Reliability of the straight leg raise test for suspected lumbar radicular pain

Nee, Robert J.; Coppieters, Michel W.; Boyd, Benjamin S.

published in
Musculoskeletal Science and Practice
2022

DOI (link to publisher)
10.1016/j.msksp.2022.102529

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

Musculoskeletal Science and Practice 59 (2022) 102529

Contents lists available at ScienceDirect

Musculoskeletal Science and Practice

journal homepage: www.elsevier.com/locate/msksp

Systematic review

Reliability of the straight leg raise test for suspected lumbar radicular pain: A systematic review with meta-analysis

Robert J. Nee[a,*], Michel W. Coppieters[b,c], Benjamin S. Boyd[a]

[a] Department of Physical Therapy, Samuel Merritt University, Oakland, CA, USA
[b] Menzies Health Institute Queensland, Griffith University, Brisbane & Gold Coast, Australia
[c] Amsterdam Movement Sciences, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Radiculopathy
Sciatica
Nerve root disorder
Neurodynamic test

ABSTRACT

Background: The passive straight leg raise (SLR) and crossed SLR are recommended tests for lumbar radicular pain. There are no recent reviews of test reliability.
Objectives: To summarize SLR and crossed SLR reliability in patients with suspected lumbar radicular pain.
Design: Systematic review with meta-analysis.
Method: MEDLINE and CINAHL were searched for studies published before April 2021 that reported SLR or crossed SLR reliability in patients with low back-related leg pain. Supplemental analyses also included patients with low back pain only. Study selection, risk of bias assessment (QAREL), and data extraction were performed in duplicate. Kappa, intraclass correlation coefficients, and smallest detectable difference (SDD_{95}) quantified reliability. Meta-analysis was performed when appropriate. Confidence in the evidence was determined by applying GRADE principles.
Results/findings: Fifteen studies met selection criteria. One-hundred-eighty-nine participants had low back-related leg pain. Four-hundred-thirty-nine were included in supplemental analyses. Meta-analyses showed at least fair inter-rater reliability when a positive SLR required provocation of lower extremity symptoms or pain. SLR reliability was at least moderate when testing included structural differentiation (e.g., ankle dorsiflexion). A low prevalence of positive crossed SLR tests led to wide-ranging reliability estimates. Confidence in the evidence for identifying a positive SLR or crossed SLR was moderate to very low. SDD_{95} values for different raters measuring SLR range of motion ranged from 13 to 20°.
Conclusions: Reliability data support testing SLR with structural differentiation manoeuvres. Crossed SLR reliability data are inconclusive. Measurement error likely prohibits using SLR range of motion for clinical decision-making.

1. Introduction

Clinical practice guidelines recommend the passive straight leg raise (SLR) test to help detect radicular pain in patients with low back pain (Oliveira et al., 2018). Furthermore, a crossed SLR may indicate radicular pain secondary to lumbar disc herniation (van der Windt et al., 2010; Stynes et al., 2018). Detecting lumbar radicular pain is important because it is typically associated with greater activity limitations and potentially poorer outcomes (Harrisson et al., 2017; Hartvigsen et al., 2017).

The SLR aims to detect radicular pain by mechanically provoking irritated lumbosacral nerve roots (Rebain et al., 2002). Biomechanical data support this premise (Gilbert et al., 2007; Rade et al., 2017). However, the SLR may also provoke symptoms related to irritation of non-neural tissues. Assessing effects structural differentiation manoeuvres (e.g., neck flexion, ankle dorsiflexion, or hip adduction) have on symptoms provoked in the SLR position potentially helps distinguish symptoms related to irritation of neural tissues from those related to irritation of non-neural tissues (Breig and Troup, 1979; Bueno-Gracia et al., 2019, 2020). Appropriate structural differentiation manoeuvres aim to further load or unload the nervous system without changing load on non-neural structures that could be sources of SLR-related symptoms. If one or more structural differentiation manoeuvres change SLR-provoked symptoms, those symptoms are thought to be at least

[*] Corresponding author.
E-mail address: robertjnee@gmail.com (R.J. Nee).

https://doi.org/10.1016/j.msksp.2022.102529
Received 25 August 2021; Received in revised form 25 January 2022; Accepted 8 February 2022
Available online 20 February 2022

partly related to neural tissue irritation (Breig and Troup, 1979; Troup, 1981). This interpretation assumes central pain mechanisms are not substantially contributing to the patient's pain experience (Smart et al., 2012a; b).

Even though guidelines recommend SLR testing, systematic reviews suggest the SLR performs poorly in diagnosing lumbar radicular pain (van der Windt et al., 2010; Scaia et al., 2012; Tawa et al., 2017; Mistry et al., 2020). Insufficient reliability may be one factor contributing to poor diagnostic performance (Sackett, 1992). Categorizing patients as having lumbar radicular pain based on SLR findings and other clinical data may also inform expectations about prognosis (Konstantinou et al., 2018), costs (Kigozi et al., 2019), and likely treatments (Delitto et al., 2012; George et al., 2021). Insufficient SLR reliability might contribute to inconsistent patient categorization and lead to inaccurate expectations about these aspects of management.

The most recent systematic review that summarized SLR reliability only included literature published between January 1989 and January 2000 (Rebain et al., 2002). Our systematic review aimed to provide an updated summary of SLR and crossed SLR reliability in patients with suspected lumbar radicular pain.

2. Methods

This systematic review is part of a larger protocol evaluating clinimetric properties of the SLR and crossed SLR in patients with lumbar radicular pain. The protocol was prospectively registered (PROSPERO CRD42018086158). This review followed the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines (Page et al., 2021).

2.1. Eligibility criteria

Eligible studies enrolled participants (aged >16 years) who presented with low back-related leg pain in any clinical setting. The SLR, with or without structural differentiation, and the crossed SLR could be performed in isolation or as part of a clinical examination of low back-related leg pain. Reliability could address identifying a positive SLR or crossed SLR or measuring SLR range of motion (ROM). Provocation of lower extremity symptoms or pain needed to be part of defining a positive SLR. Any type of study (e.g., reliability, diagnostic accuracy, or clinical trial) was eligible as long as reliability data were reported. Studies that focused on conditions often associated with neuropathies other than lumbar radicular pain (e.g., meningitis, polyneuropathy, diabetic neuropathy, HIV/AIDS, leprosy, alcohol dependence) or used orthoses to standardize lower limb positions during SLR were excluded. Studies needed to be available in English and disseminated in peer-reviewed publications.

2.2. Study identification and selection

MEDLINE (via PubMed) and CINAHL (via EBSCOhost) were searched for eligible studies published before April 2021 (Supplemental Tables). Grey literature was not searched. One of the reviewers (BSB or RJN) and a research assistant independently screened titles and abstracts for full text assessment. Two reviewers (BSB, RJN) independently screened full text articles for final inclusion in this review. Reference lists of included studies and previously published systematic reviews on SLR reliability and diagnostic accuracy were also searched independently by two reviewers (BSB, RJN). Disagreements during screening were resolved by consensus between the two reviewers. If consensus could not be reached, a third reviewer (MWC) was consulted for a final decision. Reasons for excluding studies were recorded.

In participants with low back-related leg pain, applicability of reliability data for identifying a positive SLR or crossed SLR was limited by most participants having severe symptoms that required hospitalization (Poiraudeau et al., 2001) or bed rest (Vroomen et al., 2000). Applicability of reliability data for measuring SLR ROM was limited by a small sample (Walsh and Hall, 2009a). Supplemental analyses were therefore performed on studies enrolling "mixed" samples of participants with low back-related leg pain or low back pain only.

2.3. Risk of bias assessment

Two reviewers (BSB, RJN) independently assessed risk of bias in included studies using the Quality Appraisal Tool for Studies of Diagnostic Reliability (QAREL) (Lucas et al., 2010). Reviewers agreed upon criteria for each QAREL item *a priori* to enhance inter-rater reliability (Lucas et al., 2013). To satisfy criteria for QAREL item 9 (Suitable time interval), repeated measures needed to occur within 24 h. For intervals greater than 24 h, data demonstrating participants' symptom status was similar at each measurement session needed to be reported. For QAREL item 10 (Appropriate test application/interpretation), studies investigating reliability of identifying a positive SLR needed to incorporate structural differentiation into testing. Disagreements were resolved by consensus between the two reviewers. If consensus could not be reached, a third reviewer (MWC) was consulted for a final decision.

Summarizing risk of bias with a total score from an appraisal tool is problematic because it omits details on specific limitations in each study that may influence results (Büttner et al., 2020). Limitations of included studies were therefore presented graphically by illustrating the proportion of studies that satisfied each QAREL item. Potential risks of bias for each SLR and crossed SLR reliability outcome (e.g., provocation of symptoms, measurement of ROM) were also reported.

2.4. Data extraction

Two reviewers (BSB, RJN) independently extracted data using a customized spreadsheet. Study setting, eligibility criteria, demographic characteristics of participants, SLR or crossed SLR test performance and interpretation, prevalence of a positive test, percent agreement, tools for measuring SLR ROM, and reliability outcomes (described below) with 95%CI were recorded. Disagreements were resolved by consensus between the two reviewers. If consensus could not be reached, a third reviewer (MWC) was consulted for a final decision. Authors were contacted as needed to clarify test performance or interpretation and obtain data to permit calculation of reliability outcomes or 95%CI.

2.5. Reliability outcomes

Kappa coefficients quantified the level of agreement for identifying a positive SLR or crossed SLR (Landis and Koch, 1977). Prevalence of a positive SLR or crossed SLR and percent agreement were reported when available to provide context for interpreting Kappa values. However, prevalence and bias indices were not calculated (Sim and Wright, 2005).

Relative reliability for consistency in measuring SLR ROM was quantified by intraclass correlation coefficients (ICCs) (Shrout and Fleiss, 1979). Absolute reliability (i.e., measurement error) for SLR ROM was quantified by the standard error of measurement (SEM) (Stratford, 2004) and smallest detectable difference at a 95% confidence level (SDD_{95}) (Eliasziw et al., 1994). SEM and SDD_{95} were calculated from reported data when not provided by study authors.

Reliability coefficients were interpreted as follows: 0.81 to 1.00 = substantial; 0.61 to 0.80 = moderate; 0.41 to 0.60 = fair; 0.11 to 0.40 = slight; and 0.00 to 0.10 = virtually none (Shrout, 1998).

2.6. Data analysis

Analyses focused on group data. Meta-analysis was performed when similar criteria were used to interpret SLR or crossed SLR responses in two or more similar samples. Random effects models using generic inverse variance were calculated with MedCalc Statistical Software version 19.8 (MedCalc Software Ltd, Ostend, Belgium; https://www.me

