examination. *The Journal of Arthroplasty* 24: 693–697.

Cibere J, Bellamy N, Thorne A, Esdaile JM, McGorm KJ, Chalmers A et al (2004) Reliability of the knee examination in osteoarthritis. *Arthritis and Rheumatism* 50: 458–468.

Cibere J, Thorne A, Bellamy N, Greidanus N, Chalmers A, Mahomed N et al (2008) Reliability of the hip examination in osteoarthritis: effect of standardization. *Arthritis and Rheumatism* 59: 373–381.

Clarkson HM (2005) Joint Motion and Function Assessment. A research-based practical guide (1st edn). Philadelphia: Lippincott Williams & Wilkins.

Cleffken B, Van Breukelen G, Brink P, Van Mameren H, Olde Damink S (2007) Digital goniometric measurement of knee joint motion. Evaluation of usefulness for research settings and clinical practice. *The Knee* 14: 385–389.

Croft PR, Nahit ES, Macfarlane GJ, Silma AJ (1996) Interobserver reliability in measuring flexion, internal rotation, and external rotation of the hip using a plurimeter. *Annals of the Rheumatic Diseases* 55: 320–323.

Currier LL, Froehlich PJ, Carow SD, McAndrew RK, Cliborne AV, Boyles RE et al (2007) Development of a clinical prediction rule to identify patients with knee pain and clinical evidence of knee osteoarthritis who demonstrate a favourable short–term response to hip mobilization. *Physical Therapy* 87: 1106–1119.

Cyriax J (1982) Textbook of Orthopaedic Medicine. Volume 1: Diagnosis of soft tissue lesions (8th edn). London: Bailliere Tindall.

De Vet HC, Terwee CB, Knol DL, Bouter LM (2006) When to use agreement versus reliability measures. *Journal of Clinical Epidemiology* 59: 1033–1039.

Diamond JE, Mueller MJ, Delitto A, Sinacore DR (1989) Reliability of a diabetic foot evaluation. *Physical Therapy* 69: 797–802.

Elveru RA, Rothstein JM, Lamb RL (1988) Goniometric reliability in a clinical setting. Subtalar and ankle joint measurements. *Physical Therapy* 68: 672–677.

Erichsen N, Lund H, Møller JO, Kaiser T, Jensen ML, Märcher I et al (2006) Inter-rater and intra-rater reliability of tests of translatoric movements and range of movements in the subtalar and talocrural joints. *Advances in Physiotherapy* 8: 161–167.

Farrell JP, Jensen GM (1992) Manual therapy: a critical assessment of role in the profession of physical therapy. *Physical Therapy* 72: 843–852.

Fritz JM, Delitto A, Erhard RE, Roman M (1998) An examination of the selective tissue tension scheme, with evidence for the concept of a capsular pattern of the knee. *Physical Therapy* 78: 1046–1061.

Gow RM, Barrowman NJ, Lai L, Moher D (2008) A review of five cardiology journals found that observer variability of measured variables was infrequently reported. *Journal of Clinical Epidemiology* 61: 394–401.

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 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.

Hengeveld E, Banks K (2005) Maitland's peripheral manipulation (4th edn). Philadelphia: Butterworth-Heinemann.

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.

Hoeksma HL, Dekker J, Ronday HK, Heering A, Van der Lubbe N, Vel C et al (2004) Comparison of manual therapy and exercise therapy in osteoarthritis of the hip: A randomized clinical trial. *Arthritis and Rheumatism* 51: 722–29.

Kaltenborn FM (2002) Manual mobilization of the joints. Volume I: The extremities (6th edn). Oslo: Olaf Norlis Bokhandel.

Klässbo M, Harms-Ringdahl K, Larsson G (2003) Examination of passive ROM and capsular patterns in the hip. *Physiotherapy Research International* 8: 1–12.

Kramer MS, Feinstein AR (1981) Clinical biostatistics LIV. The biostatistics of concordance. *Clinical Pharmacology and Therapeutics* 29: 111–123.

Landis JR, Koch DG (1977) The measurement of observer agreement for categorical data. *Biometrica* 33: 159–164.

Lucas NP, Macaskill P, Irwig L, Bogduk N (2010) The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *Journal of Clinical Epidemiology* 63: 854–861.

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.

Quinn MF (1989) Relation of observer agreement to accuracy according to a two-receiver signal detection model of diagnosis. *Medical Decision Making* 9: 196–206.

Rothstein JM, Echternach JL (1993) Primer on measurement: an introductory guide to measurement issues. Alexandria: American Physical Therapy Association.

Rothstein JM, Miller PJ, Roettger RF (1983) Goniometric reliability in a clinical setting. Elbow and knee measurements. *Physical Therapy* 63: 1611–1615.

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.

Smith-Oricchio K, Harris BA (1990) Interrater reliability of subtalar neutral, calcaneal inversion and eversion. *Journal of Orthopaedic and Sports Physical Therapy* 12: 10–15.

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.

Sutlive TG, Lopez HP, Schnitker DE, Yawn SE, Halle RJ, Mansfield LT et al (2008) Development of a clinical prediction rule diagnosing hip osteoarthritis in individuals with unilateral hip pain. *Journal of Orthopaedic and Sports Physical Therapy* 38: 542–550.

Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F (2007) The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost effectiveness is needed. *Journal of Clinical Epidemiology* 60: 1116–1122.

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.

Van der Wees PJ, Lenssen AF, Hendriks EJM, Stomp DJ, Dekker J, De Bie RA (2006) Effectiveness of exercise therapy and manual mobilisation in acute ankle sprain and functional instability: A systematic review. *Australian Journal of Physiotherapy* 52: 27–37.

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 Gheluwe B, Kirby KA, Roosen P, Phillips RD (2002) Reliability and accuracy of biomechanical measurements of the lower extremities. *Journal of the American Podiatric Medical Association* 92: 317–326.

Van Ravensberg CDD, Oostendorp RAB, Van Berkel LM, Scholten-Peeters GGM, Pool JJM, Swinkels RAHM et al (2005) Physical therapy and manual physical therapy: Differences in patient characteristics. *Journal of Manual and Manipulative Therapy* 13: 113–124.

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.

Van Trijffel E, Oostendorp RAB, Lindeboom R, Bossuyt PMM, Lucas C (2009) Perceptions and use of passive intervertebral motion assessment of the spine. A survey of Dutch physiotherapists specializing in manual therapy. *Manual Therapy* 14: 243–251.

Watkins MA, Riddle DL, Lamb RL, Personius WJ (1991) Reliability of goniometric measurements and visual estimates of knee range of motion obtained in a clinical setting. *Physical Therapy* 71: 90–97.

Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, 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.

Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J (2004) Sources of variation and bias in studies of diagnostic accuracy. A systematic review. *Annals of Internal Medicine* 140: 189–202.