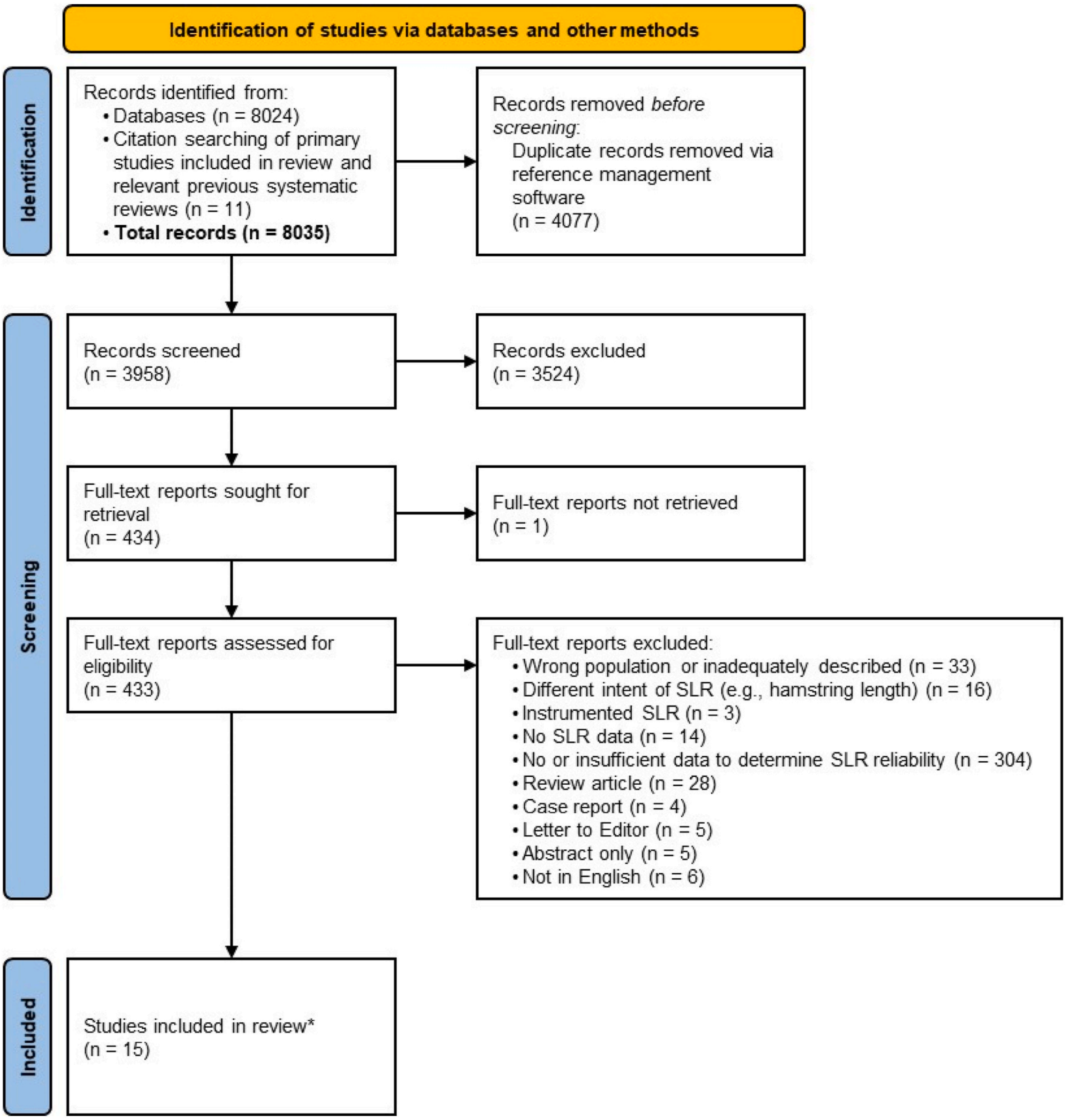

Identification of studies via databases and other methods
Identification
Records identified from:
• Databases (n = 8024)
• Citation searching of primary studies included in review and relevant previous systematic reviews (n = 11)
• Total records (n = 8035)
Records removed *before screening*:
Duplicate records removed via reference management software
(n = 4077)
Screening
Records screened
(n = 3958)
Records excluded
(n = 3524)
Full-text reports sought for retrieval
(n = 434)
Full-text reports not retrieved
(n = 1)
Full-text reports assessed for eligibility
(n = 433)
Full-text reports excluded:
• Wrong population or inadequately described (n = 33)
• Different intent of SLR (e.g., hamstring length) (n = 16)
• Instrumented SLR (n = 3)
• No SLR data (n = 14)
• No or insufficient data to determine SLR reliability (n = 304)
• Review article (n = 28)
• Case report (n = 4)
• Letter to Editor (n = 5)
• Abstract only (n = 5)
• Not in English (n = 6)
Included
Studies included in review*
(n = 15)

Fig. 1. Summary of study identification and selection process.

dcalc.org; 2021). Variance for Kappa values was calculated from percent agreement when reported (Sun, 2011). Otherwise, variance for Kappa and ICC values was calculated from reported 95%CI. Statistical heterogeneity was interpreted as low, moderate, and high when I^2 values were 25%, 50%, and 75%, respectively (Higgins et al., 2003).

2.7. Confidence in the evidence

Confidence in estimates of inter-rater reliability for identifying a positive SLR or crossed SLR was determined by applying Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) principles for diagnostic tests and strategies (Schünemann et al., 2020a, 2020b). Confidence was downgraded due to risk of bias, inconsistency, indirectness, or imprecision. Publication bias was not considered because of the small number of studies. Confidence was downgraded for risk of bias when more than 25% of participants providing data on a reliability outcome came from studies with important risks of bias. Confidence was downgraded for inconsistency when there were wide-ranging reliability estimates with minimal overlap in 95%CI or when I^2 estimates of statistical heterogeneity were greater than 50%. Confidence was downgraded for indirectness when data came from "mixed" samples that included participants with low back pain only. McHugh (2012) proposed Kappa values below 0.60 reflect insufficient agreement among raters. Confidence was therefore downgraded for imprecision when 95%CI for pooled estimates of Kappa spanned across this 0.60 threshold. A review information size was calculated to help assess imprecision when meta-analysis was not possible (Schünemann, 2016). When prevalence of a positive test is between 30% and 70%, 165 to 191 participants are needed to have 80% power ($p \leq 0.05$) to detect a Kappa value of 0.60 (fair reliability) that is significantly different from a null value of 0.40 (slight reliability) (Sim and Wright, 2005). In the absence of meta-analysis, confidence was downgraded for imprecision when reliability data came from less than 165 participants. Determinations of confidence in the evidence focused on definitions of a positive SLR or crossed SLR commonly used to detect lumbar radicular pain.

Table 1
Characteristics of included studies. “Mixed” samples involved inclusion of participants with low back-related leg pain or just low back pain.

A – Low back-related leg pain

Study	Setting, Sampling, Participants	Raters and Type of Reliability	Test(s) and Interpretation
Poiraudeau et al. (2001)	Hospital. Consecutive sample. Patients hospitalized with acute or chronic sciatica of mechanical origin. Pain below knee or pain above knee combined with “hard” neurological signs (e.g., DTR, strength, sensation). (n = 78) 58% female Mean age = 50 (SD 16)	R1: rheumatology trainee; 3 years of experience. R2: full-time physician; 10 years of experience. R3: full-time physician; 25 years of experience. Intra-rater and inter-rater reliability.	SLR to onset of pain. Positive SLR defined as provocation or exacerbation of patient’s specific sciatica symptoms. No structural differentiation. Excluded data on reliability of SLR ROM at onset of pain measured only in participants who had a positive SLR that provoked patient’s specific sciatica symptoms because of conflicting descriptions between methods and results on whether goniometric measurement or visual estimate of SLR ROM. Attempts to contact authors for clarification unsuccessful. Crossed SLR to onset of pain. Positive crossed SLR defined as provocation or exacerbation of patient’s specific sciatica symptoms.
Vroomen et al. (2000)	Setting unclear. Random sample of patients referred by 50 GPs. Patients with new episode of sciatica defined as pain referred into lower extremity. Symptoms of sufficient intensity to warrant 14 days bed rest as treatment option. (n = 91) 47% female Mean age = 46 (SD 11.2)	One neurologist and two neurologic residents created two neurologist/neurologic resident rater pairs. Experience level unclear. Inter-rater reliability. Data from each neurologist/neurologic resident rater pair pooled for overall Kappa estimates.	SLR to onset of pain. Separate analyses for different definitions of positive SLR: (1) provocation of typical dermatomal pain, (2) provocation of any pain in the limb, and (3) provocation of pain below 45 degrees. No structural differentiation. Bragard test. Lower limb five degrees below angle of onset of pain and add ankle DF. Positive test defined as provocation of pain with addition of DF. Crossed SLR. Positive test defined as provocation of back pain only, non-dermatomal limb pain, or dermatomal limb pain.
Walsh and Hall (2009a)*	Physiotherapy clinic. Consecutive sample of patients recruited from back pain screening clinic affiliated with hospital. Patients with low back-related unilateral leg pain. (n = 45 for overall sample) 51% female Mean age = 46 (SD 11) (n = 20 for reliability sample) First 20 participants recruited for overall sample. Demographics not described.	No clear description of raters.† Inter-rater reliability.	SLR to reproduction of patient’s presenting symptoms or until examiner perceived significant resistance to movement. SLR ROM measured with bubble inclinometer attached to limb just proximal to ankle. Rater blinded to inclinometer measurement that was read by independent observer.
Walsh and Hall (2009b)*	Same setting, sampling method, and participants as described above for Walsh and Hall (2000a).	Two physiotherapists. One with 3 months experience, one with 12 months experience. Inter-rater reliability.	SLR to reproduction of patient’s presenting symptoms or until examiner perceived significant resistance to movement. Positive SLR defined as provocation of patient’s specific limb symptoms that were increased with addition of ankle DF for structural differentiation.

B – “Mixed” samples of participants

Study	Setting, Sampling, Participants	Raters and Type of Reliability	Test(s) and Interpretation
Bertilson et al. (2006)	Private outpatient back clinic. Consecutive sample. Patients with low back pain ± radiation into the leg. (n = 50) 64% female Mean age females = 38 (range 18-61) Mean age males = 34 (range 16-61)	Two physiotherapists with > 20 years experience and international certification in orthopaedic manual therapy. Inter-rater reliability.	SLR with no specific description of end-point of test. Positive SLR defined as provocation of pain radiating below the knee that increased with addition of neck flexion or ankle DF for structural differentiation.
Billis et al. (2012)	Physiotherapy clinics. Convenience sample. Patients with non-specific low back pain. (n = 30) 60% female Mean age = 27.7 (SD 10.3)	Seven physiotherapists with mean clinical experience 11.8 years (range 7-19) in treating patients who had low back pain. Four were musculoskeletal specialists. Inter-rater reliability.	SLR focused between 30 and 70 degrees. Separate analyses for different definitions of positive SLR: (1) provocation of pain (no structural differentiation) and (2) reproduction of patient’s symptoms that were altered with addition of neck flexion, ankle DF/PF, hip ER/IR, or hip ABD/ADD for structural differentiation.

Boland and Adams (2000)	Private outpatient physiotherapy department. Sampling method unclear. Patients with unilateral lumbar spine pain ± ipsilateral leg pain. Group A (n = 20) 50% female Mean age = 50 (SD 18)	Group A: Two physiotherapists. One with four years of graduate experience, the other a manipulative physiotherapist with one year postgraduate experience. Inter-rater reliability.	SLR to onset of any pain in the back or leg. SLR ROM measured with a pendulum-type goniometer (pendulometer) strapped over the head of the fibula. DF+SLR to onset of any pain in the back or leg. DF+SLR ROM measured with a pendulum-type goniometer (pendulometer) strapped over the head of the fibula.
	Same setting and sampling method as described above for Boland and Adams (2000) Group A. Patients with unilateral lumbar spine pain ± ipsilateral leg pain. Group B (n = 15) 20% female Mean age = 33 (SD 9)	Group B: Two manipulative physiotherapists. One with four years postgraduate experience, one with seven years postgraduate experience. Inter-rater reliability.	Same as described above for Boland and Adams (2000) Group A.
Chow et al. (1994)	Setting unclear. Sampling method unclear. Patients recruited from one of three private practices plus two physical therapy students and one physical therapy teaching staff. Patients with low back pain ± leg pain that was provoked between 30 and 70 degrees on passive SLR. (n = 16) 31% female Mean age females = 32.6 (SD not reported) Mean age males = 44.7 (SD not reported)	One of the authors was the rater. All authors were physiotherapists. Experience not reported. Intra-rater reliability.	SLR to onset of pain or increase in resting pain. SLR ROM measured with bubble inclinometer attached 5cm proximal to inferior margin of lateral malleolus.
McCombe et al. (1989)	Research clinic. Sampling method unclear. Patients recruited from standard orthopaedic referral practice. Patients with low back pain. (n = 50) 48% female Mean age = 44.3 (SD 12.2)	Two orthopaedic surgeons. Experience level unclear. Inter-rater reliability.	SLR to pain onset and pain tolerance. Separate analyses for different definitions of positive SLR: (1) “sciatic stretch” (provocation of back and leg pain), (2) reproduction of symptoms, (3) provocation of leg pain, and (4) provocation of back pain. No structural differentiation. Unclear whether these analyses focused on response at pain onset or pain tolerance. Excluded data on reliability of SLR ROM at pain onset and pain tolerance because authors reported Pearson correlation coefficients, rather than intraclass correlation coefficients. Crossed SLR to pain onset and pain tolerance. Positive test defined as reproduction of symptoms.
	Special clinic within a hospital physiotherapy department. Random sample of patients referred from GPs or hospital specialists. Patients with low back pain. (n = 33) 21% female Mean age = 46.1 (SD 14.6)	Orthopaedic surgeon and physiotherapist. Experience level unclear. Inter-rater reliability	Same as described above for McCombe et al. (1989) orthopaedic surgeon rater pair.
Paatelma et al. (2010)	Private occupational healthcare center. Consecutive sample. Patients with low back pain lasting < 3 months. (n = 15) 73% female Mean age = 37.9 (SD 4.5)	Two physiotherapists specializing in orthopaedic manual therapy. One with 20 years experience (25 years total) and one with 2 years experience (8 years total) in orthopaedic manual therapy. Intra-rater and inter-rater reliability.	SLR to onset of buttock pain or more distal pain. No description of end point if these did not occur. Positive test defined as pain provoked in buttock or more distally. No structural differentiation.

Pesonen et al. (2021)	Institutional spine center. Consecutive sample. Patients selected by Study Controller to be part of sciatic or control group. Sciatic group (n = 20) = unilateral leg pain worse than back pain, clinical neurological deficits (strength, sensation, reflexes), and positive SLR. Control group (n = 20) = pain in low back and/or greater trochanter and/or hip region with or without posterior thigh tightness, no clinical neurological deficits, and negative SLR. (n = 40) 63% female Mean age 41 (range 22-64)	Two physiatry residents. Experience level unclear. Examiners unaware of whether participants were in sciatic or control group. Inter-rater reliability.	SLR to onset of symptoms (or 30% increase in resting symptoms) or to maximum 90 degrees hip flexion. Positive SLR defined as provocation of patient's symptoms that increased with ankle DF or hip IR for structural differentiation.
Strender et al. (1997)	Private outpatient clinic specializing in back pain. Consecutive sample. Patients with low back pain. (n = 50) 66% female Mean age = 37.7 (SD 11.7)	Two physiotherapists who passed the highest examination in manual medicine in Sweden. Both had "long clinical experience" with patients with low back pain. Inter-rater reliability. (Study also included a physician rating pair on separate sample of patients (n = 21). However, unable to calculate inter-rater reliability because no positive findings on SLR.)	SLR to onset of pain below the knee with an upper limit of 80 degrees. Positive SLR defined as provocation of pain radiating below the knee that increased with neck flexion and/or ankle DF.
van den Hoogen et al. (1996)	Eleven general practices. Consecutive sample. Patients with low back pain ± leg pain. (n = 50; 49 in analysis) 50% female Mean age = 46 (SD not reported)	One of 15 GPs at the 11 general practices performed the first examination. The second examination performed by the primary author who was also a GP. Experience level unclear. Inter-rater reliability.	SLR until the leg could not be raised further. Positive SLR defined as provocation of sciatic pain below the knee (Laségue's sign). No structural differentiation.
Waddell et al. (1982)	Orthopaedic outpatient clinic. Sampling unclear. Patients recruited from those referred to an orthopaedic outpatient clinic for backache. Patients with backache. (n = 30) Percent female not reported. Mean age not reported.	Two orthopaedic surgeons. Experience level unclear. Inter-rater reliability.	SLR to pain tolerance (maximum tolerated SLR). Positive SLR defined as pain tolerance (maximum tolerated SLR) < 75 degrees. (Although patients could distinguish between SLR limited by hamstring tightness, back pain, or radiating leg pain, this was not part of data reported for inter-rater reliability.)
Waddell et al. (1992)	Orthopaedic outpatient clinic. Sampling unclear. Reliability subset 1 taken from clinic sample of 120 patients with primary complaint of chronic low back pain (> 3 months duration) ± buttock or thigh pain. Patients with nerve root pain or neurological symptoms or signs excluded. (n = 20) Percent female unclear. Mean age not reported. 51% female full clinic sample. Mean age primary referrals (n = 94) full clinic sample = 35.3 (SD 9.9); tertiary referrals (n = 26) full clinic sample = 34.5 (SD 8.7).	Reliability subset 1: Two of the authors. Experience level unclear.	SLR to pain tolerance (maximum tolerated SLR). SLR ROM measured with inclinometer placed on crest of tibia just below tibial tuberosity.
	Same setting and sampling method as described above for Waddell et al. (1992) Reliability subset 1. Reliability subset 3 taken from same clinic sample as described above for Waddell et al. (1992) Reliability subset 1. (n = 20) Percent female not reported. Mean age not reported.	Reliability subset 3: Two of the authors. Unclear whether same or different from authors who examined patients in reliability subset 1. Experience level unclear.	Same as described above for Waddell et al. (1992) reliability subset 1.

* Data from same sample of 20 participants with low back-related unilateral leg pain.

† Although raters not clearly described, they were likely the same two physiotherapist raters with 3 and 12 months experience in the study reported by Walsh and Hall (2009b) because data from same sample of 20 participants.

ABD, abduction; ADD, adduction; DF, dorsiflexion; DTR, deep tendon reflex; ER, external rotation; GP, general practitioner; IR, internal rotation; PF, plantarflexion; R, rater; ROM, range of motion; SD, standard deviation; SLR, straight leg raise.

Table 2
Risk of bias of included studies according to the Quality Appraisal Tool for Studies of Diagnostic Reliability (QAREL). "Mixed" samples involved inclusion of participants with low back-related leg pain or low back pain only.

Study (Low back-related leg pain)	Representative patients?	Representative raters?	Inter-rater blinding?	Blind to own prior findings?	Blind to reference standard?	Blind to clinical information?	Blind to additional cues?	Varied order of examination?	Suitable time interval?	Appropriate test application/ interpretation?	Appropriate statistical analysis?
Poiraudeau (2001)	🟢	🟢	🟢	🟡	🟢	🟡	🟡	🟢	🟡	🔴	🔴
Vroomen (2000) *	🟢	🟢	🟡	⚪	⚪	🟡	🟡	🟢	🟡	🟢	🟢
Walsh (2009a) SLR ROM	🟢	🟡	🟢	⚪	⚪	🟡	🟡	🟡	🟢	🟢	🟢
Walsh (2009b)	🟢	🟢	🟢	⚪	⚪	🟡	🟡	🔴	🟢	🟢	🟢
("Mixed" samples of participants)											
Bertilson (2006) *	🟢	🟢	🟢	⚪	⚪	🔴	🟡	🟢	🟢	🟢	🟢
Billis (2012) *	🟢	🟢	🟢	⚪	⚪	🟢	🟡	🟢	🟢	🟢	🟢
Boland (2000)	🟢	🟢	🟢	⚪	⚪	🔴	🟡	🟢	🟢	🟢	🟢
Chow (1994)	🟢	🟡	⚪	🟡	⚪	🟡	🟡	⚪	🟢	🟢	🟢
McCombe (1989) *	🟡	🟢	🟡	⚪	⚪	🟡	🟡	🟡	🟡	🔴	🟢
Paatelma (2010) *	🟢	🟢	🟢	🟡	⚪	🟡	🟡	🔴	🟢	🔴	🟢
Pesonen (2021)	🔴	🟢	🟢	⚪	🟢	🟢	🟡	🟡	🟡	🟢	🟢
Strender (1997) *	🟢	🟢	🟢	⚪	⚪	🟢	🟡	🟢	🟢	🟢	🟢
van den Hoogen (1996)	🟢	🟢	🟢	⚪	🟡	🔴	🟡	🔴	🔴	🔴	🟢
Waddell (1982) *	🟡	🟢	🟢	⚪	⚪	🟡	🟡	🟢	🟡	🔴	🔴
Waddell (1992) *	🟢	🟢	🟢	⚪	⚪	🟡	🟡	🟡	🟢	🟢	🔴

Yes 🟢 No 🔴 Unclear 🟡 N/A ⚪

* Studies looked at SLR as part of composite clinical examination. ROM, range of motion; SLR, straight leg raise.

3. Results

3.1. Study selection

Results of the study identification and selection process are summarized in Fig. 1. Four studies reported SLR or crossed SLR reliability in three distinct samples of participants with low back-related leg pain (n = 189) (Table 1A). Walsh and Hall (2009a; b) reported reliability for identifying a positive SLR and measuring SLR ROM in separate publications based on data from the same sample (n = 20). As noted previously, most participants (169/189) had severe low back-related leg pain that required hospitalization (Poiraudeau et al., 2001) or bed rest (Vroomen et al., 2000). Eleven studies were included in supplemental analyses and reported SLR reliability in 14 distinct "mixed" samples of participants with low back-related leg pain or low back pain only (n = 439) (Table 1B).

3.2. Risk of bias

QAREL ratings for each study are reported in Table 2. Proportions of studies that satisfied each QAREL item are presented in Fig. 2.

3.2.1. Low back-related leg pain

QAREL ratings suggested two main risks of bias (Fig. 2A). Only two of four studies clearly prevented an order effect by varying the order of raters. Similarly, only two of four studies clearly ensured each participant's condition remained stable by using a suitable time interval between examinations.

No studies were clear about blinding raters to other clinical information or to additional cues about participants (e.g., tattoos, voice accent). Lack of clarity in blinding raters to other clinical information was a minor concern because SLR and crossed SLR are interpreted clinically within the context of a full examination. Lack of clarity about blinding raters to additional cues about participants was also a minor concern because three of the four studies reported only inter-rater reliability where each rater examined each participant only once. However, this could be a source of bias for estimates of intra-rater reliability reported by Poiraudeau et al. (2001). Lack of clarity about blinding raters to their own prior test findings is another potential source of bias for estimates of intra-rater reliability from this study.

3.2.2. "Mixed" samples

There were two main risks of bias in studies that enrolled participants with low back-related leg pain or low back pain only (Fig. 2B). Only five of 11 studies varied the order of raters, and only seven used structural differentiation to categorize the SLR as positive or negative.

Only three of 11 studies blinded raters to other clinical information and no studies were clear about blinding raters to additional cues about participants. These were minor concerns for reasons outlined previously. However, lack of clarity about blinding raters to additional cues about participants and to their own prior test findings are potential sources of bias for estimates of intra-rater reliability reported by Chow et al. (1994) and Paatelma et al. (2010).

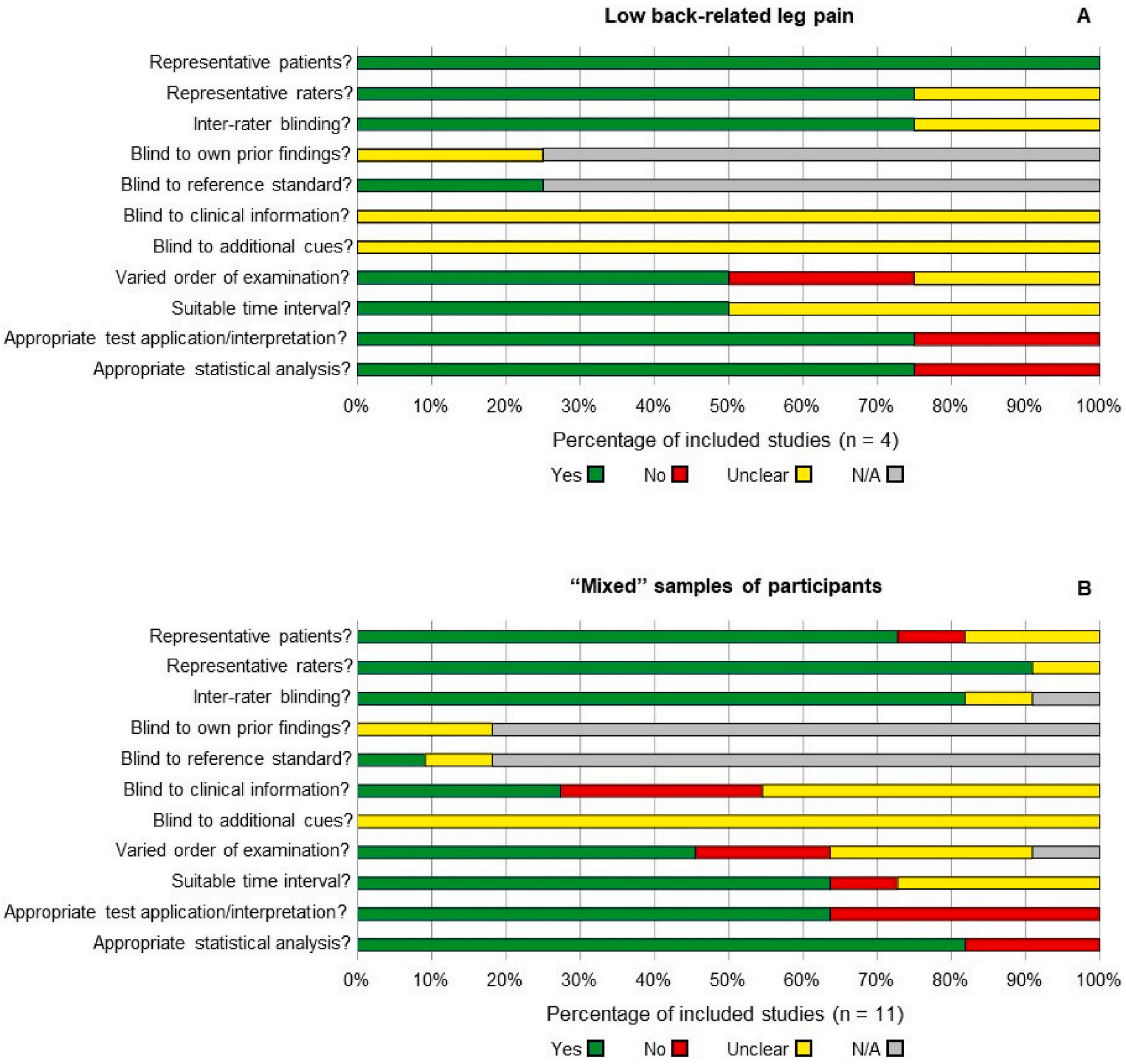


Fig. 2. Summary of risk of bias for each item on the Quality Appraisal Tool for Studies of Diagnostic Reliability (QAREL).

3.3. *Reliability outcomes for SLR and crossed SLR*

Because of inconsistent terminology, studies reporting provocation of participants' specific symptoms were interpreted separately from studies reporting provocation of pain. Confidence in the evidence for inter-rater reliability when a positive SLR included provocation of lower extremity pain or symptoms was moderate to very low (Table 3). Confidence in the evidence for inter-rater reliability for identifying a positive crossed SLR was moderate to very low (Table 3).

3.3.1. *SLR provokes symptoms with structural differentiation*

Meta-analysis showed moderate inter-rater reliability for identifying a positive SLR based on provocation of symptoms that changed with structural differentiation in patients with low back-related leg pain (n = 111) (Vroomen et al., 2000; Walsh and Hall, 2009b) (Fig. 3A). Meta-analysis showed substantial inter-rater reliability for this definition of a positive SLR in "mixed" samples of participants (n = 170) (Strender et al., 1997; Bertilson et al., 2006; Billis et al., 2012; Pesonen et al., 2021) (Fig. 3B).

3.3.2. *SLR provokes symptoms without structural differentiation*

Intra-rater reliability for identifying a positive SLR based on provocation of symptoms without structural differentiation in patients with low back-related leg pain was moderate to substantial (n = 78) (Poiraudeau et al., 2001). Inter-rater reliability was slight to moderate (n = 169) (Vroomen et al., 2000; Poiraudeau et al., 2001) (Fig. 4A). Meta-analysis for inter-rater reliability was not possible because Poiraudeau et al. (2001) did not report adequate data to calculate variance and the authors could not be reached. Meta-analysis showed fair inter-rater reliability for this definition of a positive SLR in "mixed" samples of participants (n = 113) (McCombe et al., 1989; Billis et al., 2012) (Fig. 4B).

3.3.3. *SLR provokes pain below the knee without structural differentiation*

No studies on patients with low back-related leg pain reported reliability for identifying a positive SLR based on provocation of pain below the knee without structural differentiation. Meta-analysis showed fair inter-rater reliability for this definition of a positive SLR in "mixed" samples of participants (n = 132) (McCombe et al., 1989; van den Hoogen et al., 1996) (Fig. 5).

3.3.4. *SLR provokes low back and/or lower extremity pain without structural differentiation*

Vroomen et al. (2000) reported slight inter-rater reliability for

Table 3
Confidence in the evidence for inter-rater reliability for definitions of a positive SLR or crossed SLR commonly used to detect lumbar radicular pain.

Inter-rater reliability outcome	Kappa (95% CI)	Number of participants (studies)	Confidence in the evidence (GRADE)*
SLR provokes symptoms with structural differentiation			
Low back-related leg pain	0.67† (0.60, 0.75)	111 (2 studies)	⊕⊕⊕⊖ **Moderate** due to risk of bias[1]
"Mixed" samples	0.83† (0.70, 0.96)	170 (4 studies)	⊕⊕⊕⊖ **Moderate** due to indirectness[2]
SLR provokes symptoms without structural differentiation			
Low back-related leg pain	Range 0.29 – 0.68‡	169 (2 studies)	⊕⊕⊖⊖ **Low** due to risk of bias[1], inconsistency[3]
"Mixed" samples	0.53† (0.23, 0.83)	113 (2 studies; 3 samples)	⊕⊖⊖⊖ **Very low** due to risk of bias[1], inconsistency[4], indirectness[2], imprecision[5]
SLR provokes pain below the knee without structural differentiation			
"Mixed" samples	0.54† (0.37, 0.72)	132 (2 studies; 3 samples)	⊕⊖⊖⊖ **Very low** due to risk of bias[1], indirectness[2], imprecision[5]
SLR provokes low back and/or lower extremity pain without structural differentiation			
Low back-related leg pain	0.36 (0.11, 61)	91 (1 study)	⊕⊕⊖⊖ **Low** due to risk of bias[1], imprecision[5]
"Mixed" samples	0.53† (0.30, 0.75)	98 (2 studies; 3 samples)	⊕⊖⊖⊖ **Very low** due to risk of bias[1], inconsistency[4], indirectness[2], imprecision[5]
Crossed SLR provokes low back and/or lower extremity symptoms			
Low back-related leg pain	Range 0.43 – 0.72‡	169 (2 studies)	⊕⊕⊕⊖ **Moderate** due to risk of bias[1]
"Mixed" samples	Range -0.02 – 0.74‡	83 (1 study; 2 samples)	⊕⊖⊖⊖ **Very low** due to risk of bias[1], inconsistency[3], indirectness[2], imprecision[6]

Footnotes

[1] > 25% of participants providing data came from studies with important risks of bias.
[2] Samples included participants who had low back pain only.
[3] Wide-ranging point estimates for Kappa with minimal or no overlap in 95% CI.
[4] Statistical heterogeneity (I^2) > 50%.
[5] 95% CI for pooled estimate of Kappa contained values above and below 0.60.
[6] Given that meta-analysis not possible, total number of participants did not meet review information size of 165.

* **High confidence**: Further research is very unlikely to change our confidence in the estimate of reliability.
Moderate confidence: Further research is likely to have an important impact on our confidence in the estimate of reliability and may change the estimate.
Low confidence: Further research is very likely to have an important impact on our confidence in the estimate of reliability and is likely to change the estimate.
Very low confidence: We have very little confidence in the estimate of reliability.
† Meta-analysis
‡ Meta-analysis not possible
CI, confidence interval

identifying a positive SLR based on provocation of lower extremity pain without structural differentiation in patients with low back-related leg pain (n = 91) (Fig. 6A).

Two studies on "mixed" samples of participants (n = 98) reported intra-rater and inter-rater reliability for identifying a positive SLR based on provocation of pain in the buttock or more distally (Paatelma et al., 2010), and inter-rater reliability when a positive SLR was defined as provocation of back and lower extremity pain ("sciatic stretch") (McCombe et al., 1989). Intra-rater reliability was moderate while meta-analysis showed fair inter-rater reliability (Fig. 6B).

3.3.5. *SLR provokes pain below a ROM threshold without structural differentiation*

Vroomen et al. (2000) reported fair inter-rater reliability for identifying a positive SLR based on provoking onset of pain below 45° in patients with low back-related leg pain (n = 91) (Fig. 7A). Waddell et al. (1982) reported fair inter-rater reliability for identifying a positive SLR based on pain tolerance below 75° in a "mixed" sample of participants (n = 30) (Fig. 7B).

3.3.6. *Crossed SLR*

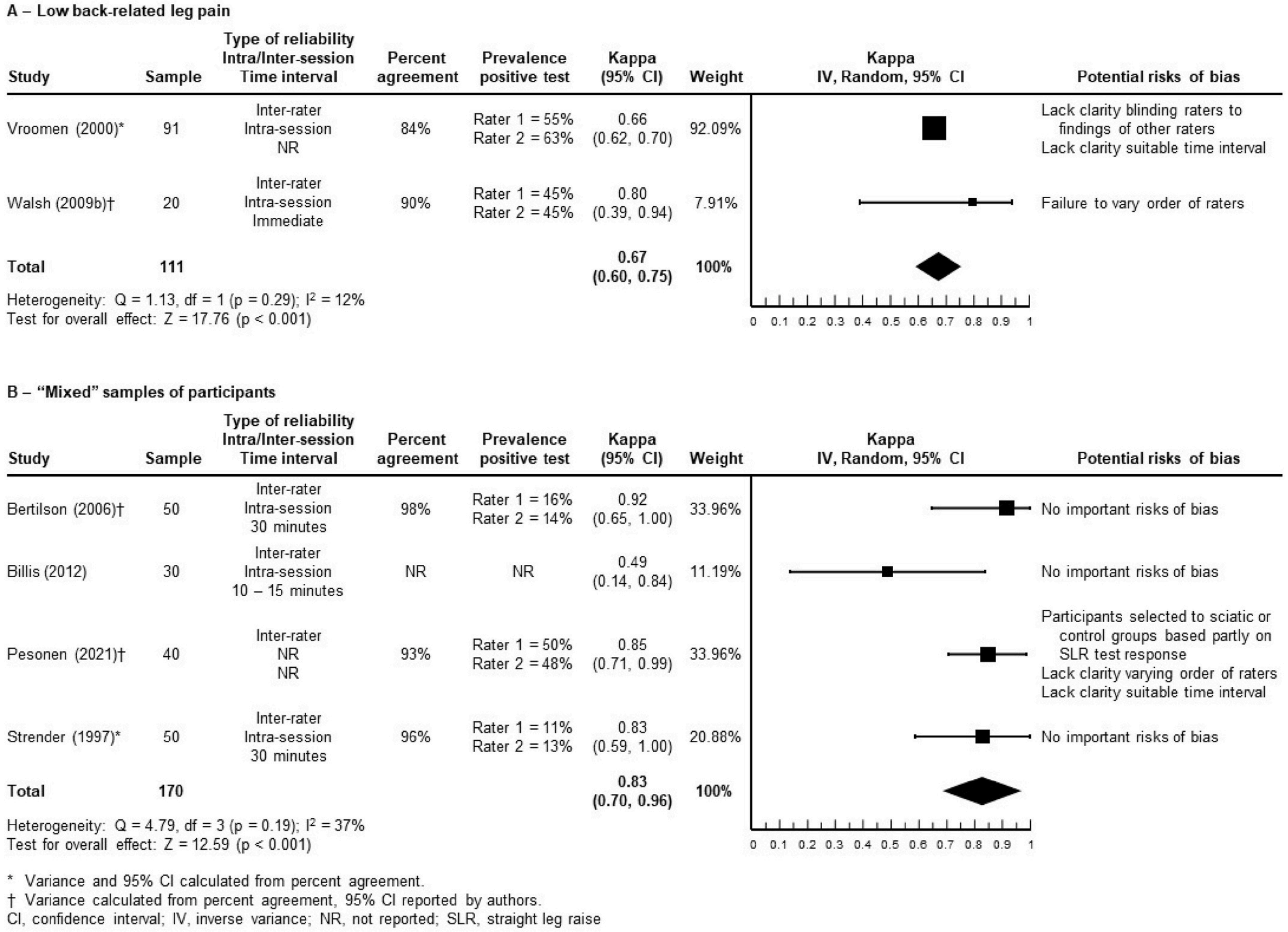
Intra-rater reliability for identifying a positive crossed SLR based on provocation of low back and/or lower extremity symptoms in patients with low back-related leg pain was moderate to substantial (n = 78) (Poiraudeau et al., 2001). Inter-rater reliability was fair to moderate (n = 169) (Vroomen et al., 2000; Poiraudeau et al., 2001) (Fig. 8A). Meta-analysis for inter-rater reliability was not possible because of previously stated issues regarding Poiraudeau et al. (2001).

Inter-rater reliability for this definition of a positive crossed SLR in two "mixed" samples of participants was virtually none to moderate (n = 83) (McCombe et al., 1989) (Fig. 8B). No data were reported to provide insight into the large discrepancy in inter-rater reliability between the surgeon (moderate) and surgeon/physiotherapist (virtually none) rater pairs. Meta-analysis was therefore not performed.

3.3.7. *Measuring SLR ROM*

Walsh and Hall (2009a) reported moderate to substantial inter-rater reliability for measuring SLR ROM in the asymptomatic and symptomatic limbs, respectively, in patients with low back-related leg pain (n = 20) (Table 4A). Corresponding SDD_{95} values ranged from 16 to 20° (Table 4A).

Three studies reported intra-rater (Chow et al., 1994) or inter-rater (Waddell et al., 1992; Boland and Adams, 2000) reliability for measuring SLR ROM in "mixed" samples of participants (n = 91). Intra-rater reliability for measuring SLR ROM in the symptomatic limb was substantial with an intra-session SDD_{95} value of 6.1° (Table 4B). The pooled estimate of inter-rater reliability for measuring SLR ROM in the symptomatic limb was substantial with an intra-session SDD_{95} value of 17.7° (Boland and Adams, 2000) (Table 4B). Although Waddell et al. (1992) reported substantial inter-rater reliability for measuring SLR



A – Low back-related leg pain

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Vroomen (2000)*	91	Inter-rater Intra-session NR	84%	Rater 1 = 55% Rater 2 = 63%	0.66 (0.62, 0.70)	92.09%	Lack clarity blinding raters to findings of other raters Lack clarity suitable time interval
Walsh (2009b)†	20	Inter-rater Intra-session Immediate	90%	Rater 1 = 45% Rater 2 = 45%	0.80 (0.39, 0.94)	7.91%	Failure to vary order of raters
Total	**111**				**0.67** **(0.60, 0.75)**	**100%**	

Heterogeneity: Q = 1.13, df = 1 (p = 0.29); I^2 = 12%
Test for overall effect: Z = 17.76 (p < 0.001)

B – "Mixed" samples of participants

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Bertilson (2006)†	50	Inter-rater Intra-session 30 minutes	98%	Rater 1 = 16% Rater 2 = 14%	0.92 (0.65, 1.00)	33.96%	No important risks of bias
Billis (2012)	30	Inter-rater Intra-session 10 – 15 minutes	NR	NR	0.49 (0.14, 0.84)	11.19%	No important risks of bias
Pesonen (2021)†	40	Inter-rater NR NR	93%	Rater 1 = 50% Rater 2 = 48%	0.85 (0.71, 0.99)	33.96%	Participants selected to sciatic or control groups based partly on SLR test response Lack clarity varying order of raters Lack clarity suitable time interval
Strender (1997)*	50	Inter-rater Intra-session 30 minutes	96%	Rater 1 = 11% Rater 2 = 13%	0.83 (0.59, 1.00)	20.88%	No important risks of bias
Total	**170**				**0.83** **(0.70, 0.96)**	**100%**	

Heterogeneity: Q = 4.79, df = 3 (p = 0.19); I^2 = 37%
Test for overall effect: Z = 12.59 (p < 0.001)

* Variance and 95% CI calculated from percent agreement.
† Variance calculated from percent agreement, 95% CI reported by authors.
CI, confidence interval; IV, inverse variance; NR, not reported; SLR, straight leg raise

Fig. 3. Reliability for categorizing SLR as positive or negative based on provocation of symptoms with structural differentiation.

ROM, results were reported for right and left limbs, rather than asymptomatic and symptomatic limbs (Table 4B). Furthermore, no data for calculating absolute reliability (measurement error) were reported. The authors could not be reached. Consequently, no meta-analysis was performed on data from Boland and Adams (2000) and Waddell et al. (1992).

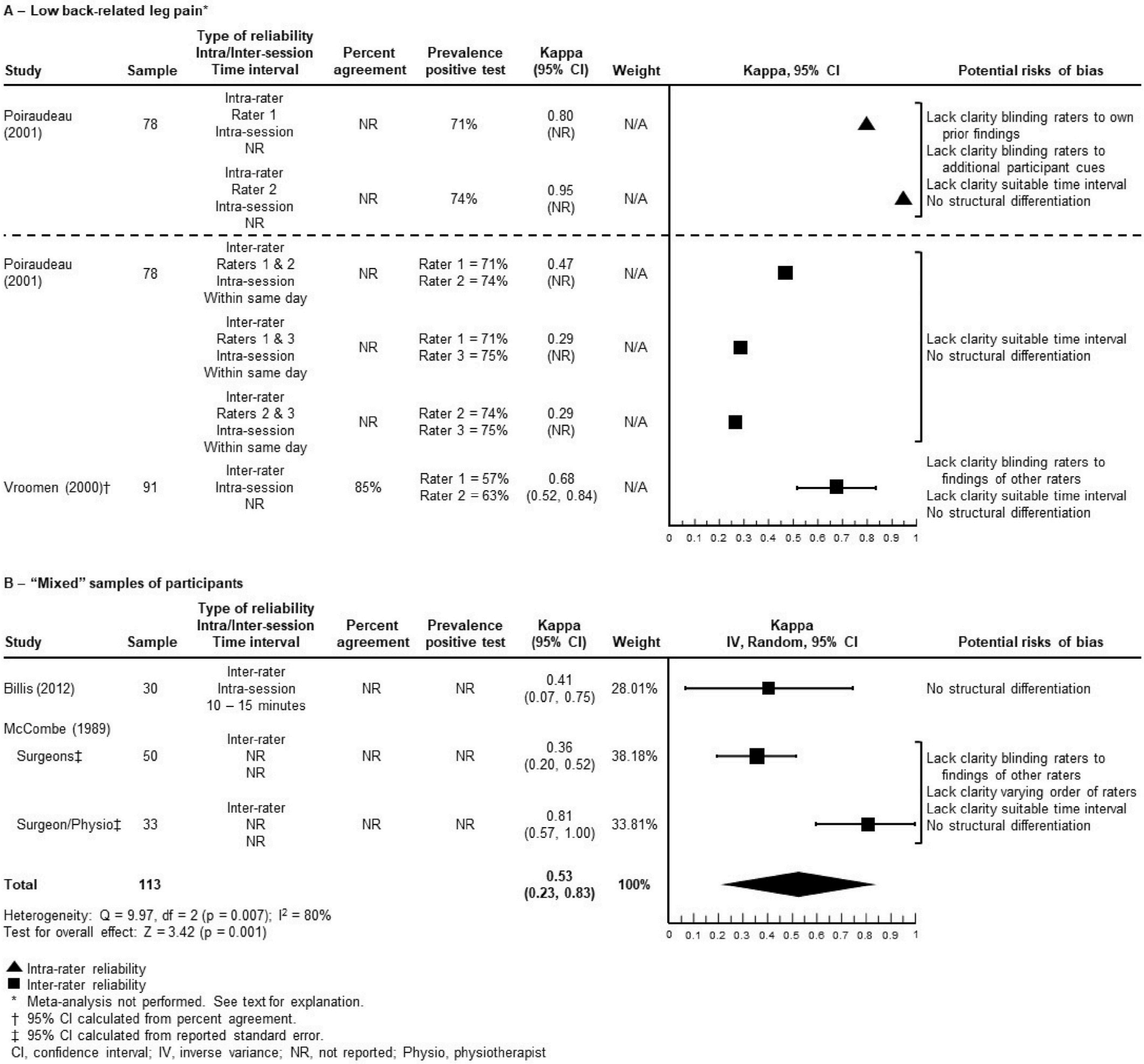
Boland and Adams (2000) also reported inter-rater reliability for measuring ROM during a modified SLR where ankle dorsiflexion preceded hip flexion (n = 35). The pooled estimate of inter-rater reliability was substantial with a SDD_{95} of 16.9° (Table 4B).

4. Discussion

This systematic review summarized SLR and crossed SLR reliability in patients with suspected lumbar radicular pain. In samples including only participants with low back-related leg pain, applicability of reliability data was limited by most participants having severe symptoms that warranted hospitalization or bed rest or by small samples. Supplemental analyses were therefore performed on "mixed" samples of participants with low back-related leg pain or low back pain only. Findings were typically consistent between primary and supplemental analyses. Confidence in the evidence for definitions of a positive SLR or crossed SLR commonly used to detect lumbar radicular pain was moderate to very low. Inter-rater reliability for identifying a positive SLR was moderate (low back-related leg pain samples) to substantial ("mixed" samples) when a positive test provoked the patient's symptoms and those symptoms changed with structural differentiation. Large errors occurred when different clinicians measured SLR range of motion on the same patient.

Interpreting test reliability requires additional context provided by diagnostic performance data (Fritz and Wainner, 2001). As stated previously, systematic reviews suggest the SLR performs poorly in diagnosing lumbar radicular pain (van der Windt et al., 2010; Scaia et al., 2012; Tawa et al., 2017; Mistry et al., 2020). When described, the most common definition of a positive test was provocation of pain/symptoms below the knee. Structural differentiation was rarely used, and these reviews did not address whether structural differentiation impacted diagnostic performance. Our meta-analyses showed that provoking pain below the knee without structural differentiation had fair reliability (Fig. 5) while provoking a patient's specific symptoms and changing those symptoms with structural differentiation had at least moderate reliability (Fig. 3). Limitations in SLR reliability therefore seem less likely to be the main reason for this poor diagnostic performance.

Categorizing patients as having lumbar radicular pain typically relies on findings from a full clinical examination. Using history and physical examination to diagnose nerve root involvement in patients with low back-related leg pain in primary care has only slight inter-rater reliability (Kappa 0.35; 95%CI 0.07, 0.63) (Stynes et al., 2016). However, reliability was more acceptable when clinicians had greater confidence in their diagnosis. The authors did not explore factors associated with clinicians' diagnostic confidence. It is therefore unclear whether a more reliable definition of a "positive" SLR that incorporates structural differentiation would improve diagnostic confidence and consistency in categorizing patients as having lumbar radicular pain.



A – Low back-related leg pain*

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Poiraudeau (2001)	78	Intra-rater Rater 1 Intra-session NR	NR	71%	0.80 (NR)	N/A	Lack clarity blinding raters to own prior findings Lack clarity blinding raters to additional participant cues Lack clarity suitable time interval No structural differentiation
		Intra-rater Rater 2 Intra-session NR	NR	74%	0.95 (NR)	N/A	
Poiraudeau (2001)	78	Inter-rater Raters 1 & 2 Intra-session Within same day	NR	Rater 1 = 71% Rater 2 = 74%	0.47 (NR)	N/A	Lack clarity suitable time interval No structural differentiation
		Inter-rater Raters 1 & 3 Intra-session Within same day	NR	Rater 1 = 71% Rater 3 = 75%	0.29 (NR)	N/A	
		Inter-rater Raters 2 & 3 Intra-session Within same day	NR	Rater 2 = 74% Rater 3 = 75%	0.29 (NR)	N/A	
Vroomen (2000)†	91	Inter-rater Intra-session NR	85%	Rater 1 = 57% Rater 2 = 63%	0.68 (0.52, 0.84)	N/A	Lack clarity blinding raters to findings of other raters Lack clarity suitable time interval No structural differentiation

B – "Mixed" samples of participants

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Billis (2012)	30	Inter-rater Intra-session 10 – 15 minutes	NR	NR	0.41 (0.07, 0.75)	28.01%	No structural differentiation
McCombe (1989)							
Surgeons‡	50	Inter-rater NR NR	NR	NR	0.36 (0.20, 0.52)	38.18%	Lack clarity blinding raters to findings of other raters Lack clarity varying order of raters Lack clarity suitable time interval No structural differentiation
Surgeon/Physio‡	33	Inter-rater NR NR	NR	NR	0.81 (0.57, 1.00)	33.81%	
Total	**113**				**0.53** **(0.23, 0.83)**	**100%**	

Heterogeneity: Q = 9.97, df = 2 (p = 0.007); I^2 = 80%
Test for overall effect: Z = 3.42 (p = 0.001)

▲ Intra-rater reliability
■ Inter-rater reliability
* Meta-analysis not performed. See text for explanation.
† 95% CI calculated from percent agreement.
‡ 95% CI calculated from reported standard error.
CI, confidence interval; IV, inverse variance; NR, not reported; Physio, physiotherapist

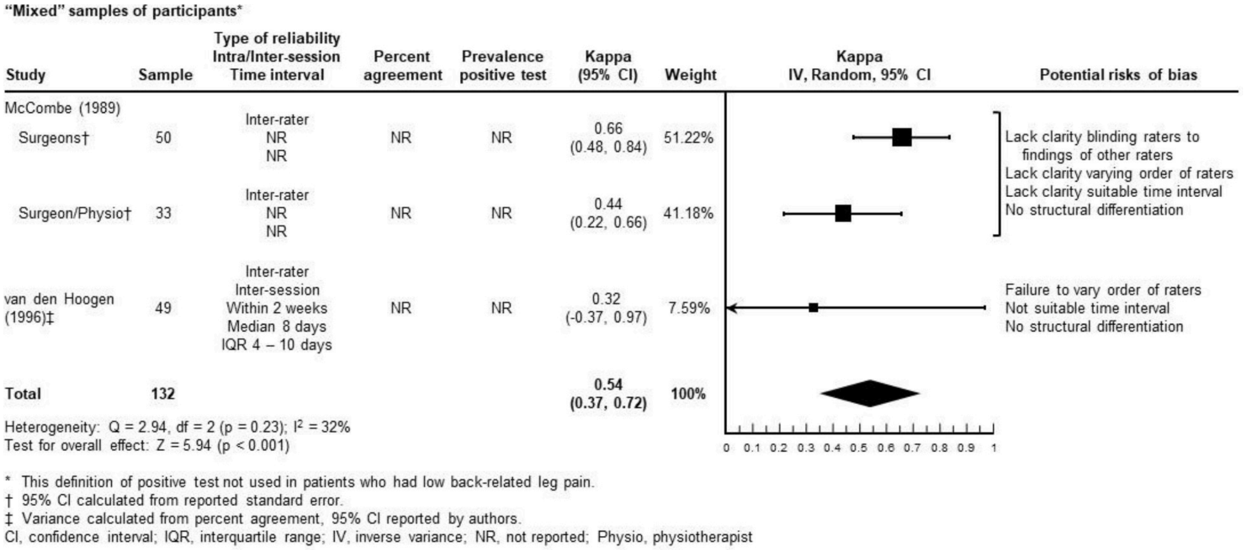
Fig. 4. Reliability for categorizing SLR as positive or negative based on provocation of symptoms without structural differentiation.

Wide-ranging estimates of inter-rater reliability for identifying a positive crossed SLR prevent definitive conclusions about test reliability. Part of the difficulty in estimating crossed SLR reliability is that a low prevalence of positive tests can deflate Kappa values (Sim and Wright, 2005). A low prevalence of positive tests was evident in this review (Fig. 8). Even though data suggest a positive crossed SLR helps identify lumbar radicular pain related to disc herniation (van der Windt et al., 2010; Stynes et al., 2018), lack of clarity about inter-rater reliability requires caution when using this test to confirm this diagnosis.

Despite substantial relative reliability, absolute reliability between raters for measuring SLR ROM in the symptomatic limb is poor. Corresponding SDD_{95} values ranged from 13 to 20° (Table 4). This means that SLR ROM measurements on the same patient obtained by two clinicians can differ by up to 20° because of measurement error. Measurement error is therefore a significant barrier to establishing a ROM threshold for defining a positive SLR.

SLR ROM is sometimes used as an outcome to assess treatment effects for nerve-related back and leg pain (Basson et al., 2017; Pourahmadi et al., 2019). Several questions may need consideration if this practice is to continue. Is SLR ROM a relevant outcome for patients? Although SLR ROM improvements of more than 6° within a treatment session are unlikely to be due to measurement error (Chow et al., 1994) (Table 4), how much within-session change in SLR ROM is clinically important? How much error occurs when the same clinician measures SLR ROM on different days and what constitutes a clinically important between-session change in SLR ROM?

Assessing risk of bias with QAREL was affected by incomplete reporting in many studies. For inter-rater reliability, failure to use structural differentiation to define a positive SLR was a common risk of bias. Even when structural differentiation was used, studies usually did not clearly describe whether the location of SLR-provoked symptoms dictated the chosen structural differentiation manoeuvre. Future studies should provide these details to allow judgment of the quality of the structural differentiation process. Other common risks of bias were lack



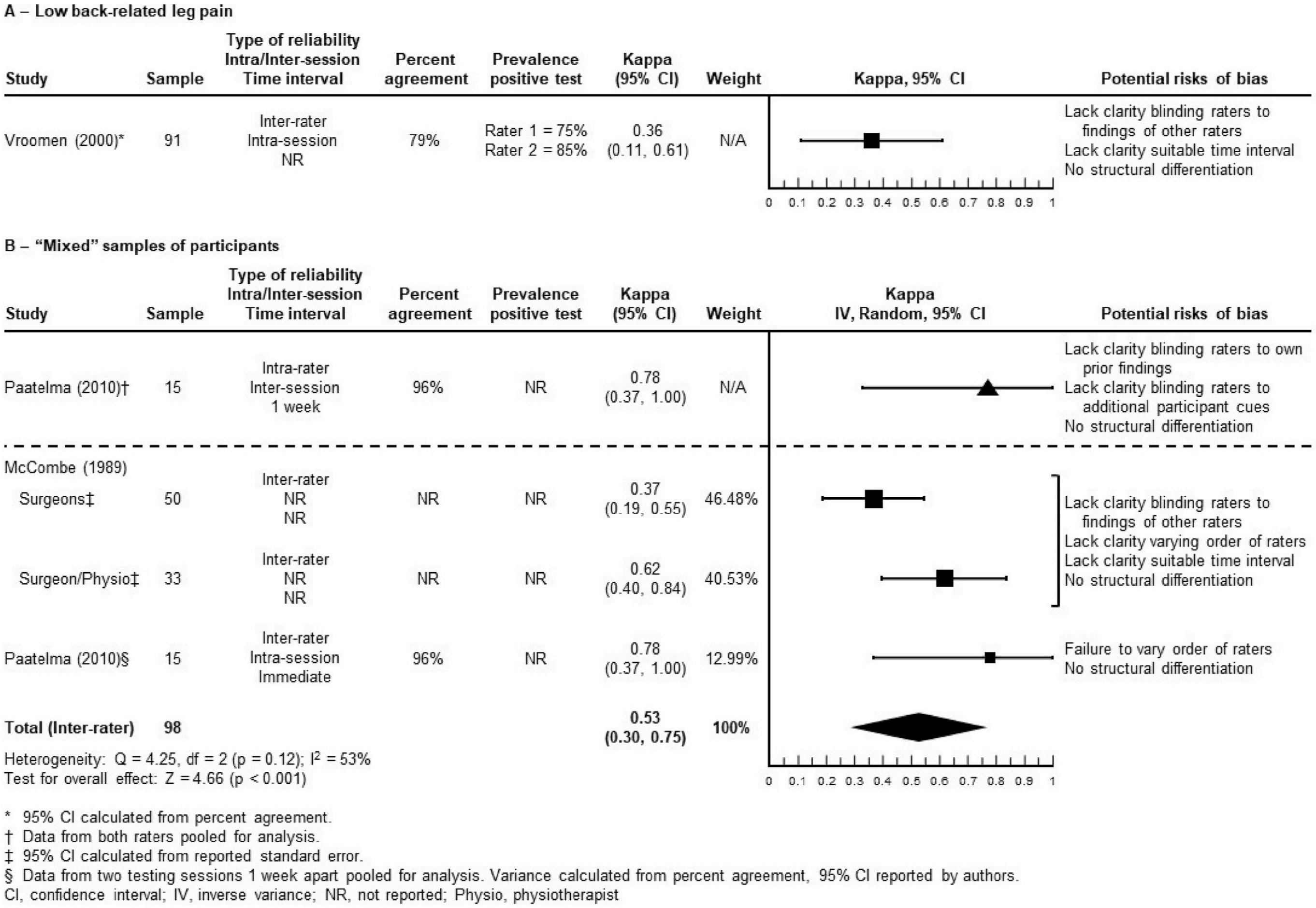
"Mixed" samples of participants*

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
McCombe (1989)							
Surgeons†	50	Inter-rater NR NR	NR	NR	0.66 (0.48, 0.84)	51.22%	Lack clarity blinding raters to findings of other raters Lack clarity varying order of raters Lack clarity suitable time interval No structural differentiation
Surgeon/Physio†	33	Inter-rater NR NR	NR	NR	0.44 (0.22, 0.66)	41.18%	
van den Hoogen (1996)‡	49	Inter-rater Inter-session Within 2 weeks Median 8 days IQR 4 – 10 days	NR	NR	0.32 (-0.37, 0.97)	7.59%	Failure to vary order of raters Not suitable time interval No structural differentiation
Total	**132**				**0.54 (0.37, 0.72)**	**100%**	

Heterogeneity: Q = 2.94, df = 2 (p = 0.23); I^2 = 32%
Test for overall effect: Z = 5.94 (p < 0.001)

* This definition of positive test not used in patients who had low back-related leg pain.
† 95% CI calculated from reported standard error.
‡ Variance calculated from percent agreement, 95% CI reported by authors.
CI, confidence interval; IQR, interquartile range; IV, inverse variance; NR, not reported; Physio, physiotherapist

Fig. 5. Reliability of categorizing SLR as positive or negative based on provocation of pain below the knee without structural differentiation.



A – Low back-related leg pain

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Vroomen (2000)*	91	Inter-rater Intra-session NR	79%	Rater 1 = 75% Rater 2 = 85%	0.36 (0.11, 0.61)	N/A	Lack clarity blinding raters to findings of other raters Lack clarity suitable time interval No structural differentiation

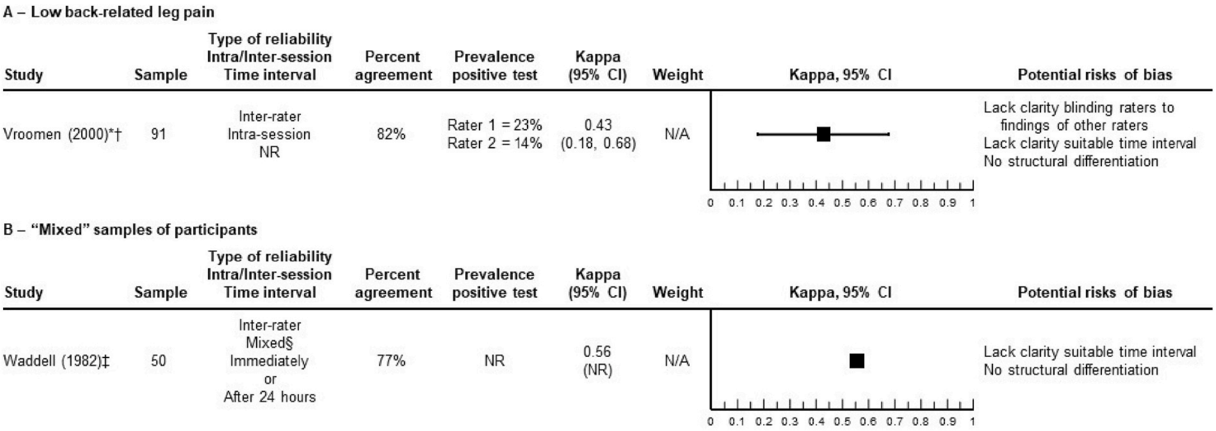
B – "Mixed" samples of participants

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Paatelma (2010)†	15	Intra-rater Inter-session 1 week	96%	NR	0.78 (0.37, 1.00)	N/A	Lack clarity blinding raters to own prior findings Lack clarity blinding raters to additional participant cues No structural differentiation
McCombe (1989)							
Surgeons‡	50	Inter-rater NR NR	NR	NR	0.37 (0.19, 0.55)	46.48%	Lack clarity blinding raters to findings of other raters Lack clarity varying order of raters Lack clarity suitable time interval No structural differentiation
Surgeon/Physio‡	33	Inter-rater NR NR	NR	NR	0.62 (0.40, 0.84)	40.53%	
Paatelma (2010)§	15	Inter-rater Intra-session Immediate	96%	NR	0.78 (0.37, 1.00)	12.99%	Failure to vary order of raters No structural differentiation
Total (Inter-rater)	**98**				**0.53 (0.30, 0.75)**	**100%**	

Heterogeneity: Q = 4.25, df = 2 (p = 0.12); I^2 = 53%
Test for overall effect: Z = 4.66 (p < 0.001)

* 95% CI calculated from percent agreement.
† Data from both raters pooled for analysis.
‡ 95% CI calculated from reported standard error.
§ Data from two testing sessions 1 week apart pooled for analysis. Variance calculated from percent agreement, 95% CI reported by authors.
CI, confidence interval; IV, inverse variance; NR, not reported; Physio, physiotherapist

Fig. 6. Reliability of categorizing SLR as positive or negative based on provocation of low back and/or lower extremity pain without structural differentiation.



A – Low back-related leg pain

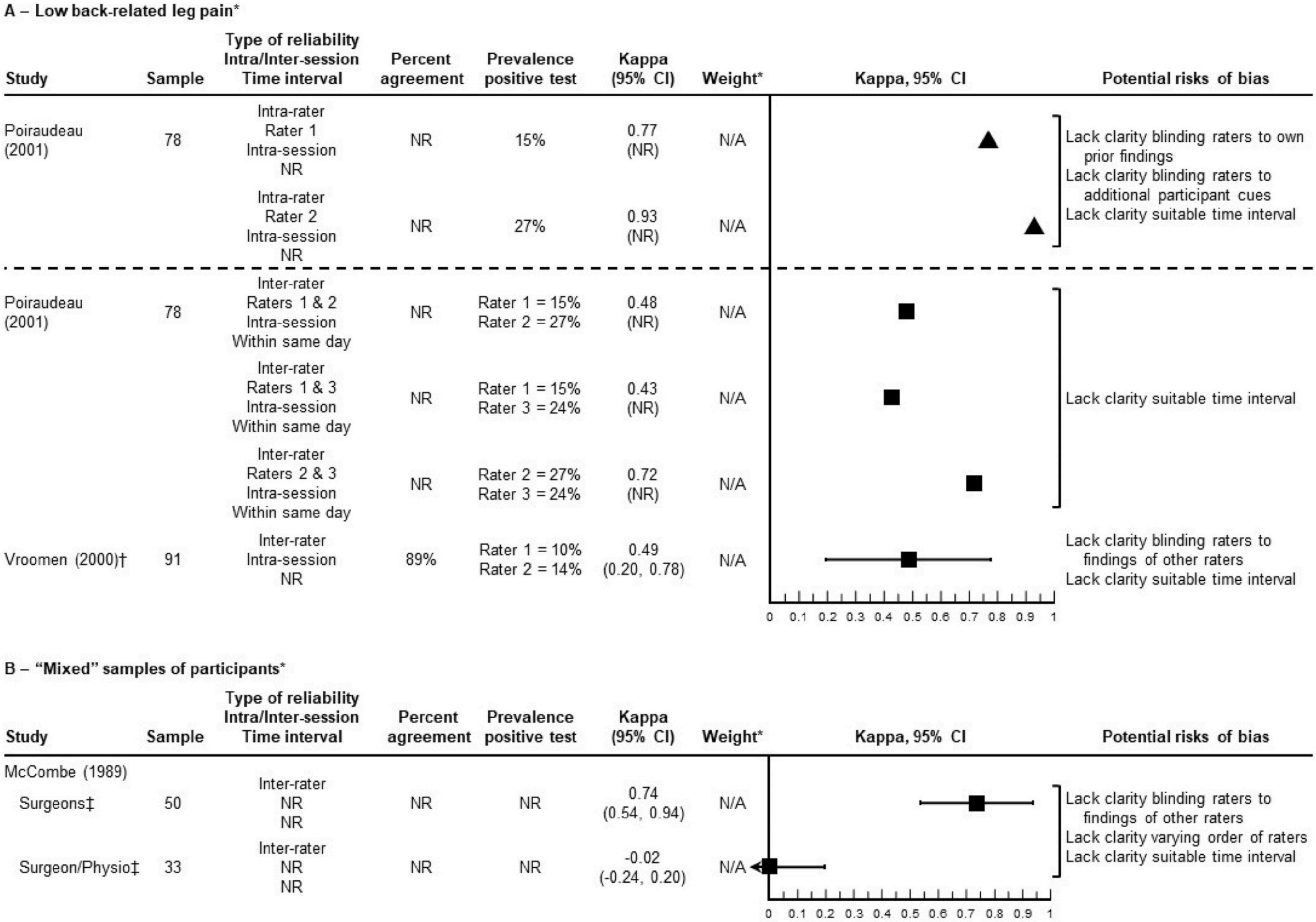
Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Vroomen (2000)*†	91	Inter-rater Intra-session NR	82%	Rater 1 = 23% Rater 2 = 14%	0.43 (0.18, 0.68)	N/A	Lack clarity blinding raters to findings of other raters Lack clarity suitable time interval No structural differentiation

B – "Mixed" samples of participants

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight	Potential risks of bias
Waddell (1982)‡	50	Inter-rater Mixed§ Immediately or After 24 hours	77%	NR	0.56 (NR)	N/A	Lack clarity suitable time interval No structural differentiation

* Positive SLR = provocation of pain below 45 degrees.
† 95% CI calculated from percent agreement.
‡ Positive SLR = pain tolerance below 75 degrees.
§ Proportions of participants examined under each time interval (immediately or after 24 hours) not reported.
CI, confidence interval; NR, not reported

Fig. 7. Reliability of categorizing SLR as positive or negative based on provocation of pain below a ROM threshold without structural differentiation.



A – Low back-related leg pain*

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight*	Potential risks of bias
Poiraudeau (2001)	78	Intra-rater Rater 1 Intra-session NR	NR	15%	0.77 (NR)	N/A	Lack clarity blinding raters to own prior findings Lack clarity blinding raters to additional participant cues Lack clarity suitable time interval
		Intra-rater Rater 2 Intra-session NR	NR	27%	0.93 (NR)	N/A	
Poiraudeau (2001)	78	Inter-rater Raters 1 & 2 Intra-session Within same day	NR	Rater 1 = 15% Rater 2 = 27%	0.48 (NR)	N/A	Lack clarity suitable time interval
		Inter-rater Raters 1 & 3 Intra-session Within same day	NR	Rater 1 = 15% Rater 3 = 24%	0.43 (NR)	N/A	
		Inter-rater Raters 2 & 3 Intra-session Within same day	NR	Rater 2 = 27% Rater 3 = 24%	0.72 (NR)	N/A	
Vroomen (2000)†	91	Inter-rater Intra-session NR	89%	Rater 1 = 10% Rater 2 = 14%	0.49 (0.20, 0.78)	N/A	Lack clarity blinding raters to findings of other raters Lack clarity suitable time interval

B – "Mixed" samples of participants*

Study	Sample	Type of reliability Intra/Inter-session Time interval	Percent agreement	Prevalence positive test	Kappa (95% CI)	Weight*	Potential risks of bias
McCombe (1989)							
Surgeons‡	50	Inter-rater NR NR	NR	NR	0.74 (0.54, 0.94)	N/A	Lack clarity blinding raters to findings of other raters Lack clarity varying order of raters Lack clarity suitable time interval
Surgeon/Physio‡	33	Inter-rater NR NR	NR	NR	-0.02 (-0.24, 0.20)	N/A	

▲ Intra-rater reliability
■ Inter-rater reliability
* Meta-analysis not performed. See text for explanation.
† 95% CI calculated from percent agreement.
‡ 95% CI calculated from reported standard error.
CI, confidence interval; IV, inverse variance; NR, not reported; Physio, physiotherapist

Fig. 8. Reliability of categorizing crossed SLR as positive or negative based on provocation of low back and/or lower extremity symptoms.

Table 4
Relative and absolute reliability for measuring SLR ROM.

Study Instrument	Limb tested Type of reliability Intra-/Inter-session Time interval	ICC_{Model} (95% CI) Level of reliability	SEM (degrees)	SDD_{95} (degrees)	Potential risks of bias
A – Low back-related leg pain					
Walsh and Hall (2009a) Bubble inclinometer	Symptomatic limb Inter-rater Intra-session Immediate	$ICC_{3,1}$ 0.82 (0.49, 0.94) substantial	7.2*	20.0†	Lack of clarity varying order of examiners
	Asymptomatic limb Inter-rater Intra-session Immediate	$ICC_{3,1}$ 0.77 (0.35, 0.92) moderate	5.8*	16.1†	Lack of clarity varying order of examiners
B – "Mixed" samples of participants					
Chow et al. (1994) Bubble inclinometer	Symptomatic limb Intra-rater Intra-session 30 – 120 seconds	$ICC_{3,1}$ 0.95 (0.87, 0.92)‡ substantial	2.2‡	6.1‡	Lack of clarity blinding raters to own prior findings Lack of clarity blinding raters to additional cues about participants
Boland and Adams (2000) Pendulometer Group A	Symptomatic limb Inter-rater Intra-session Immediate	$ICC_{2,1}$ 0.86 (0.67, 0.94) substantial	7.2	20.0†	No important risks of bias
Group B	Symptomatic limb Inter-rater Intra-session Immediate	$ICC_{2,1}$ 0.91 (0.72, 0.97) substantial	4.8	13.3†	No important risks of bias
Pooled Groups A/B	Symptomatic limb Inter-rater Intra-session Immediate	$ICC_{1,1}$ 0.88 (0.80, 0.94) substantial	6.4	17.7†	No important risks of bias
Waddell et al. (1992) Inclinometer Subset 1§	Right limb Inter-rater Intra-session Immediate	ICC 0.87 (Not reported)‖ substantial	Not reported	Not reported	Lack of clarity varying order of examiners
Subset 1§	Left limb Inter-rater Intra-session Immediate	ICC 0.94 (Not reported)‖ substantial	Not reported	Not reported	Lack of clarity varying order of examiners
Subset 3§	Right limb Inter-rater Intra-session Immediate	ICC 0.94 (Not reported)‖ substantial	Not reported	Not reported	Lack of clarity varying order of examiners
Subset 3§	Left limb Inter-rater Intra-session Immediate	ICC 0.96 (Not reported)‖ substantial	Not reported	Not reported	Lack of clarity varying order of examiners
Boland and Adams (2000) Pendulometer DF + SLR Group A	Symptomatic limb Inter-rater Intra-session Immediate	$ICC_{2,1}$ 0.89 (0.59, 0.96) substantial	6.7	18.6†	No important risks of bias
DF + SLR Group B	Symptomatic limb Inter-rater Intra-session Immediate	$ICC_{2,1}$ 0.91 (0.75, 0.97) substantial	4.8	13.3†	No important risks of bias
DF + SLR Pooled Groups A/B	Symptomatic limb Inter-rater Intra-session Immediate	$ICC_{1,1}$ 0.89 (0.80, 0.94) substantial	6.1	16.9†	No important risks of bias

* Calculated from data reported by the authors using highest standard deviation of the two raters for most conservative estimate.
† Calculated from data reported by the authors.
‡ Calculated from raw data reported by the authors.
§ Pain tolerance, rather than pain onset.
‖ ICC model not specified.
CI, confidence interval; ICC, intraclass correlation coefficient; SDD_{95}, smallest detectable difference at 95% confidence level; SEM, standard error of measurement.

of clarity about blinding raters to findings of other raters, failure or lack of clarity about varying the order of raters, and lack of clarity about a suitable time interval between examinations. Except for structural differentiation, these issues affected a relatively small proportion of included studies (Table 2, Fig. 2). However, they impacted most estimates of inter-rater reliability for identifying a positive SLR or crossed SLR because three studies (McCombe et al., 1989; Vroomen et al., 2000; Poiraudeau et al., 2001) provided data for multiple definitions of a positive test (Figs. 3–8). This illustrates the importance of reporting potential risks of bias for each reliability outcome, not just for each study included in the review or each item on the assessment tool (Büttner et al., 2020). Additional risks of bias for intra-rater reliability were lack of clarity about blinding raters to own prior findings and to additional cues about participants. Incomplete reporting may be due to most studies being published prior to the dissemination of QAREL (Lucas et al., 2010) and Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (Kottner et al., 2011). Future reliability research should incorporate these resources into study design and reporting of results.

Limitations of our review need to be acknowledged. Only MEDLINE and CINAHL were searched and there was no search of grey literature. However, we believed these were the two most relevant databases based on our familiarity with SLR test literature. The comprehensiveness of the search strategy is supported by the fact that only two of the included studies (Waddell et al., 1982; Billis et al., 2012) were not identified by database searches. Only English language studies were included because we did not have translation resources. There is not established methodology for applying GRADE principles when judging confidence in the evidence for reliability outcomes. Consistent with GRADE principles, we have been transparent about the decision-making process for judging confidence in the evidence (Guyatt et al., 2011; Santesso et al., 2016). Researchers and clinicians should consider this information when interpreting our results. Lastly, the small number of studies for various reliability outcomes may have biased estimates of statistical heterogeneity and prevented assessment of publication bias.

5. Conclusions

Reliability data suggest clinicians should use structural differentiation manoeuvres during SLR testing. Lack of clarity about crossed SLR reliability means this test should be interpreted cautiously. Measurement error likely prohibits using SLR ROM for clinical decision-making. Confidence in the evidence for SLR and crossed SLR reliability could increase if future research adheres to published guides for improving reliability study design and reporting.

Funding sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Prospective registration

PROSPERO CRD42018086158.

Declaration of interest

The straight leg raise (SLR) is a neurodynamic test. The authors (RJN, MWC, BSB) present continuing education courses on neurodynamic testing and treatment to clinicians (mainly physical therapists) on behalf of the Neuro Orthopaedic Institute (NOI) Group. They are paid as independent contractors for their teaching. The outcomes of this review will not impact the authors' relationship with NOI Group, and therefore this affiliation did not affect judgments the authors made for this review. The authors have no other financial relationships with the NOI Group nor their products. This review is completely separate from NOI Group. The authors declare that they have no other known conflicts of interest.

Acknowledgments

The authors thank Eva Sierra-Silvestre for assistance in screening titles and abstracts.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.msksp.2022.102529.

References

Basson, A., Olivier, B., Ellis, R., Coppieters, M., Stewart, A., Mudzi, W., 2017. The effectiveness of neural mobilization for neuro-musculoskeletal conditions: a systematic review and meta-analysis. J. Orthop. Sports Phys. Ther. 47, 593–615. https://doi.org/10.2519/jospt.2017.7117.

Bertilson, B., Bring, J., Sjöblom, A., Sundell, K., Strender, L., 2006. Inter-examiner reliability in the assessment of low back pain (LBP) using the Kirkaldy-Willis classification (KWC). Eur. Spine J. 15, 1695–1703. https://doi.org/10.1007/s00586-005-0050-3.

Billis, E., McCarthy, C., Gliatis, J., Gittins, M., Papandreou, M., Oldham, J., 2012. Inter-tester reliability of discriminatory examination items for sub-classifying non-specific low back pain. J. Rehabil. Med. 44, 851–857. https://doi.org/10.2340/16501977-0950.

Boland, R., Adams, R., 2000. Effects of ankle dorsiflexion on range and reliability of straight leg raising. Aust. J. Physiother. 46, 191–200.

Breig, A., Troup, J., 1979. Biomechanical considerations in the straight-leg-raising test: cadaveric and clinical studies of the effects of medial hip rotation. Spine 4, 242–250.

Bueno-Gracia, E., Estébanez-de-Miguel, E., López-de-Celis, C., Shacklock, M., Caudevilla-Polo, S., González-Rueda, V., et al., 2020. Effect of ankle dorsiflexion on displacement and strain in the tibial nerve and biceps femoris muscle at the posterior knee during the straight leg raise: investigation of specificity of nerve movement. Clin. BioMech. 75, 105003 https://doi.org/10.1016/j.clinbiomech.2020.105003.

Bueno-Gracia, E., Pérez-Bellmunt, A., Estébanez-de-Miguel, E., López-de-Celis, C., Shacklock, M., Caudevilla-Polo, S., et al., 2019. Differential movement of the sciatic nerve and hamstrings during the straight leg raise with ankle dorsiflexion: implications for diagnosis of neural aspect to hamstring disorders. will angungand au 43, 91–95. https://doi.org/10.1016/j.msksp.2019.07.011.

Büttner, F., Winters, M., Delahunt, E., Elbers, R., Lura, C., Khan, K., et al., 2020. Identifying the 'Incredible'! Part 2: spot the difference - a regirous risk of bias assessment can alter the main findings of a systematic review. Br. J. Sports Med. 54, 801–808. https://doi.org/10.1136/bjsports-2019-101675.

Chow, R., Adams, R., Herbert, R., 1994. Straight leg raise test high reliability is not a motor memory artefact. Aust. J. Physiother. 40, 107–111. https://doi.org/10.1016/S0004-9514(14)60457-8.

Delitto, A., George, S., Van Dillen, L., Whitman, J., Sowa, G., Shekelle, P., et al., 2012. Low back pain: Clinical Practice Guidelines linked to the International Classification of Functioning, Disability, and Health from the Orthopaedic Section of the American Physical Therapy Association. J. Orthop. Sports Phys. Ther. 42, A1–A57. https://doi.org/10.2519/jospt.2012.0301.

Eliasziw, M., Young, S., Woodbury, M., Fryday-Field, K., 1994. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys. Ther. 74, 777–788.

Fritz, J., Wainner, R., 2001. Examining diagnostic tests: an evidence-based perspective. Phys. Ther. 81, 1546–1564. https://doi.org/10.1093/ptj/81.9.1546.

George, S., Fritz, J., Silfies, S., Schneider, M., Beneciuk, J., Lentz, T., et al., 2021. Interventions for the management of acute and chronic low back pain: revision 2021. J. Orthop. Sports Phys. Ther. 51, CPG1–CPG60. https://doi.org/10.2519/jospt.2021.0304.

Gilbert, K., Brismee, J., Collins, D., James, C., Shah, R., Sawyer, S., et al., 2007. 2006 Young Investigator Award Winner: lumbosacral nerve root displacement and strain; Part 1. A novel measurement technique during straight leg raise in unembalmed cadavers. Spine 32, 1513–1520. https://doi.org/10.1097/BRS.0b013e318067dd55.

Guyatt, G., Oxman, A., Akl, E., Kunz, R., Vist, G., Brozek, J., et al., 2011. GRADE Guidelines: 1. Introduction - GRADE evidence profiles and summary of findings tables. J. Clin. Epidemiol. 64, 383–394. https://doi.org/10.1016/j.jclinepi.2010.04.026.

Harrisson, S., Stynes, S., Dunn, K., Foster, N., Konstantinou, K., 2017. Neuropathic pain in low back-related leg pain patients: what is the evidence of prevalence, characteristics, and prognosis in primary care? A systematic review of the literature. J. Pain 18, 1295–1312. https://doi.org/10.1016/j.jpain.2017.04.012.

Hartvigsen, L., Hestbaek, L., Lebouef-Yde, C., Vach, W., Kongsted, A., 2017. Leg pain location and neurological signs relate to outcomes in primary care patients with low back pain. BMC Muscoskel. Disord. 18, 133. https://doi.org/10.1186/s12891-017-1495-3.

Higgins, J., Thompson, S., Deeks, J., Altman, D., 2003. Measuring inconsistency in meta-analyses. BMJ 327, 557–560. https://doi.org/10.1136/bmj.327.7414.557.

Kigozi, J., Konstantinou, K., Ogollah, R., Dunn, K., Martyn, L., Jowett, S., 2019. Factors associated with costs and health outcomes in patients with back and leg pain in primary care: a prospective cohort analysis. BMC Health Serv. Res. 19, 406. https://doi.org/10.1186/s12913-019-4257-0.

Konstantinou, K., Dunn, K., Ogollah, R., Lewis, M., van der Windt, D., Hay, E., 2018. Prognosis of sciatica and back-related leg pain in primary care: the ATLAS cohort. Spine J. 18, 1030–1040. https://doi.org/10.1016/j.spinee.2017.10.071.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B., Hróbjartsson, A., et al., 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J. Clin. Epidemiol. 64, 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002.

Landis, J., Koch, G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lucas, N., Macaskill, P., Irwig, L., Bogduk, N., 2010. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J. Clin. Epidemiol. 63, 854–861. https://doi.org/10.1016/j.jclinepi.2009.10.002.

Lucas, N., Macaskill, P., Irwig, L., Moran, R., Rickards, L., Turner, R., et al., 2013. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). BMC Med. Res. Methodol. 13, 111. https://doi.org/10.1186/1471-2288-13-111.

McCombe, P., Fairbank, J., Cocersole, B., Pynsent, P., 1989. Volvo Award in clinical sciences. Reproducibility of physical signs in low-back pain. Spine 14, 908–918. https://doi.org/10.1097/00007632-198909000-00002, 1989.

McHugh, M., 2012. Interrater reliability: the kappa statistic. Biochem. Med. 22, 276–282.

Mistry, J., Heneghan, N., Noblet, T., Falla, D., Rushton, A., 2020. Diagnostic utility of patient history, clinical examination and screening tool data to identify neuropathic pain in low back related leg pain: a systematic review and narrative synthesis. BMC Muscoskel. Disord. 21, 532. https://doi.org/10.1186/s12891-020-03436-6.

Oliveira, C., Maher, C., Pinto, R., Traeger, A., Lin, C., Chenot, J., et al., 2018. Clinical practice guidelines for the management of non-specific low back pain in primary care: an updated overview. Eur. Spine J. 27, 2791–2803. https://doi.org/10.1007/s00586-018-5673-2.

Paatelma, M., Karvonen, E., Heinonen, A., 2010. Inter- and intra-tester reliability of selected clinical tests in examining patients with early phase lumbar spine and sacroiliac joint pain and dysfunction. Adv. Physiother. 12, 74–80. https://doi.org/10.3109/14038190903582154.

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., et al., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n1. https://doi.org/10.1136/bmj.n71.

Pesonen, J., Shacklock, M., Rantanen, P., Mäki, J., Karttunen, L., Kankaanpää, M., et al., 2021. Extending the straight leg raise test for improved clinical evaluation of sciatica: reliability of hip internal rotation or ankle dorsiflexion. BMC Muscoskel. Disord. 22, 303. https://doi.org/10.1186/s12891-021-04159-y.

Poiraudeau, S., Foltz, V., Drapé, J., Fermanian, J., Lefèvre-Colau, M., Mayoux-Benhamou, M., et al., 2001. Value of the bell test and the hyperextension test for diagnosis in sciatica associated with disc herniation: comparison with Lasègue's sign and the crossed Lasègue's sign. Rheumatology 40, 460–466. https://doi.org/10.1093/rheumatology/40.4.460.

Pourahmadi, M., Hesarikia, H., Keshtkar, A., Zamani, H., Bagheri, R., Ghanjal, A., et al., 2019. Effectiveness of slump stretching on low back pain: a systematic review and meta-analysis. Pain Med. 20, 378–396. https://doi.org/10.1093/pm/pny208.

Rade, M., Shacklock, M., Könönen, M., Marttila, J., Vanninen, R., Kankaanpää, M., et al., 2017. Normal multiplanar movement of the spinal cord during unilateral and bilateral straight leg raise: quantification, mechanisms, and overview. J. Orthop. Res. 35, 1335–1342. https://doi.org/10.1002/jor.23385.

Rebain, R., Baxter, G., McDonough, S., 2002. A systematic review of the passive straight leg raising test as a diagnostic aid for low back pain (1989 to 2000). Spine 27, E388–E395. https://doi.org/10.1097/00007632-200209010-00025.

Sackett, D., 1992. The rational clinical examination. A primer on the precision and accuracy of the clinical examination. JAMA 267, 2638–2644.

Santesso, N., Carrasco-Labra, A., Lagnendam, M., Brignardello-Petersen, R., Mustafa, R., Heus, P., et al., 2016. Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments. J. Clin. Epidemiol. 74, 28–39. https://doi.org/10.1016/j.jclinepi.2015.12.006.

Scaia, V., Baxter, D., Cook, C., 2012. The pain provocation-based straight leg raise test for diagnosis of lumbar disc herniation, lumbar radiculopathy, and/or sciatica: a systematic review of clinical utility. J. Back Musculoskelet. Rehabil. 25, 215–223. https://doi.org/10.3233/BMR-2012-0339.

Schünemann, H., 2016. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J. Clin. Epidemiol. 75, 6–15. https://doi.org/10.1016/j.jclinepi.2016.03.018.

Schünemann, H., Mustafa, R., Brozek, J., Steingart, K., Leeflang, M., Murad, M., et al., 2020a. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J. Clin. Epidemiol. 122, 129–141. https://doi.org/10.1016/j.jclinepi.2019.12.020.

Schünemann, H., Mustafa, R., Brozek, J., Steingart, K., Leeflang, M., Murad, M., et al., 2020b. GRADE guidelines: 21 part 2. Test accuracy: inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiels and summary of findings tables. J. Clin. Epidemiol. 122, 142–152. https://doi.org/10.1016/j.jclinepi.2019.12.021.

Shrout, P., 1998. Measurement reliability and agreement in psychiatry. Stat. Methods Med. Res. 7, 301–317.

Shrout, P., Fleiss, J., 1979. Intraclass correlations: uses in assessing rater relaibility. Psychol. Bull. 86, 420–428.

Sim, J., Wright, C., 2005. The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys. Ther. 85, 257–268.

Smart, K., Blake, C., Staines, A., Thacker, M., Doody, C., 2012a. Mechanisms-based classifications of musculoskeletal pain: Part 1 of 3: symptoms and signs of central sensitisation in patients with low back (± leg) pain. Man. Ther. 17, 336–344. https://doi.org/10.1016/j.math.2012.03.013.

Smart, K., Blake, C., Staines, A., Thacker, M., Doody, C., 2012b. Mechanisms-based classifications of musculoskeletal pain: Part 2 of 3: symptoms and signs of peripheral neuropathic pain in patients with low back (± leg) pain. Man. Ther. 17, 345–351. https://doi.org/10.1016/j.math.2012.03.003.

Stratford, P., 2004. Getting more from the literature: estimating the standard error of measurement from reliability studies. Physiother. Can. 56, 27–30.

Strender, L., Sjoblom, A., Sundell, K., Ludwig, R., Taube, A., 1997. Interexaminer reliability in physical examination of patients with low back pain. Spine 22, 814–820.

Stynes, S., Konstantinou, K., Dunn, K., Lewis, M., Hay, E., 2016. Reliability among clinincians diagnosing low back-related leg pain. Eur. Spine J. 25, 2734–2740. https://doi.org/10.1007/s00586-015-4359-2.

Stynes, S., Konstantinou, K., Ogollah, R., Hay, E., Dunn, K., 2018. Clinical diagnostic model for sciatica developed in primary care patients with low back-related leg pain. PLoS One 13, e0191852. https://doi.org/10.1371/journal.pone.0191852.

Sun, S., 2011. Meta-analysis of Cohen's kappa. Health Serv. Outcome Res. Methodol. 11, 145–163. https://doi.org/10.1007/s10742-011-0077-3.

Tawa, N., Rhoda, A., Diener, I., 2017. Accuracy of clinical neurological examination in diagnosing lumbo-sacral radiculopathy: a systematic literature review. BMC Muscoskel. Disord. 18, 93. https://doi.org/10.1186/s12891-016-1383-2.

Troup, J., 1981. Straight-leg-raising (SLR) and the qualifying tests for increased root tension: their predictive value after back and sciatic pain. Spine 6, 526–527.

van den Hoogen, H., Koes, B., Devillé, W., van Eijk, J., Bouter, L., 1996. The inter-observer reproducibility of Lasègue's sign in patients with low back pain in general practice. Br. J. Gen. Pract. 46, 727–730.

van der Windt, D., Simons, E., Riphagen, I., Ammendolia, C., Verhagen, A., Laslett, M., et al., 2010. Physical examination for lumbar radiculopathy due to disc herniation in patients with low back pain. Cochrane Database Syst. Rev. (2), CD007431 https://doi.org/10.1002/14651858.CD007431.pub2.

Vroomen, P., de Krom, M., Knotterus, J., 2000. Consistency of history taking and physical examination in patients with suspected lumbar nerve root involvement. Spine 25, 91–96. https://doi.org/10.1097/00007632-200001010-00016 discussion 7.

Waddell, G., Main, C., Morris, E., Venner, R., Rae, P., Sharmy, S., et al., 1982. Normality and reliability of the clinical assessment of backache. Br. Med. J. 284, 1519–1523. https://doi.org/10.1136/bmj.284.6328.1519.

Waddell, G., Somerville, D., Henderson, I., Newton, M., 1992. Objective clinical evaluation of physical impairment in chronic low back pain. Spine 17, 617–628. https://doi.org/10.1097/00007632-199206000-00001.

Walsh, J., Hall, T., 2009a. Agreement and correlation between the straight leg raise and slump tests in subjects with leg pain. J. Manipulative Physiol. Therapeut. 32, 184–192. https://doi.org/10.1016/j.jmpt.2009.02.006.

Walsh, J., Hall, T., 2009b. Reliability, validity and diagnostic accuracy of palpation of the sciatic, tibial and common peroneal nerves in the examination of low back-related leg pain. Man. Ther. 14, 623–629. https://doi.org/10.1016/j.math.2008.12.007.